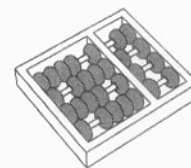


Fábio Kenji Masago

# “Odysseýs: Sistema para Análise de Documentos de Patentes”

CAMPINAS  
2013





Universidade Estadual de Campinas  
Instituto de Computação

Fábio Kenji Masago

## “Odysseýs: Sistema para Análise de Documentos de Patentes”

Orientador(a): Prof. Dr. Jacques Wainer

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Mestre em Ciência da Computação.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA POR FÁBIO KENJI MASAGO, SOB ORIENTAÇÃO DE PROF. DR. JACQUES WAINER.

  
Assinatura do Orientador(a)

CAMPINAS

2013

iii

Ficha catalográfica  
Universidade Estadual de Campinas  
Biblioteca do Instituto de Matemática, Estatística e Computação Científica  
Ana Regina Machado - CRB 8/5467

M37o Masago, Fábio Kenji, 1984-  
Odysseýs sistema para análise de documentos de patentes / Fábio Kenji  
Masago. – Campinas, SP : [s.n.], 2013.

Orientador: Jacques Wainer.  
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de  
Computação.

1. Inteligência artificial - Programas de computador. 2. Mineração de dados  
(Computação). 3. Processamento de textos (Computação). 4. Programas de  
computador - Patentes. 5. Análise de algoritmos. I. Wainer, Jacques, 1958-. II.  
Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

**Título em inglês:** Odysseýs system for analysis of patent documents

**Palavras-chave em inglês:**

Artificial intelligence - Computer programs

Data mining

Text processing (Computer science)

Computer programs - Patents

Algorithm analysis

**Área de concentração:** Ciência da Computação

**Titulação:** Mestre em Ciência da Computação

**Banca examinadora:**

Jacques Wainer [Orientador]

Maria Ester Soares Dal Poz

Siome Klein Goldenstein

**Data de defesa:** 08-04-2013

**Programa de Pós-Graduação:** Ciência da Computação

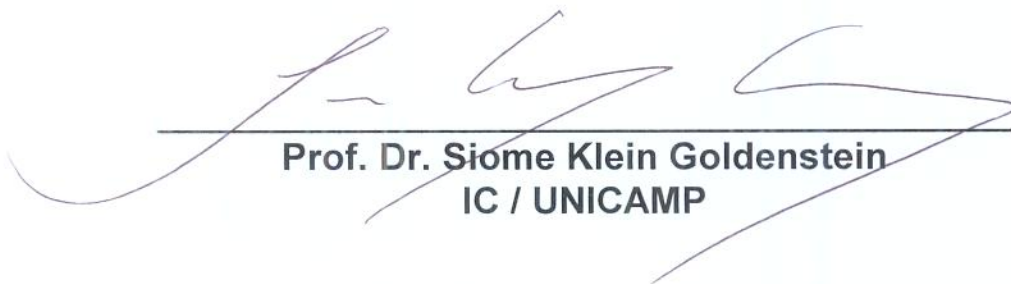
## TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 08 de Abril de 2013, pela  
Banca examinadora composta pelos Professores Doutores:



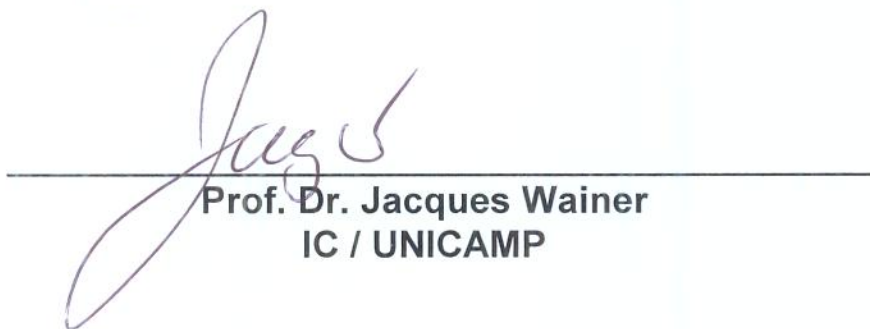
---

**Profª. Drª. Maria Ester Soares Dal Poz**  
**FCA / UNICAMP**



---

**Prof. Dr. Siome Klein Goldenstein**  
**IC / UNICAMP**



---

**Prof. Dr. Jacques Wainer**  
**IC / UNICAMP**



# Odysseýs: Sistema para Análise de Documentos de Patentes

Fábio Kenji Masago

08 de abril de 2013

## Banca Examinadora:

- Prof. Dr. Jacques Wainer (*Orientador*)
- Prof. Dr. Siome Klein Goldenstein  
Instituto de Computação - UNICAMP
- Prof. Dr. Anderson de Rezende Rocha  
Instituto de Computação - UNICAMP
- Profa. Dra. Maria Ester Soares Dal Poz  
Faculdade de Ciências Aplicadas - UNICAMP
- Prof. Dr. Jaime Simão Sichman  
Departamento de Engenharia de Computação e Sistemas Digitais - USP





# Abstract

A patent is a document about an invention's property given by the state to authors, preventing others from producing, using, commercialize, importing and exporting the described invention without a permission of the document's owner.

A study in the economic area frequently used is the use of patents to measure importance or technological impact of an innovative field of an entity or nation. Thus, can be asserted that patents are a kind of inventive level meter and their citations is a form of measuring a country's or firm's flow or the impact of knowledge, as well as evaluate trends in a certain technological field.

This thesis presents a computational tool to assist in the process of patents analysis, approaching the applicability of the method *Latent Dirichlet Allocation (LDA)* for the similarity of patents.

The computational system called *Odysseýs* evaluates the similarity between a patent given by the user and a group of documents, ordering them according to their similarity degree in relation to evaluated patent. In addition, the software allows, in an unsupervised manner, generate a patent citation's network by searches for a set of related patents in the database *United States Patent and Trademark Office (USPTO)* through a query designated by the user applying those patents to the similarity analysis, and also for generation of a knowledge flow network.

The inexistence of national software for patent processing and only a few auxiliary tools for the analysis of such documents were the main motivations for the development of this project.



# Resumo

Uma patente é um documento sobre uma propriedade de criação concedida pelo Estado aos autores, que impede terceiros a produzir, utilizar, comercializar, importar e exportar a invenção descrita sem a devida autorização do titular do documento.

Um estudo na área econômica muito empregado é a utilização de patentes para medir a importância ou impacto tecnológico de um campo inovativo de uma entidade ou nação. Pode-se afirmar que as patentes são como uma espécie de medidores do nível inventivo e as citações contidas nas patentes são um meio para medir o fluxo ou os impactos do conhecimento de um país ou firma, assim como, avaliar tendências de um campo tecnológico.

A presente dissertação de mestrado apresenta o desenvolvimento de uma ferramenta para auxiliar no procedimento de análise de patentes, abordando a aplicabilidade do método *Latent Dirichlet Allocation (LDA)* para o processo de similaridade de patentes.

O sistema computacional denominado *Odysseys* verifica a similaridade entre uma determinada patente dada pelo usuário e um grupo de documentos, ordenando-os conforme o seu grau de semelhança em relação à patente em avaliação. Além disso, o software permite, de forma não supervisionada, a geração de redes de citações de patentes por meio de buscas de um conjunto de patentes correlacionadas na base de dados do *United States Patent and Trademark Office (USPTO)* a partir de uma consulta designada pelo usuário, utilizando essas patentes para a análise de similaridade e, também, para a geração da rede de fluxo de conhecimento.

A inexistência de *softwares* nacionais específicos para o processamento de patentes e as poucas ferramentas auxiliares para a análise de tais documentos foram as principais motivações para o desenvolvimento do projeto.



# Agradecimentos

Agradeço primeiramente a Deus por me proporcionar transpassar mais um desafio em minha vida, ajudando-me a tomar, mesmo em momentos de solidão e desesperança, as decisões corretas para se abrir novos caminhos de sucesso a serem desbravados.

Assim, acredito que: *No tear que tece a nossa vida, não há pontas soltas. Todos os fios estão entremeados entre si e revestidos de significado.* Isso me faz pensar que a cada momento que passamos sozinhos ou na companhia de outras pessoas, mesmo pequeno que esse o seja, é simplesmente único e irreversível. E com isso, um simples gesto é o suficiente para que esse momento se torne especial em nossas vidas. Então, podemos dizer que a felicidade que buscamos é formada pela união desses pequenos momentos únicos de alegria que compartilhamos ou são compartilhadas por outras pessoas.

Com isso, para esse *momento único* de alegria, gostaria de agradecer principalmente aos meus pais pelo apoio incondicional em relação às minhas decisões tomadas durante a minha vida e por todos os sacrifícios feitos para que eu alcançasse a devida formação profissional com ética e dignidade.

Um agradecimento especial ao meu orientador Prof Jacques Wainer pelos momentos de paciência e apoio durante todo o meu período de mestrado.

Ao Alexandre Passos pela ajuda com o método *Latent Dirichlet Allocation*.

E, não menos importante, a todos os meus amigos e pessoas que contribuíram para a realização desse trabalho. Acharia injusto citar nomes para mostrar o quão grato me sinto para superar mais esse desafio e, assim, creio que de todas as palavras que vêm à minha mente apenas estas podem expressar todo o meu sentimento nesse exato momento em relação a todos: *Muito obrigado.*



# Sumário

Abstract	ix
Resumo	xi
Agradecimentos	xiii
<b>1 Introdução</b>	<b>1</b>
<b>2 Propriedade Intelectual</b>	<b>3</b>
2.1 Direitos de Autor . . . . .	3
2.2 Propriedade Industrial . . . . .	5
2.3 <i>Proteção Sui Generis</i> . . . . .	7
<b>3 Patentes</b>	<b>9</b>
3.1 Patentes: Conceitos . . . . .	9
3.2 Sistema de Patentes . . . . .	14
3.3 Sistema de Patentes Americano: USPTO . . . . .	15
3.4 Procedimentos para a Obtenção de uma Patente . . . . .	17
3.5 Problemas nas referências das Patentes . . . . .	19
3.6 Valor Econômico de uma patente . . . . .	20
3.7 Tendências de Mercado . . . . .	24
3.8 Redes de Citações de Patentes . . . . .	25
<b>4 Similaridade</b>	<b>27</b>
4.1 Similaridade de Texto . . . . .	27
4.2 Similaridade de Patentes . . . . .	28
4.3 Representação de Patentes . . . . .	29
4.3.1 Espaço Vetorial Binário . . . . .	30
4.3.2 <i>Term Frequency–Inverse Document Frequency (TF-IDF)</i> . . . . .	30
4.4 Métricas de Similaridade . . . . .	31





4.4.1	Distância Euclidiana . . . . .	32
4.4.2	Distância dos Cossenos . . . . .	32
4.4.3	Earth Mover’s Distance . . . . .	32
<b>5</b>	<b>Extração de Tópicos</b>	<b>35</b>
5.1	<i>Latent Dirichlet Allocation</i> . . . . .	36
5.2	<i>Collapsed Gibbs Sampling</i> . . . . .	38
<b>6</b>	<b>Sistema</b>	<b>41</b>
6.1	Objetivo . . . . .	41
6.2	Implementação . . . . .	42
6.3	Metodologia . . . . .	43
6.4	Testes e Avaliação . . . . .	47
6.4.1	Análise de grupos tecnológicos distintos de patentes . . . . .	48
6.4.1.1	Experimento 1 . . . . .	50
6.4.1.2	Experimento 2 . . . . .	51
6.4.2	Análise de grupos tecnológicos similares de patentes . . . . .	53
6.4.2.1	Experimento 1 . . . . .	54
6.4.2.2	Experimento 2 . . . . .	57
6.4.2.3	Experimento 3 . . . . .	60
6.5	Discussão dos Resultados . . . . .	62
6.5.1	Avaliação para grupos tecnológicos distintos de patentes . . . . .	63
6.5.2	Avaliação para grupos tecnológicos similares de patentes . . . . .	63
<b>7</b>	<b>Trabalhos Relacionados</b>	<b>67</b>
<b>8</b>	<b>Conclusão</b>	<b>69</b>
	<b>Referências Bibliográficas</b>	<b>71</b>
	<b>Glossário</b>	<b>77</b>
<b>A</b>	<b>Propriedade Intelectual no Brasil</b>	<b>89</b>
A.1	Quadro Atual . . . . .	89
<b>B</b>	<b>Análise Manual de Similaridade de Patentes</b>	<b>91</b>
B.1	1ª tentativa . . . . .	92
B.2	2ª tentativa . . . . .	93
B.3	3ª tentativa . . . . .	94
B.4	4ª tentativa . . . . .	94



B.5	5ª tentativa . . . . .	95
B.6	6ª tentativa . . . . .	96
<b>C</b>	<b>Tabelas de Similaridade</b>	<b>99</b>
C.1	Análise de grupos tecnológicos distintos de patentes . . . . .	99
C.1.1	Experimento 1 . . . . .	99
C.1.2	Experimento 2 . . . . .	103
C.2	Análise de grupos tecnológicos similares de patentes . . . . .	106
C.2.1	Experimento 1 . . . . .	106
C.2.2	Experimento 2 . . . . .	107
C.2.3	Experimento 3 . . . . .	107
<b>D</b>	<b>Tabelas e Listas de Tópicos de <i>LDA with Gibbs Sampling</i></b>	<b>109</b>



# Lista de Tabelas

6.1	Patentes utilizadas para a execução de testes de similaridade . . . . .	49
6.2	Grau de similaridade da patente 4876196 em relação aos documentos dos grupos de análise . . . . .	50
6.3	Grau de similaridade do modelo parcial de patente 4876196 em relação aos documentos dos grupos de análise . . . . .	52
6.4	Experimento 1: Grupo tecnológico utilizado para os testes de similaridade	55
6.5	Experimento 1: Ordem decrescente de similaridade para a patente 4940835 segundo o analista de patentes . . . . .	56
6.6	Experimento 1: Percentagem de similaridade das patentes em relação a patente de análise 4940835 . . . . .	56
6.7	Experimento 1: Valores do coeficiente de Spearman para os métodos aplicados	57
6.8	Experimento 2: Grupo tecnológico utilizado para os testes de similaridade	58
6.9	Experimento 2: Ordem decrescente de similaridade para a patente 5107065 segundo o analista de patentes . . . . .	58
6.10	Experimento 2: Percentagem de similaridade das patentes em relação a patente de análise 5107065 . . . . .	59
6.11	Experimento 2: Valores do coeficiente de Spearman para os métodos aplicados	59
6.12	Experimento 3: Grupo tecnológico utilizado para os testes de similaridade	60
6.13	Experimento 3: Ordem decrescente de similaridade para a patente 5352605 segundo o analista de patentes . . . . .	61
6.14	Experimento 3: Percentagem de similaridade das patentes em relação a patente de análise 5352605 . . . . .	62
6.15	Experimento 3: Valores do coeficiente de Spearman para os métodos aplicados	62
6.16	Valores do coeficiente de correlação de postos de <i>Spearman</i> para os métodos aplicados . . . . .	64
C.1	Experimento 1: Grau de similaridade obtido pelo método <i>LDA+EMD</i> para os documentos de análise em relação a patente 4876196 . . . . .	100



C.2	Experimento 1: Grau de similaridade obtido pelo método Espaço Vetorial Binário+Distância Euclidiana para os documentos de análise em relação a patente 4876196 . . . . .	101
C.3	Experimento 1: Grau de similaridade obtido pelo método <i>TF-IDF</i> +Distância dos Cossenos para os documentos de análise em relação a patente 4876196 . . . . .	102
C.4	Experimento 2: Grau de similaridade obtido pelo método <i>LDA+EMD</i> para os documentos de análise em relação ao modelo parcial de patente 4876196 . . . . .	103
C.5	Experimento 2: Grau de similaridade obtido pelo método Espaço Vetorial Binário+Distância Euclidiana para os documentos de análise em relação ao modelo parcial de patente 4876196 . . . . .	104
C.6	Experimento 2: Grau de similaridade obtido pelo método <i>TF-IDF</i> +Distância dos Cossenos para os documentos de análise em relação ao modelo parcial de patente 4876196 . . . . .	105
C.7	Experimento 1: <i>Ranking</i> de similaridade de patentes em relação ao documento <i>USPTO</i> 4940835 . . . . .	106
C.8	Experimento 2: <i>Ranking</i> de similaridade de patentes em relação ao documento <i>USPTO</i> 5107065 . . . . .	107
C.9	Experimento 3: <i>Ranking</i> de similaridade de patentes em relação ao documento <i>USPTO</i> 5352605 . . . . .	107





# Lista de Figuras

3.1	Exemplo simplificado de uma patente do <i>USPTO</i> . . . . .	12
3.2	Página de buscas avançadas do <i>USPTO</i> . . . . .	16
3.3	Resultados de busca de patentes para a <i>query TTL/(fruit) AND ABST/(DNA)</i> . . . . .	17
5.1	Modelo gráfico para <i>LDA</i> . . . . .	36
6.1	Rede de citações para a tecnologia <i>ethanol</i> conforme <i>IPC=C07H</i> . . . . .	45
6.2	Interpretação do coeficiente de correlação de postos de <i>Spearman</i> . . . . .	54



# Capítulo 1

## Introdução

Desde o aparecimento dos primeiros instrumentos para auxiliar nas tarefas manuais realizadas no cotidiano, o homem sempre procurou ferramentas mais sofisticadas para realizar tais serviços com muito mais eficácia e em um curto intervalo de tempo. Esse fato se mostrou ainda mais evidente com a Revolução Industrial em meados do século XVIII, período em que se iniciou a migração dos processos antes completamente manuais para a forma de produção automatizada.

Com a mesma concepção de automatização das atividades, a partir do crescente aumento do número de informações, tornaram-se necessários meios digitais para armazenagem de informações, além de técnicas computacionais para organizar os milhares de documentos existentes. Porém, somente utilizar as informações já organizadas da base de dados não compreende o requisito único para a extração das informações desejadas. Além disso, muitas vezes se tornam inviáveis as análises manuais de cada uma das informações retornadas, acarretando em uma tarefa completamente exaustiva e repetitiva de ser executada [10]. A partir disso, resultou-se na necessidade de utilização de ferramentas para realizar tais funções de forma automática, sem a necessidade de uma supervisão sistemática pelo ser humano.

A presente dissertação apresenta o desenvolvimento de um sistema para auxiliar no processo de análise de documentos de patentes.

O sistema computacional denominado *Odysseýs* realiza uma comparação de similaridade entre uma determinada patente dada pelo usuário e um grupo de documentos, ordenando-os conforme o seu grau de semelhança entre as tecnologias apresentadas nos documentos. Além disso, o software permite, de forma não supervisionada, buscas de um conjunto de patentes correlacionadas na base de dados do *United States Patent and Trademark Office (USPTO)*<sup>1</sup> a partir de uma consulta designada pelo usuário, utilizando essas patentes para a análise de similaridade ou para a geração de uma rede de fluxo de

---

<sup>1</sup>*United States Patent and Trademark Office - USPTO. Disponível em: <http://www.uspto.gov/>*

conhecimento.

A inexistência de *softwares* nacionais específicos para o processamento de patentes e as poucas ferramentas auxiliares para a análise de tais documentos são as principais motivações para o desenvolvimento do projeto.

Os capítulos estão organizados da seguinte maneira: a próxima seção (Capítulo 2) apresenta uma introdução aos conceitos de Propriedade Intelectual e suas principais subdivisões – Direitos de Autor, Propriedade Industrial e *Proteção Sui Generis*; na seção seguinte (Capítulo 3) realiza-se uma breve descrição sobre patentes, abordando conceitos essenciais, histórico, aquisição, importância e redes de citações. Mais à frente (Capítulo 4), apresenta-se uma introdução sobre similaridade de documentos e principais métodos utilizados; na próxima seção (Capítulo 5), aborda-se basicamente o método *Latent Dirichlet Allocation* para a Extração de tópicos; já no Capítulo 6, descreve-se o sistema, a sua metodologia de desenvolvimento e os resultados obtidos para a similaridade de documentos por meio do método *Latent Dirichlet Allocation*; no Capítulo 7, descrevem-se brevemente alguns trabalhos similares desenvolvidos na área; por fim (Capítulo 8) encontram-se as conclusões para o projeto.

# Capítulo 2

## Propriedade Intelectual

Segundo a definição da Organização Mundial da Propriedade Intelectual (*OMPI ou WIPO - World Intellectual Property Organization*), o termo Propriedade Intelectual (*PI*) pode ser definido como:

*“às obras literárias, artísticas e científicas, às interpretações dos artistas intérpretes e às execuções dos artistas executantes, aos fonogramas e às emissões de radiodifusão, às invenções em todos os domínios da atividade humana, às descobertas científicas, aos desenhos e modelos industriais, às marcas industriais, comerciais e de serviço, bem como às firmas comerciais e denominações comerciais, à proteção contra a concorrência desleal e todos os outros direitos inerentes à atividade intelectual nos domínios industrial, científico, literário e artístico.”* (Estocolmo, 14 de julho de 1967. Artigo 2, VIII)<sup>1</sup>.

O termo *propriedade* pode ser generalizado como um direito estabelecido, conforme as legislações nacionais, sobre a posse legal de alguma coisa. O seu titular possui o livre arbítrio para utilização da propriedade da forma que lhe bem convier, inclusive cedendo os direitos ou licenciando a terceiros, contanto que esteja de acordo com as leis e não ponha em risco a saúde e a segurança humana. Assim, de forma geral, o termo *propriedade intelectual* pode ser entendido como todo o tipo de propriedade gerado pela criação do espírito humano [7].

Atualmente, a propriedade intelectual se divide em três categorias principais: Direitos de Autor, Propriedade Industrial e *Proteção Sui Generis*.

### 2.1 Direitos de Autor

Os direitos de autor compreendem a proteção de obras de origem científica, literária e artística, isto é, a proteção de expressão e pensamentos, e não a ideia em si da invenção.

---

<sup>1</sup>WIPO – World Intellectual Property Organization. Disponível em: <http://www.wipo.int/>

É nessa categoria de propriedade intelectual que se incluem os textos como, por exemplo, livros, traduções e enciclopédias; as músicas; as obras de arte e as obras tecnológicas, como os programas de computador. Para a obtenção da proteção dos direitos autorais, a obra deve obrigatoriamente ser *original*, podendo esse requisito de originalidade variar conforme as leis de cada país onde se deseja obter a proteção da invenção [7].

É por meio dos direitos de autor que o titular da obra obtém os direitos exclusivos concedidos pela legislação nacional. Com isso, o detentor pode autorizar ou inibir terceiros da utilização, reprodução, interpretação, radiodifusão e publicação de trechos completos ou parciais da obra original, inclusive os seus derivados. As obras derivadas englobam as traduções para outros idiomas, as adaptações, arranjos musicais e qualquer outro tipo de alteração que possa ocorrer na obra original.

Para uma melhor organização, os direitos de autor podem ser subdivididos em duas categorias:

- Direitos autorais: abrangem os direitos de obras literárias, científicas e artísticas. Disposto na legislação brasileira conforme a lei nº 9.610 de 19 de fevereiro de 1998 [8];
- Direitos Conexos: originam-se de uma obra protegida pelo direito autoral, sendo esse conceito relacionado com a proteção de direitos autorais. Os direitos Conexos protegem a pessoa física e jurídica, contribuindo para tornar as obras autorais acessíveis ao público [7]. É nesse contexto que se enquadram as interpretações, as transmissões de rádio (meios de radiodifusão) e os programas de computador. Segundo a legislação da *PI*, os direitos de autor para sistemas computacionais abrangem o código fonte, o objeto ou o executável. Na legislação brasileira, as leis para os direitos conexos estão dispostos no decreto nº 57.125 de 19 de outubro de 1965 e no decreto nº 4.533 de 19 de dezembro de 2002 [8].

Normalmente, ao se desenvolver uma determinada obra, apenas o fato de sua criação já a torna suficiente para a sua proteção, não sendo necessário realizar qualquer outro procedimento em específico. Uma exceção a essa regra, seria a aquisição dos direitos de autor em países com tradição de *common law*, nos quais não basta somente a criação da obra para ela se tornar protegida, mas também se necessita o procedimento de *fixação* da obra a ser protegida. O termo *fixação* significa que a obra deve ser registrada de uma forma escrita ou gravada, isto é, em forma de um documento para comprovação de autoria perante a um tribunal jurídico. A Convenção de Berna, uma das mais antigas convenções internacionais (criada em 1886), estabelece a seguinte definição para os direitos de autorais:

*“Os termos ‘obras literárias e artísticas’ abrangem todas as produções do domínio literário, científico e artístico, qualquer que seja o modo ou a forma de expressão, tais como*

os livros, brochuras e outros escritos; as conferências, alocações, sermões e outras obras da mesma natureza; as obras dramáticas ou dramático-musicais; as obras coreográficas e as pantomimas; as composições musicais, com ou sem palavras; as obras cinematográficas e as expressas por processo análogo ao da cinematografia; as obras de desenho, de pintura, de arquitetura, de escultura, de gravura e de litografia; as obras fotográficas e as expressas por processo análogo ao da fotografia; as obras de arte aplicada; as ilustrações e os mapas geográficos; os projetos, esboços e obras plásticas relativos à geografia, à topografia, à arquitetura ou às ciências” [4].

Dessa forma, pode-se dizer resumidamente que a *expressão humana* é o fator determinante para a condição de proteção de uma determinada obra derivada da criatividade intelectual do ser humano. Além disso, a Convenção de Berna adota critérios como a não existência de qualquer tipo de discriminação de obras vindas de outros países membros [7].

Também, outro tratado complementar a Convenção de Berna vem sendo adotado recentemente, o Acordo Relativo aos Aspectos do Direito da Propriedade Intelectual Relacionados com o Comércio (*ADPIC* ou, em inglês, *Trade Related Aspects of Intellectual Property Rights - TRIPs*), responsável pelos tratados de direitos de propriedade intelectual ligados ao comércio entre os seus países membros. O *ADPIC* é administrado pela Organização Mundial do Comércio (*OMC*, ou em inglês, *World Trade Organization - WTO*) e possui como principal requisito para a admissão de países membros a sujeição às normas da Convenção de Berna [7] [4].

## 2.2 Propriedade Industrial

A propriedade Industrial abrange os inventos de aplicação industrial e comercial, concebendo, assim, a proteção jurídica dos bens tecnicamente aplicáveis aos segmentos da produção industrial. Nessa categoria se encontra o conteúdo relacionado às Patentes, ao Desenho Industrial, as Marcas e as Indicações Geográficas.

Dessa forma, ressalta-se que é nesse capítulo se encontra o principal foco de análise utilizado para toda a dissertação de mestrado – as patentes.

- **Patentes:** Patente é um título temporário de propriedade sobre uma invenção ou modelo de utilidade, outorgado pelo Estado aos inventores, autores ou outras pessoas físicas ou jurídicas detentoras de direitos sobre a criação. Em contrapartida aos privilégios recebidos, o inventor se obriga a revelar detalhadamente todo o conteúdo técnico da matéria protegida pela patente. A obtenção de uma patente para um invento, segundo o Art. 6º (LPI, 9.279/96), o autor da invenção deve apresentar um resultado completamente novo para um problema existente, visando um efeito

técnico em uma determinada área tecnológica [8]. Além disso, deve-se também obedecer aos critérios de atividade inventiva e aplicabilidade industrial. No capítulo seguinte (Capítulo 3) serão descritos mais detalhadamente os conceitos e a forma de aquisição de patentes.

- **Desenho Industrial:** Desenho industrial pode ser definido como qualquer forma ou característica de um objeto apresentando um novo aspecto visual e, dessa forma, servindo como um tipo de fabricação industrial [8]. Assim como as patentes, a proteção do desenho industrial oferece direitos exclusivos para o seu titular, excluindo terceiros de produzir, comercializar ou importar produtos que possuam ou contenham cópias do modelo protegido. A proteção dos desenhos industriais é feita por meio do registro do *design* do objeto no país onde se deseja obter a proteção, podendo variar conforme a legislação vigente em cada região [7]. O tempo de vigência da proteção para os países concordantes do *ADPIC* é de no mínimo 10 anos, mas, também, pode variar conforme as leis locais de cada país<sup>2</sup>.
- **Marcas:** Uma Marca pode ser definida como um conjunto de sinais visuais que fornece a individualidade para uma empresa em comparação aos seus concorrentes, possuindo um caráter distintivo e não enganoso dos produtos e serviços. O termo *distintivo* refere-se à necessidade da marca não ser de caráter descritivo, pois, uma vez que a marca seja um sinal comum para uma área de produtos e serviços da empresa, os seus concorrentes não poderão utilizá-la para descrever os seus produtos [7]. Isso significa que para uma empresa que vende frutas, por exemplo, esta não poderá utilizar a marca *fruta* para designar seus produtos, uma vez que os seus concorrentes que vendem frutas também possam empregar o mesmo termo para descrever os seus produtos. Para a proteção de uma marca, assim como os desenhos industriais, pode-se realizar de um registro em um órgão nacional destinado a proteção de marcas, podendo a forma de aquisição desse registro variar conforme a legislação onde será registrado a marca do produto. Resumidamente, pode-se dizer que as marcas simplificam a identificação de certos produtos e serviços pelos consumidores, além de se tornar uma forma de comunicação para fins de *marketing* em uma empresa.
- **Indicações Geográficas:** Indicação Geográfica é um meio de indicar a procedência de produção ou o local de extração de um determinado produto ou serviço conforme o nome da região, da cidade ou do país de origem que implicou na fama ou reconhecimento desse pelo público [8]. No Brasil, o *INPI* (Instituto Nacional da Propriedade

---

<sup>2</sup>WTO – World Trade Organization. Disponível em: <http://www.wto.org/>



Industrial) é o órgão nacional responsável pelo registro das indicações geográficas, sendo amparado pela Lei nº 9.279, de 14 de maio de 1996 [7].

## 2.3 *Proteção Sui Generis*

*Proteção Sui Generis*, também conhecida como *híbridos jurídicos*, é a forma de denominação das novas criações intelectuais que não se enquadram por completo em nenhuma das modalidades anteriormente fixadas de Propriedade Intelectual, podendo a invenção possuir características tanto dos Direitos de Autor quanto da Propriedade Industrial. É nessa categoria que se encontram a topologia dos circuitos integrados, a proteção de cultivares (ou obtenção de vegetais ou variedades de vegetais) e os conhecimentos relacionados aos recursos genéticos [8].

Assim como os demais tipos de proteção intelectual, a requisição da *proteção sui generis* deve ser realizada junto ao órgão nacional responsável pelo registro da invenção, podendo esse decidir pela aprovação ou denegação do invento apresentado. Além disso, os procedimentos e o prazo de vigência do direito concedido também ficam a cargo de decisão do órgão nacional de cada país onde se deseja obter a *proteção sui generis*.

No Brasil, o pedido de registro da proteção *sui generis* é realizado perante o Nacional da Propriedade Industrial (*INPI*) que concede o prazo mínimo de proteção de 10 anos, sendo contados a partir da data do depósito do pedido de registro ou da primeira exploração da invenção.



# Capítulo 3

## Patentes

Patentes são consideradas um dos meios mais antigos de proteção da propriedade intelectual, possuindo como finalidade incentivar o desenvolvimento tecnológico e econômico de uma nação.

Neste capítulo abordam-se conceitos essenciais sobre patentes, sistemas de patentes, processos de aquisição, valor de mercado e a importância da construção de uma rede de citações de patentes.

### 3.1 Patentes: Conceitos

Uma invenção é caracterizada como o resultado de uma nova solução para um problema técnico, podendo se utilizar meios completamente inexistentes na literatura, ou mesmo, a combinação de ideias nunca antes aplicadas para gerar uma nova solução para um problema em averiguação. Dessa forma, com o intuito principal de proteção aos processos tecnológicos de uma invenção, surgem diversas formas de proteção ao direito de propriedade intelectual (*DPI*), destacando-se, assim, os documentos em forma de patentes.

Uma patente, ou carta-patente, é um documento sobre uma propriedade de criação concedida pelo Estado aos autores, onde se encontra o conteúdo detalhadamente relatado sobre a invenção a ser protegida. Por meio disso, somente os titulares ou os procuradores possuem direitos exclusivos sobre a ideia concebida, monopolizando, assim, todos os processos relacionados ao invento, tais como a fabricação, a utilização e a comercialização em âmbito nacional ou local de registro da patente.

Mesmo com os direitos de invenção concedidos pelo Estado em mãos, cabe ao dono da patente a responsabilidade de averiguar a existência da quebra de direitos da patente e entrar com os passos iniciais de processo judicial.

Porém, mesmo não ocorrendo à expiração da data de licenciamento do documento, existem algumas exceções em que o Estado ou o escritório de patentes autoriza o uso da

invenção a terceiros através do regime chamado de licenciamento compulsório. Conforme a Convenção de Paris e *ADPIC*, esse regime impede os abusos comerciais que podem resultar da aquisição dos direitos exclusivos conferidos a uma patente, como também, a não utilização de uma invenção patenteada [7]. Outra exemplificação para o licenciamento compulsório seria o uso de patentes na área da saúde, em que o Estado permite o uso da invenção de uma determinada patente para a produção de medicamentos ou necessidades de impacto social.

Além disso, pode ocorrer da invenção patenteada ser considerada essencial para a continuidade no desenvolvimento de outras novas tecnologias, podendo ser essa patente enquadrada como um novo padrão industrial. Porém, diferentemente do licenciamento compulsório, o padrão industrial proporciona retornos financeiros por meio de acordos entre o titular da patente e o grupo industrial interessado na invenção, conforme será abordado mais adiante no Capítulo 3.5.

Assim, uma patente pode ser resumidamente definida como uma das formas de propriedade intelectual (*PI*) para se reconhecer a inovação e, principalmente, proteger os interesses dos seus inventores, garantindo ao seu titular os direitos exclusivos para a utilização de sua invenção por um período limitado.

A partir do momento em que o registro de patente é adquirido mediante a uma repartição governamental responsável pelos trâmites de *DPI* para a concessão de patentes (normalmente um Escritório de Patentes), cria-se um documento de domínio público contendo, de forma detalhada, as informações sobre a invenção a ser protegida. Dessa maneira, nesse documento apresentam-se informações como o conteúdo da patente, os dados dos inventores, as citações de patentes ou de outros materiais anteriormente publicados de aspecto relevante para a invenção e, também caso seja necessário, os arquivos anexos com a descrição do invento em forma de ilustrações [25].

As patentes são de domínio público, podendo ser utilizadas tanto para uma simples consulta de informações ou invenções concedidas pelo núcleo de inovação de um país ou, até mesmo, para fins de pesquisa por empresas ou universidades.

As leis para o gerenciamento dos direitos da propriedade intelectual são exclusivas de cada país ou região [25], podendo variar conforme aspectos culturais e organizacionais. Dessa forma, uma patente pode ser de âmbito nacional, como no caso do *INPI* e *USPTO*; regional, como no caso do *esp@cenet* que abrange direitos de patentes sobre vários países europeus; ou de tratado cooperativo (*PCT – Patent Cooperation Treaty*) [37].

Considerando-se os diversos escritórios de gerenciamento de patentes existentes no mundo, atualmente nos *EUA*, um dos primeiros países a estabelecerem leis mais precisas sobre a proteção de *DPI* em forma de patentes, existem mais de cinco milhões de patentes cadastradas [29], gerando no total cerca de 100 a 200 *gigabytes* de arquivos de texto. Somando-se essa quantia com aproximadamente 40 milhões de páginas de arquivos *bitmap*,

obtém-se mais de quatro *terabytes* de dados.

Geralmente, uma patente (como é mostrada simplificada na Figura 3.1), possui em torno de poucos *kilobytes* a 1,5 *megabytes*, podendo ser dividida em dois níveis básicos para uma melhor estruturação. No primeiro nível encontram-se as informações principais do documento em questão, como, por exemplo, o número da aplicação, o número da patente, dados da aplicação, dados da emissão e número de figuras. Já o último nível contém as informações específicas do conteúdo da invenção e informações a respeito dos autores, isto é, nome completo e endereço dos inventores, procurador, examinador e representante da patente. Além disso, nesse nível é onde se encontram os campos de texto que definem a patente, como por exemplo, o título, o resumo e as descrições detalhadas do documento.

<b>United States Patent</b>	<b>7,423,143</b>		
<b>McGall, et al.</b>	<b>September 9, 2008</b>		
Nucleic acid labeling compounds			
<b>Abstract</b>			
Nucleic acid labeling compounds containing heterocyclic derivatives are disclosed. The heterocyclic derivative containing compounds are synthesized by condensing a heterocyclic derivative with a cyclic group (e.g. a ribofuranose derivative). The labeling compounds are suitable for enzymatic attachment to a nucleic acid, either terminally or internally, to provide a mechanism of nucleic acid detection.			
Inventors:	<b>McGall; Glenn H.</b> (Palo Alto, CA), <b>Barone; Anthony D.</b> (San Jose, CA)		
Assignee:	<b>Affymetrix, Inc.</b> (Santa Clara, CA)		
Appl. No.:	<b>11/125,338</b>		
Filed:	<b>May 10, 2005</b>		
<b>Related U.S. Patent Documents</b>			
<u>Application Number</u>	<u>Filing Date</u>	<u>Patent Number</u>	<u>Issue Date</u>
09952387	Sep., 2001	6965020	
09780574	Feb., 2001	6596856	
09126645	Jul., 1998		
<b>Current U.S. Class:</b>	<b>536/26.6</b> ; 435/6; 536/23.1; 536/25.1; 536/25.3		
<b>Current International Class:</b>	C07H 21/04 (20060101); C07H 21/02 (20060101); C12Q 1/68 (20060101); C07H 21/00 (20060101)		
<b>Field of Search:</b>	435/6 536/23.1,25.1,26.6,25.3		
<b>References Cited [Referenced By]</b>			
<b>U.S. Patent Documents</b>			
<a href="#">3352849</a>	November 1967	Shen et al.	
<a href="#">3817837</a>	June 1974	Rubenstein et al.	
<a href="#">3850752</a>	November 1974	Schuurs et al.	
<a href="#">3891623</a>	June 1975	Vorbruggen et al.	
<b>Other References</b>			
Ramzaeva, N. , et al., "Oligonucleotides Functionalized by Fluorescein and Rhodamine Dyes: Michael Addition of Methyl Acrylate to 2'-Deoxypseudouridine", Helvetica Chimica Acta, 83 (2000), 1108-1126. cited by other .			
Akita, Y. , et al., "Cross-Coupling Reaction of Chloropyrazines with Acetylenes", Chemical & Pharmaceutical Bulletin, 34 (4), (Apr. 1986),pp. 1447-1458. cited by other .			
Aoyagi, M. , et al., "Nucleosides and Nucleotides. 115. Synthesis of 3-Alkyl-3-Deazainosines via Palladium-Catalyzed Intramolecular Cyclization: A New Conformational Lock with the			
<b>Claims</b>			
The invention claimed is:			
1. A nucleic acid labeling compound of the following structure: ##STR00074## wherein A is H, monophosphate, diphosphate, triphosphate, phosphoramidite ((R.sub.2N)(R'O)P) wherein R is linear, branched or cyclic alkyl and R' is a protecting group, or H-phosphonate (HP(O)O--HNR.sub.3), wherein R is linear, branched or cyclic alkyl; X is O, S, NR.sub.1 or CHR.sub.2, wherein R.sub.1 and R.sub.2 are, independently, H, alkyl or aryl; Y is H, N.sub.3, F, OR.sub.9, SR.sub.9 or NHR.sub.9, wherein R.sub.9 is H, alkyl or aryl; Z is H, N.sub.3, F or OR.sub.10, wherein R.sub.10 is H, alkyl or aryl; L is an amido alkyl group; Q is a detectable moiety; and, M is a connecting group, and wherein n is an integer ranging from 0 to about 3.			
2. The nucleic acid labeling compound of claim 1, wherein A is H or H.sub.4O.sub.9P.sub.1.sup.- with an appropriate counterion.			
3. The nucleic acid labeling compound of claim 1, wherein X is O.			
<b>Description</b>			
BACKGROUND OF THE INVENTION			
Gene expression in diseased and healthy individuals is oftentimes different and characterizable. The ability to monitor gene expression in such cases provides medical professionals with a powerful diagnostic tool. This form of diagnosis is especially important in the area of oncology, where it is thought that the overexpression of an oncogene, or the underexpression of a tumor suppressor gene, results in tumorigenesis. See Mikkelsen et al. J. Cell. Biochem. 1991, 46, 3-8.			
* * * * *			

Figura 3.1: Exemplo simplificado de uma patente do *USPTO*

Com a finalidade de um tratamento mais detalhado para a enorme quantia de pedidos de patentes depositados diariamente, as leis de regulamentação de *DPI* norte-americanas distribuem as patentes conforme três grupos distintos dependendo do tipo de invenção a ser patenteada: *Design*, Plantas ou Utilidade.

Normalmente, o período de vigência de uma patente é em média de 20 anos, possuindo como marco inicial a sua data de depósito no escritório de patentes, sendo sujeita a pagamentos de taxas de concessão e manutenção ao órgão expedidor do documento [25]. Ao se extinguirem os direitos patentários, a técnica contida no documento cai em domínio público, ficando aberta para utilização em benefício próprio de outras pessoas físicas ou jurídicas sem o pagamento de quaisquer ônus ao titular da patente. Assim, uma patente pode ser especificamente considerada como uma espécie de acordo entre o público e detentor dos direitos patentários [7].

Deve-se notar ainda que existem certos requisitos fundamentais para que a invenção alcance a sua patenteabilidade no escritório de patentes do cada país onde será feito o pedido de patenteamento. Segundo o *ADPIC*, para um pedido de aquisição de patente, a invenção basicamente deve seguir os critérios [7] de:

- **Novidade:** a invenção deve ser nova para o seu campo científico, sendo nunca usada e realizada anteriormente. Dependendo das regras de cada centro de inovação, necessita-se também que a invenção nunca tenha sido revelada ao público anteriormente à data de pedido da patente.
- **Atividade Inventiva:** a invenção deve possuir um estágio de desenvolvimento suficiente para que seja patenteável. Além disso, a invenção deve ser não óbvia, isto é, não trivial - o invento deve possuir um nível substancial de inventividade ao ponto de que uma pessoa, com conhecimentos técnicos no assunto, possa reproduzir fielmente o invento descrito no documento.
- **Aplicação Industrial ou Utilidade:** a invenção deve ser passível de utilização na prática, devendo obrigatoriamente alcançar resultados aceitáveis e úteis, incluindo um potencial valor comercial [14]. Devido ao fato desse critério ser bastante amplo, pois, quase tudo desenvolvido pelo indivíduo é passível de utilização, cabem a cada escritório de patentes a avaliação e a responsabilidade da decisão de aprovação para o invento.

Além dos critérios anteriormente mencionados, para o *USPTO*, acrescenta-se também o critério de *capacitação*, ou seja, o inventor deve revelar ao público a sua invenção e a sua melhor forma de utilização e criação do invento [25].

Porém, mesmo obedecendo todos os requisitos descritos, nem todos os objetos são passíveis de patenteamento, sendo excluídos do escopo de patenteabilidade. Exemplos

disso seriam o genoma humano, os seres vivos e os materiais já existentes na natureza. Excluem-se também máquinas que desafiam as leis da natureza, salvo o fato de comprovar o êxito de sua funcionalidade na prática. Também não são possíveis de patenteamento as fórmulas matemáticas, as teorias científicas, as ideias abstratas, os métodos, os esquemas e todas as invenções que possam apresentar risco a segurança, a saúde ou questões étnicas e morais para a nação [7].

## 3.2 Sistema de Patentes

O modelo do vigente sistema de patentes ou escritório de patentes teve seu desenvolvimento durante vários séculos, variando conforme aspectos culturais e organizacionais de cada nação [7]. Devido a esse fato, ainda não se possui um escritório internacional para o depósito de patentes, porém, vários grupos têm realizados esforços para a implantação de um sistema internacional para procurar contornar essa situação [8].

Cada país possui um sistema de patentes responsável pelo pedido de aquisição de patentes. No caso dos *EUA*, o *USPTO* seria o órgão responsável pelos trâmites de registro da patente, enquanto que no Brasil, essa atividade ficaria a cargo do Instituto Nacional da Propriedade Industrial (*INPI*)<sup>1</sup>.

Vinculada ao Ministério do Desenvolvimento, Indústria e Comércio Exterior (*MDIC*), o *INPI* foi criado em 1970, sendo responsável pelo aprimoramento, disseminação e gestão do sistema brasileiro de concessão de direitos de propriedade intelectual para a indústria. O órgão é responsável por registros de marcas, desenhos industriais, indicações geográficas, programas de computador, topografias de circuitos, concessão de patentes, contratos de franquia e transferência de tecnologia.

Além disso, existem os sistemas regionais de patentes a fim de se obter simultaneamente vários pedidos de patentes de uma mesma invenção em um conjunto de nações abrangidas pelo escritório regional sem a necessidade da realização de pedidos individuais para cada um dos países pertencentes ao sistema. Para isso, o detentor da patente deve iniciar os trâmites de pedido obedecendo às normas da legislação do sistema de patentes regional. Um exemplo de escritório regional seria o *esp@cenet* que possibilita o registro da patente em todos os países europeus-membros a partir de um único pedido de patentes.

---

<sup>1</sup>Instituto Nacional da Propriedade Industrial - INPI. Disponível em: <http://www.inpi.gov.br/portal/>



### 3.3 Sistema de Patentes Americano: USPTO

Com a crescente demanda por aquisição de patentes, até o ano de 2007, segundo uma pesquisa realizada por *Kasrav e Risov* [25], a base de dados do *USPTO*<sup>2</sup> possuía mais de oito milhões de patentes concebidas e pendentes para validação. Calcula-se que atualmente, em média, mais de 120000 pedidos de patentes são realizados anualmente apenas no escritório de patentes americano. Dentre todas essas requisições, estima-se que aproximadamente 15% dos pedidos são relacionadas a *software* [11]. Esse fato tem se intensificado e transcorrido desde o final da década de 90 até os dias atuais, devido, principalmente, a liberação da possibilidade de patenteamento de programas computacionais pela constituição americana [27].

Segundo as leis e as normas dos Estados Unidos, conforme consta no Artigo I, Seção 8, Cláusula 8, pode-se definir a essência da *PI* da seguinte forma: “*Para promover o progresso da ciência e artes, assegura-se por um tempo limitado aos autores e inventores os exclusivos direitos sobre suas respectivas escritas e descobertas*”.

Por meio disso, ficam extremamente claras as preocupações do governo americano em instigar os avanços nas áreas científicas e tecnológicas. Pois, por meio de estímulos aos autores da invenção, através de remunerações, faz com que também se promova uma expansão da economia, já que quanto maior o apoio ao desenvolvimento da *C&T&I* maiores serão os benefícios em pró ao crescimento científico e tecnológico da nação.

Cada país, ou mesmo uma determinada região, possui um órgão responsável pela regulamentação das leis para a obtenção dos direitos vinculados a propriedade intelectual de uma invenção. Nos *EUA*, país considerado um dos pioneiros a desenvolver e implantar um sistema de patentes estruturado, o órgão governamental para a implementação e regulamentação das leis de aquisição de patentes é o *USPTO*. A entidade se apoia no *MPEP (Manual of Patent Examination and Prosecution)* e nos *37 CFR (Code of Federal Regulations)* para prover as regras e processos para a operação do sistema de patentes conforme a legislação federal do país.

O *MPEP* é um manual publicado para prover aos examinadores de patentes do *USPTO*, aos advogados, aos agentes e aos representantes das aplicações, uma referência de trabalho para as práticas e processos de patentes. O *MPEP* contém instruções e outros materiais de informação, resumindo os processos que os examinadores de patentes são outorgados a seguir em determinados casos no exame de patentes.

Já o *CRF* é a codificação das regras gerais e permanentes publicados no registro federal pelo departamento executivo e agências do governo federal. É dividido em 50 *títulos* que representam vastas áreas sujeitas ao regulamento federal<sup>3</sup>.

---

<sup>2</sup>United States Patent and Trademark Office - USPTO. Disponível em: <http://www.uspto.gov/>

<sup>3</sup>US Government Printing Office. Disponível em: <http://www.gpo.gov/>

A Figura 3.2 mostra a página de pesquisas avançadas do *USPTO* que possibilita o usuário digitar a *query* para a busca das patentes desejadas na base de dados do sistema.

**USPTO PATENT FULL-TEXT AND IMAGE DATABASE**

[Home](#)   [Quick](#)   [Advanced](#)   [Pat Num](#)   [Help](#)  
[View Cart](#)

**Data current through August 3, 2010.**

Query [\[Help\]](#)

Examples:  
**ttl/(tennis and (racquet or racket))**  
**isd/1/8/2002 and motorcycle**  
**in/newmar-julie**

Select Years [\[Help\]](#)

1976 to present [full-text]

Patents from 1790 through 1975 are searchable only by Issue Date, Patent Number, and Current US Classification.  
 When searching for specific numbers in the Patent Number field, patent numbers must be seven characters in length, excluding commas, which are optional.

Field Code	Field Name	Field Code	Field Name
PN	<a href="#">Patent Number</a>	IN	<a href="#">Inventor Name</a>
ISD	<a href="#">Issue Date</a>	IC	<a href="#">Inventor City</a>
TTL	<a href="#">Title</a>	IS	<a href="#">Inventor State</a>
ABST	<a href="#">Abstract</a>	ICN	<a href="#">Inventor Country</a>
ACLM	<a href="#">Claim(s)</a>	LREP	<a href="#">Attorney or Agent</a>
SPEC	<a href="#">Description/Specification</a>	AN	<a href="#">Assignee Name</a>
CCL	<a href="#">Current US Classification</a>	AC	<a href="#">Assignee City</a>
ICL	<a href="#">International Classification</a>	AS	<a href="#">Assignee State</a>
APN	<a href="#">Application Serial Number</a>	ACN	<a href="#">Assignee Country</a>
APD	<a href="#">Application Date</a>	EXP	<a href="#">Primary Examiner</a>
PARN	<a href="#">Parent Case Information</a>	EXA	<a href="#">Assistant Examiner</a>
RLAP	<a href="#">Related US App. Data</a>	REF	<a href="#">Referenced By</a>
REIS	<a href="#">Reissue Data</a>	FREF	<a href="#">Foreign References</a>
PRIR	<a href="#">Foreign Priority</a>	OREF	<a href="#">Other References</a>
PCT	<a href="#">PCT Information</a>	GOVT	<a href="#">Government Interest</a>
APT	<a href="#">Application Type</a>		

Figura 3.2: Página de buscas avançadas do *USPTO*

A imagem seguinte (Figura 3.3) apresenta uma lista de patentes com os seus respectivos números para uma busca realizada na base de dados do *USPTO*. Para a exemplificação utilizou-se da *query* *TTL/(fruit) AND ABST/(DNA)*, isto é, deseja-se obter todas as patentes que apresentem a palavra *fruit* (fruto) no campo *title* e *DNA* no campo *abstract* das patentes.

**USPTO PATENT FULL-TEXT AND IMAGE DATABASE**

[Home](#)
[Quick](#)
[Advanced](#)
[Pat Num](#)
[Help](#)

[Bottom](#)
[View Cart](#)

Searching US Patent Collection...

**Results of Search in US Patent Collection db for:**  
**(TTL/fruit AND ABST/dna): 19 patents.**  
*Hits 1 through 19 out of 19*

Jump To

Refine Search

PAT. NO.	Title
1 7,202,355	<a href="#">DNA sequence regulating plant fruit-specific expression</a>
2 7,084,321	<a href="#">Isolated nucleic acid molecules relating to papaya fruit ripening</a>
3 7,071,377	<a href="#">Method to control the ripening of papaya fruit and confer disease resistance to papaya plants</a>
4 6,483,012	<a href="#">Methods for producing parthenocarpic or female sterile transgenic plants and methods for enhancing fruit setting and development</a>
5 6,284,946	<a href="#">Banana DNA associated with fruit development</a>
6 6,271,033	<a href="#">Method for modifying production of fruit ripening enzyme</a>
7 6,268,552	<a href="#">Transgenic seedless fruit comprising AGL or GH3 promoter operably linked to isopentenyl transferase or tryptophan monooxygenase coding DNA</a>
8 6,107,548	<a href="#">DNA sequences from muskmelon (Cucumis melo) related to fruit ripening</a>
9 6,043,410	<a href="#">Strawberry fruit promoters for gene expression</a>
10 6,031,154	<a href="#">Fructokinase genes and their use in metabolic engineering of fruit sweetness</a>
11 6,011,199	<a href="#">Method for producing fruiting plants with improved fruit flavour</a>
12 5,908,973	<a href="#">DNA encoding fruit-ripening-related proteins, DNA constructs, cells and plants derived therefrom</a>
13 5,859,344	<a href="#">Modified fruit containing galactanase transgene</a>
14 5,859,330	<a href="#">Regulated expression of heterologous genes in plants and transgenic fruit with a modified ripening phenotype</a>
15 5,702,933	<a href="#">Control of fruit ripening and senescence in plants</a>
16 5,608,150	<a href="#">Fruit specific promoters</a>
17 5,545,815	<a href="#">Control of fruit ripening in plants</a>
18 5,512,466	<a href="#">Control of fruit ripening and senescence in plants</a>
19 5,304,490	<a href="#">DNA constructs containing fruit-ripening genes</a>

Figura 3.3: Resultados de busca de patentes para a *query TTL/(fruit) AND ABST/(DNA)*

### 3.4 Procedimentos para a Obtenção de uma Patente

O procedimento completo para a obtenção de uma patente é um pouco complexo e pode tramitar de no mínimo dois anos, perpetuando-se até mesmo por vários anos.

Basicamente, ao se definir a ideia a ser desenvolvida, procura-se concretizar essa representação à utilidade. Esse processo é conhecido no campo do *DPI* como *reduzindo a ideia à prática (reducing the idea to practice)*. Hoje em dia esse princípio não se torna um requisito necessário para se obter o patenteamento de determinado invento. Porém,

até o ano de 1998, o princípio de reduzir a ideia à prática era extremamente obrigatório para a obtenção de patenteamento, sendo modificado apenas após a decisão ocorrida no caso *Pfaff vs. Wells Electronics*. Nesse ato judicial, a corte suprema dos Estados Unidos observou e comparou que uma patente pode ser obtida sem ser reduzida à prática, pois, tendo como base o caso da requisição de patenteamento do telefone por *Graham Bell*, esse foi concebido sem mesmo possuir um modelo funcional para o invento [11].

Assim, para a obtenção de uma patente no *USPTO*, se aceita de uma construtiva redução à prática, isto é, uma documentação bem descrita sobre o invento, podendo o mesmo possuir desenhos para facilitar o entendimento de uma pessoa não leiga no assunto a fim de reconstruir a invenção descrita.

Feito isso, o inventor pode dar entrada ao processo de depósito da patente. Assim, o escritório de patentes pode dar início ao exame do documento a fim de verificar se todos os critérios de patenteamento foram cumpridos conforme a legislação local de depósito. Por fim, o sistema de patentes decide pela concessão ou não da patente requisitada.

O início da data de validade da patente começa a contar a partir da entrada de pedido do documento, podendo esse período variar conforme as legislações do órgão expedidor onde a patente está sendo depositada. Normalmente, a validade é em média de 20 anos, como no caso do *INPI* e *USPTO*, existindo a possibilidade de prorrogação antes do vencimento da patente. Outro fato a ser destacado é que, como regra geral, a concessão de uma patente obedece ao sistema de *primeiro a depositar (First to file)*, sendo essa a razão da importância da data do início do pedido de depósito presente no documento (*Filing Date*).

Estabelecida em 1883, a Convenção de Paris para a proteção da propriedade industrial é uma das legislações mais antigas para a proteção da propriedade industrial em que se estabelece o princípio de *primeiro a depositar* [8].

Sendo administrada pela *OMPI*, essa legislação provê o direito de prioridade a todos os seus Estados membros. E, dessa maneira, o solicitante do pedido de depósito pode também, no período de 12 meses, solicitar o pedido de proteção em quaisquer um dos outros Estados membros. Isto é, os demais pedidos de patentes feitos em outros Estados membros também são concedidos com a mesma data de pedido de depósito da primeira patente solicitada. Esse processo implica em uma prioridade do pedido de patente sobre outras solicitações que possam ser depositadas durante o mesmo período por outros solicitantes para uma mesma invenção [7]. No entanto, enfatiza-se que a aprovação de concessão depende exclusivamente do Estado membro onde se deseja depositar a patente.

Via de regra, o fator *primeiro a depositar* torna-se algo de grande importância nos dias atuais. Pois, em uma realidade empresarial extremamente competitiva, a diferença de poucos dias ou mesmo de algumas horas no pedido de concessão de patentes pode resultar em grandes perdas na aquisição de patentes importantes para uma determinada

empresa, uma vez que outras empresas rivais podem desenvolver a mesma invenção e realizar primeiramente o pedido de patente.

Além das etapas mencionadas, deve-se prestar atenção ao fato de que em certos países não se concedem direitos de patente às invenções publicamente reveladas antes da data de depósito. Nos Estados Unidos, a legislação permite um período de no máximo um ano para que a invenção possa ser revelada ao público ou posta à venda [11] antes da data de registro, essa fase é conhecida como *período de graça* (*Grace period*). Assim, os inventores podem optar por manterem secretas suas invenções ou disponíveis à venda antes da data de depósito da patente.

Geralmente uma patente se torna disponível ao meio público após 18 meses da data de depósito, recebendo finalmente uma data de emissão para a patente. Durante esse período, o *USPTO* investiga o estado da arte, realiza avaliações, comunica-se com o inventor, emite ações e garante ou rejeita o documento patentário. Enquanto isso, o inventor possui o direito de revelar por completo a sua invenção ao público, período chamado de *janela de predição* (*prediction window*) [25].

Após a data de emissão, o inventor ou o detentor da patente possui proteção total das leis videntes de propriedade intelectual no local onde foi concedida a patente.

Atualmente, como ainda não existe um sistema internacional de patentes, é impossível de se obter uma única patente mundial para a invenção. O sistema de patentes é um esquema territorial e, para a obtenção da proteção patentária em vários países, deve-se exclusivamente obedecer às normas para o depósito de patente em cada um dos países em se deseja obter o registro de patente [8].

No entanto, para facilitar esse processo de vários pedidos de patentes, existe um acordo administrado pela *OMPI* chamado de *Tratado de Cooperação em Matéria de patentes* (*PCT – Patent Cooperation Treaty*). Por meio deste, com apenas um único pedido internacional feito pelo titular da patente, resultam-se em vários pedidos nacionais de patente. Porém, as taxas de anuidade e manutenção dos serviços deverão ser quitadas individualmente para todos os países em seus respectivos escritórios de patentes onde a patente foi registrada. Caso contrário, o não pagamento implicará na perda da proteção patentária nos territórios onde não foram realizados os devidos pagamentos [7].

### 3.5 Problemas nas referências das Patentes

Com o início da disponibilidade de informações em meio eletrônico a partir da década de 70 e a facilidade de aquisição dos dados das patentes, instigaram-se grandes interesses em análises econométricas por instituições públicas e privadas. Mais especificamente no final dos anos 80 esse interesse adquiriu uma maior ascensão, pois, a partir de então, as informações de citações de patentes também começaram a serem disponibilizadas em

forma eletrônica. Desse modo, a análise econométrica poderia usar os dados de relacionamento entre patentes para melhorar a investigação sobre o conteúdo informativo desses documentos, além de procurar desvendar um conjunto adicional de questões relacionadas ao fluxo de conhecimento no tempo, espaço e limites organizacionais [14].

No entanto, devido à forma como os dados de patentes são coletados e armazenados na base de dados do *USPTO*, correlacionar diretamente patentes não se torna uma tarefa simplesmente trivial. A análise das informações deve ser realizada com muita cautela, pois, na realidade existem vários problemas substanciais nos dados de citações de patentes [14], possivelmente devido ao propósito com que os dados de patentes são inicialmente adicionados.

Algumas vezes, muitas das citações contidas nas patentes são adicionadas pelo advogado do inventor ou pelo examinador da patente, decorrendo, por exemplo, em citar documentos que são bastante importantes para uma área de invenção ou até mesmo patentes que possuem grande valor econômico. Nessa etapa, muitas das patentes podem apresentar citações de invenções que são completamente desconhecidas e não utilizadas para o desenvolvimento da tecnologia pelo inventor. Devido a isso, alguns erros podem ser cometidos nas referências da invenção, já que normalmente as empresas detentoras de patentes procuram visar interesses comerciais e econômicos para a referenciação de suas patentes.

Conforme mostra uma pesquisa realizada com mais de 150 proprietários de patentes [14], ao se questionar sobre o relacionamento de suas invenções com as citações mencionadas nos documentos, verificou-se que existem vários problemas no processo de citação patentária, ocasionando dessa forma, certas imperfeições no fluxo de conhecimento das patentes. Nessa investigação constatou-se que, dentre 10 citações realizadas, alguns inventores citam em média 2 patentes as quais não possuem nenhum tipo de relação com a invenção. No geral, apenas um quarto das citações corresponde a um correto fluxo de conhecimento com significado real para o invento.

Outro fato observável é o de que empresas detentoras de patentes possuem livremente o direito de mudar a sua identificação comercial durante o registro de patenteamento, uma vez que o *USPTO* não procura manter um identificador único para a mesma entidade patenteadora [14]. Muitas vezes essas variações de nomenclaturas ocorrem devido ao fato de que grandes empresas, normalmente multinacionais, possuem a propriedade de englobarem vastos setores distintos ou por razões estratégicas de mercado.

### 3.6 Valor Econômico de uma patente

Em um primeiro momento pode parecer não existir vantagens para o registro de uma patente, já que a invenção torna-se publicamente divulgada e acessível ao público. No



entanto, como mencionado no capítulo sobre conceito de uma patente (Capítulo 3.1), uma patente é uma forma de documento de propriedade intelectual com um prazo fixo de vigência que impede terceiros a produzir, utilizar, comercializar, importar e exportar a invenção descrita sem a devida autorização do titular da patente. Com isso, o titular possui todos os direitos exclusivos para a exploração da invenção em toda a extensão territorial abrangido pelo órgão ao qual foi obtido o registro da patente.

Um estudo de patentes na área econômica muito utilizada recentemente é a utilização de patentes para medir o nível quantitativo e qualitativo de invenções geradas por organizações ou unidades geográficas, ou seja, medir a importância ou o impacto tecnológico de certo campo inovativo de uma entidade ou nação. Dessa forma, pode-se afirmar que as patentes são como uma espécie de medidores do nível inventivo e as suas citações são um meio para medir o fluxo ou os impactos do conhecimento de um país ou firma [14]. Assim, as invenções são o núcleo criador de novos produtos e serviços, além de serem considerados medidores do crescimento econômico.

Atualmente as atividades de patenteamento de produtos e novas tecnologias vêm se comportando não apenas como um meio de proteção intelectual, mas também como uma forma valorizada de atividade mercantil entre empresas. Esse comportamento se deve a inúmeras razões, sendo uma delas a de assegurar a produção exclusiva de um determinado produto no mercado. Também, possibilita obter ganhos financeiros com vendas de direitos de produção da tecnologia para outras firmas e, principalmente, servir como ponto de referência para a produtividade da equipe de pesquisa e desenvolvimento da empresa [14].

Por meio da aquisição de patentes, os países capitalistas industrializados têm observado um acelerado aumento em sua produtividade nessas últimas duas décadas comparado a todos os milênios desde os primórdios da história da humanidade. O fundamento principal para esses avanços se deve exclusivamente à busca de lucros financeiros, forçando as empresas a gerarem inovações tecnológicas a fim de manterem sua competitividade perante outras corporações [22].

Porém, com a intensificação de um mercado cada vez mais globalizado, juntamente com as terceirizações de serviços e disputas comerciais acirradas, a empresa deve estar sempre atenta a inúmeras violações de patentes decorrentes dessa transformação econômica [28]. Cabe então a ela ou ao titular da patente tomar a iniciativa em cada um dos países de registro da patente caso ocorra à infração dos direitos patentários, assim como, as atitudes sobre a detecção de violação dos direitos e a repreensão de infratores. Geralmente, o procedimento a ser tomado pelo titular da patente é o envio de uma notificação alertando o infrator sobre a existência de uma patente para a invenção. Muitas vezes essa ação acaba resultando em uma cessação da infração ou mesmo a um contrato de licenciamento.

Não obstante, nem sempre se torna possível uma solução amigável entre ambas as partes (titular e infrator da patente), resultando em casos nos quais se necessita a intervenção

da jurisdição nacional. Um exemplo disso seria o caso da *Union Carbide vs. Shell* [28]. Esse processo envolveu discussões na suprema corte federal dos Estados Unidos (*CAFC - Court of Appeals for the Federal Circuit*) sobre a quebra de direitos patentários de um método de produção de Óxido de Etileno pertencente à *Union Carbide* pela multinacional *Shell* fora dos limites territoriais dos *EUA*.

Outra exemplificação de quebra de patente seria o caso da *Albie's Foods vs Smucker Co* [22], em que a empresa *Sucker Co* acusa a *Albie's Foods* pela violação de sua patente sobre um tipo especial de sanduíche (Patente *USPTO* de N° 6004596).

O valor de mercado de uma patente, diferentemente do valor científico, é determinado pela importância que a patente em questão possui para o desenvolvimento de um produto a ser comercializado [48].

Ultimamente, com a formação de conglomerações mercantis, a utilização de tecnologias patenteadas em padrões industriais têm se tornado um fator essencial para o meio corporativo. O termo padrão industrial refere-se a um conjunto de tecnologias adotadas por empresas com a finalidade de prover a compatibilidade entre produtos.

Com isso, organizações atuantes para o processo de padronização da tecnologia têm desenvolvido políticas e regulamentações para o uso de documentos patentários, requerendo que seus membros revelem a propriedade intelectual que possa supostamente impactar em um padrão já proposto. Pois, caso a patente seja extremamente essencial para a geração e desenvolvimento de uma tecnologia, o seu veto à comercialização e a monopolização podem acarretar em trazer ganhos indiretos a uma determinada empresa devido à cessação de outros competidores de ocupar o mesmo espaço de mercado [25].

Ao se definir um padrão industrial para uma tecnologia emergente, provendo privilégios a determinadas patentes, resultam-se imediatamente em mudanças no valor de mercado das corporações [9]. Assim, dois tipos de patentes são afetados por esse processo de padronização: as patentes básicas e as patentes de desenvolvimento. As patentes básicas descrevem uma inovação tecnológica fundamental para a geração de outras tecnologias, enquanto que as patentes de desenvolvimento buscam implementar uma solução para um problema específico.

Dessa forma, a decisão de incorporar uma patente a um padrão industrial deve ser tomada com extrema cautela. Pois, escolhendo-se uma patente de desenvolvimento como padrão industrial, esta ação poderá resultar em uma acentuada valorização de mercado para a determinada patente e, conseqüentemente, surtirá em decréscimos no valor de mercado para as outras patentes com soluções semelhantes, podendo inclusive torná-las sem valor comercial. Por outro lado, caso a organização opte pela padronização de uma patente do tipo básica, a patente se torna extremamente valorizada, acarretando em uma participação de 100% de mercado pelo proprietário da patente.

Dependendo das legislações adotadas por cada organização padronizadora, o detentor



dos direitos da patente adotada como padrão pode coletar *royalties* dos membros usuários dessa patente. *Royalty*, em outras palavras, é uma forma de pagamento pela licença a um terceiro para explorar algo patenteado pelo licenciador.

Para usufruir desses benefícios, o valor arrecadado pelo licenciador da patente dever ser justo e razoável, sendo oferecido a todos os usuários do padrão de forma não discriminatória. No entanto, existem várias maneiras de interpretação do termo não discriminatório para o mercado corporativo, podendo variar conforme as regulamentações das organizações padronizadoras. Uma exemplificação para isso seria os valores de *royalty* cobrados por organizações: em uma determinada instituição, adota-se o critério de oferecer valores idênticos a todos os licenciados de uma patente sem distinção; já em outra corporação, estabelece-se tratamento semelhante a apenas grupos similarmente situados no mercado. Para a corte judicial dos Estados Unidos, ser não discriminatório não requer prover licenças idênticas a todos os licenciados, mas, fornecer o mesmo tratamento para licenciados semelhantes [9].

Muitas vezes o valor de mercado de uma patente possui algum tipo de ligação direta com os processos de padronização de uma patente, podendo esse fato prover incentivos para que as empresas pressionem as organizações padronizadoras a incorporarem suas patentes como padrões industriais e, dessa forma, receberem destaque nessas patentes (citações de patentes) para o desenvolvimento de novas tecnologias por outras empresas. Segundo *Kasravi e Risov* [25], uma citação por patente de uma empresa resulta em 3% de acréscimo no seu valor total de mercado no meio corporativo.

Estimativas também mostram que empresas que possuem patentes muito citadas, com mais que 20 citações por patente, apresentam grandes diferenças no valor de mercado. Ao se comparar uma empresa com um grande índice de citação de patentes a outra firma que possui mesmo índice de *P&D* e ações de patentes, porém, com um índice de citação de patentes médio (em torno de 10 citações por patente), o fato da empresa possuir patentes com muitas citações acarreta em um aumento de 50% a mais no seu valor de mercado. Então, devido a esse fator, percebe-se que cerca de um quarto das citações recebidas por patentes corporativas vêm de outras patentes pertencentes ou associadas à mesma companhia [14].

Um fenômeno bastante comum ao se observar o histórico das invenções patenteadas é o fato de que as patentes básicas e as patentes mais inventivas, ou seja, as patentes que se utilizam de uma idéia essencial já patenteada para a criação de uma nova tecnologia de impacto mercantil, acumulam um grande número de citações recebidas de outras patentes ou outro meio literário [25]. Conforme mostram pesquisas nas áreas de citações de patentes, se certa patente, mesmo com o passar do tempo desde o seu registro, continua a receber citações de outras novas patentes, esse fato pode significar que a patente possui grande valor inovativo [48].

As citações providas de outras patentes indicam um status de reconhecimento da ideia pela comunidade inventiva, fornecendo também informações diretas dos excedentes de impacto e conhecimento de uma tecnologia.

### 3.7 Tendências de Mercado

Nas duas últimas décadas tem-se observado uma ascensão na aquisição do número de patentes por parte de empresas e instituições de desenvolvimento tecnológico, fato que vem gerando várias pesquisas científicas a fim de entender esse novo comportamento. *Kasravi e Risov* [25] relatam esse fenômeno através de uma pesquisa sobre o número de artigos em jornais que discutem patentes na indústria da tecnologia da informação entre os anos de 1997 a 2004.

Além de descrever e proteger a invenção, patentes incluem várias informações diretamente ou indiretamente dispostas sobre o detentor da patente. Assim, por meio de diversas análises, torna-se possível realizar prospecções mercadológicas para o meio corporativo como, por exemplo, justificar investimentos realizados, prever a direção do desenvolvimento da tecnologia e avaliar empreendimentos individuais das companhias [25]. Essas diretrizes podem, assim, ajudar a liderar uma melhor decisão de negócios a serem tomadas pelas empresas.

Examinando-se os documentos patentários recentemente concedidos em um órgão expedidor de patentes, torna-se possível descobrir os mais novos avanços tecnológicos ainda não registrados pelo mercado. Uma revisão um pouco mais detalhada desses dados pode apontar a direção geral de *P&D* das indústrias e, possivelmente, os produtos e serviços que serão futuramente introduzidos no mercado, além de ajudar a identificar forças, fraquezas e oportunidades de negócios das corporações [25].

Outra forma para verificar o estado da arte e as mudanças tecnológicas do mercado pelas corporações seria por meio da análise do fluxo de conhecimento de uma rede de patentes. Segundo estudos realizados, uma rede de citações de patentes pode ser caracterizada como um fenômeno de *small world*, isto é, uma rede com um alto nível de *clustering* com menor comprimento de caminho (*short path length*) entre dois nós [18]. Completada essas exigências, torna-se possível, partindo-se de uma determinada patente, obter informações de outras patentes correlacionadas por meio de poucas intermediações, além de possibilitar a visualização de patentes de produtos derivados da patente em investigação.

## 3.8 Redes de Citações de Patentes

Devido às patentes proverem informações sobre os níveis de tecnologia em um determinado setor e a intenção comercial de um competidor em potencial, as redes de citações de patentes servem como um estanque inicial para o desenvolvimento do plano estratégico de uma corporação. Além disso, permitem medir o *acúmulo de tecnologia* e comprovar fatos econômicos de desenvolvimento de determinados países [18].

Em uma pesquisa realizada sobre a tecnologia *TFT-LCD* para a análise do fenômeno *Small World* [18], resultados mostraram que por meio da análise dessa rede de citações, podem-se comprovar certos acontecimentos históricos de mercado através da observação de certos padrões do fluxo de citações. No caso da tecnologia *TFT-LCD*, os resultados do exame da rede de citações mostraram claras evidências de como países de primeiro mundo como Japão e *EUA* proporcionaram um papel importante no desenvolvimento econômico de países considerados *Tigres Asiáticos* como a Coreia do Sul e Taiwan.

Também se observa que em uma rede de citações poucos documentos patentários possuem um alto grau de citações e a maioria apresenta um nível baixo de citações. Esse fato indica que os documentos que recebem muitas citações apresentam uma informação considerada como essencial para a continuidade do processo de inovação tecnológica, podendo também significar a existência de muitos esforços de pesquisa investidos em relação a essa tecnologia em questão [48] [6].



# Capítulo 4

## Similaridade

Similaridade é um conceito bastante complexo que tem sido estudado entre as várias áreas do conhecimento. Para a psicologia, similaridade é uma espécie de relacionamento entre dois objetos perceptuais, sendo determinados por uma reação psicológica do indivíduo. Dessa forma, pode-se dizer que o grau de semelhança entre dois objetos depende exclusivamente de suas características comuns e de suas diferenças [51].

### 4.1 Similaridade de Texto

Desde muito tempo, técnicas para determinar a similaridade de texto têm sido estudadas e utilizadas em diversas aplicações computacionais, tais como a recuperação de informação e o processamento de linguagem natural. Basicamente, os métodos para se calcular a similaridade entre documentos textuais podem ser classificados em quatro categorias: similaridade textual léxica (*Text-based lexical similarity*), similaridade textual semântica (*Text-based semantical similarity*), métodos híbridos (*Hybrid methods*) e métodos baseados em características (*Feature-based methods*) [21].

- Métodos de similaridade textual léxica: são métodos que resultam em um valor numérico indicando o grau de semelhança entre os textos analisados. Para a realização do cálculo de similaridade, utiliza-se do número de unidades léxicas que os textos avaliados possuem em comum [19]. Além disso, com o intuito de se obter melhorias para esses métodos, várias adaptações algorítmicas foram realizadas como, por exemplo, o *Stemming*, a remoção de palavras irrelevantes (*Stop words*), o *Longest subsequence matching* e o uso de vários fatores de normalização e peso [38]. Os métodos de similaridade textual léxica podem ser subdivididos em duas categorias: métodos de coocorrência de palavras e métodos baseados em corpus. Os métodos de coocorrência de palavras (*Word co-occurrence method*), também conhecidos como

métodos de modelo de documento baseados em vetor (*Vector-based document model method*), são os métodos mais comuns aplicados no campo de similaridade de textos e baseiam-se na proposição de que documentos semelhantes possuem muitas palavras em comum em seus corpora. Para a representação dos textos ou de seus segmentos, os métodos dessa categoria utilizam uma estrutura vetorial de palavras, determinando-se o grau de semelhança entre os documentos por meio da utilização de uma métrica de similaridade. Por último, diferentemente dos métodos de co-ocorrência de palavras, os métodos baseados em corpus (*Corpus-based similarity methods*) levam em conta o grau de similaridade entre palavras necessariamente usando informações derivadas de grandes corpora textuais, para isso, empregam-se métricas específicas para avaliação de tais estruturas textuais.

- Métodos de similaridade textual semântica: tomando-se uma vertente diferentemente dos métodos de similaridade textual léxica, os algoritmos dessa classe possuem a vantagem de procurar identificar a similaridade semântica entre os textos. Um exemplo de aplicação seria a identificação do conteúdo semântico para as frases que buscam expressar ideias semelhantes como, por exemplo, as sentenças *Eu estudo na universidade* e *Eu sou um aluno de uma instituição de ensino superior*. Ambas possuem o mesmo contexto de informação, demonstrando um alto grau de semelhança. Porém, para a maioria dos métodos de similaridade textual léxica essa semelhança conceitual não se torna detectável, considerando ambas as frases completamente distintas e não semelhantes. Dessa forma, os métodos de similaridade textual semântica procuram contornar essa deficiência de análise, utilizando técnicas para encontrar a similaridade entre as duas sentenças. Então, para a realização da análise textual, os métodos de similaridade textual semântica devem principalmente levar em conta a estrutura dos textos a serem utilizados [38].
- Métodos híbridos: são métodos de similaridade que utilizam a junção de vários algoritmos para verificar a semelhança entre textos, mesclando tanto técnicas de análise semântica como métodos baseados em corpus [21] [36].
- Métodos baseados em características: buscam determinar o nível de similaridade textual empregando um conjunto de características pré-definidas para a representação dos textos. Para isso, necessita-se de um classificador inicialmente treinado [21].

## 4.2 Similaridade de Patentes

Patentes são, por natureza, indicadores de inovação e crescimento econômico de uma nação. Das 50 milhões de patentes existentes, 90% possuem caráter científico e tec-

nológico. Além disso, estudos mostram que a utilização correta das informações contidas nas patentes pode minimizar em 40% os gastos com *P&D* de uma empresa [51].

Devido à ampla quantidade de documentos de patentes disponíveis em vários escritórios de patentes pelo mundo, o processo de detecção manual de similaridade entre várias patentes não se torna uma tarefa fácil de ser executada. Realizar tal tarefa pode acarretar em grandes esforços físicos e intelectuais, além de um grande consumo de tempo pelos profissionais da área.

Patentes são consideradas semelhantes quando, em seu escopo, possuem a descrição de uma mesma invenção com conceitos equivalentes ou similares. Também, para se determinar a similaridade desses tipos de documentos, deve-se também levar em conta que duas patentes podem possuir as mesmas palavras-chaves para sua definição e, ainda assim, descrever invenções diferentes. O mesmo vale para o conceito oposto de possuir diferentes palavras-chaves e descrever a mesma invenção [26].

Existem inúmeras aplicações computacionais para as técnicas de detecção de similaridade de patentes, sendo principalmente utilizadas com o intuito de minimizar o esforço manual de análise de documentos. É por meio dos métodos de similaridade que se tornam exequíveis, por exemplo, o processo de *filtragem* das patentes mais similares em relação às menos semelhantes para uma dada tecnologia em investigação, proporcionando aplicações como a busca de invenções similares, pesquisas de anterioridade tecnológica, detecções de infringimento patentário, melhorias na inteligência competitiva de uma empresa por meio da redução do intervalo de análise dos documentos, descobertas de novas oportunidades inventivas e a análise do conteúdo de patentes [26].

### 4.3 Representação de Patentes

Os documentos de patentes possuem uma extensão textual bastante variada em relação ao conteúdo, uma vez que podem descrever invenções simples como um ornamento de um objeto ou mesmo invenções mais elaboradas e complexas como um procedimento industrial para a produção de uma substância química. Assim, a escolha da técnica para a obtenção da similaridade textual entre patentes deve ser feita com muito critério, considerando-se principalmente o conteúdo ou partes do documento a serem utilizados [35].

Para se alcançar um resultado plausível para a análise de similaridade das patentes, campos como o título (*Title*), o resumo (*Abstract*) e principalmente o campo de reivindicações das patentes (*Claims*) devem ser especialmente considerados para a análise [26] [19].

Segundo *Islam e Inkpen* [21], devido à extensão textual do conteúdo de documentos de patentes, os algoritmos mais apropriados para a análise de similaridade são aqueles desenvolvidos especificamente para longas estruturas de texto. Isto é, para se conseguir resul-

tados aceitáveis para a análise de similaridade de documentos patentários recomendam-se os métodos do tipo baseados em Corpus (*Corpus-based methods*).

Assim, para o procedimento de testes e comparação do método *LDA* em relação a outros métodos de avaliação de similaridade de patentes, o sistema desenvolvido (maiores detalhes no Capítulo 6) emprega as seguintes combinações de representação de documentos e distâncias de similaridade:

- Representação de espaço vetorial binário e uso da Distância Euclidiana;
- Representação *TF-IDF* e uso da Distância dos Cossenos.

### 4.3.1 Espaço Vetorial Binário

Tradicionalmente, em campos da inteligência artificial como o reconhecimento de padrões e a aprendizagem de máquina, o paradigma mais comum para se representar um objeto seria a adoção do modelo de vetor espacial de características (*Feature vector space*) devido à simplicidade de realização de operações numéricas com o mesmo. Assim, a representação de um documento de texto pode ser feita por meio de vetores de palavras ou termos associados ao documento.

Com isso, um conceito bastante usual para a representação de um documento seria o uso de um vetor binário  $n$ -dimensional de termos. Nessa representação, cada dimensão do vetor corresponde a uma palavra e, caso essa se encontre no documento, associa-se a dimensão com o valor não zero.

### 4.3.2 *Term Frequency–Inverse Document Frequency (TF-IDF)*

*Term Frequency–Inverse Document Frequency (TF-IDF)*, também conhecido como *TFIDF*, é um método de caráter estatístico baseado em vetor, sendo bastante empregado nas áreas de recuperação de informação e mineração de texto.

Desenvolvido inicialmente para o processo de seleção de documentos relevantes contidos em uma base de dados utilizando-se de uma simples *query de busca*, esse procedimento baseia-se na proposição de quão importante é uma palavra para um determinado documento pertencente a uma coleção ou corpora de documentos textuais [41]. Com isso, os documentos e *queries* podem ser representados na forma de vetores como mostram a Função 4.1 e a Função 4.2, respectivamente [45].

$$D = (d_1, \dots, d_j) \quad j \geq 1 \quad (4.1)$$

$$Q = (t_1, \dots, t_i) \quad i \geq 1 \quad (4.2)$$



Para o cálculo do *TF-IDF* de uma coleção de documentos, deve-se a princípio calcular a *Term Frequency (TF)* e a *Inverse Document Frequency (IDF)* [46].

Utilizado por muitos anos no ramo de indexação automática de texto, *Term Frequency* pode ser essencialmente definido como o número de vezes que certo termo aparece em um documento ou *query*. A Função 4.3 apresenta o cálculo do Term Frequency, onde  $n_{i,j}$  indica o número de vezes que um termo  $t_i$  ocorre em um documento  $d_j$ . Já  $\sum_k n_{k,j}$  é a soma do número de vezes que todos os termos ocorrem no documento  $d_j$ .

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (4.3)$$

Porém, somente o uso de *Term Frequency* não se garante um resultado aceitável para a recuperação das informações necessárias de uma base de dados. Pois, no momento em que o termo está presente em praticamente todos os documentos da coleção e não em apenas poucos documentos, *TF* faz com que sejam retornados todos os documentos para a *query*, prejudicando a precisão de busca [45]. Assim, para contornar isso, utiliza-se de *Inverse Document Frequency* que varia inversamente com o número de documentos contendo o termo associado ao número total de documentos da coleção (ver a Função 4.4).

$$idf_i = \log \frac{|D|}{1 + |\{d : t_i \in d\}|} \quad (4.4)$$

Na função anteriormente apresentada (Função 4.4),  $|D|$  é o número de documentos totais da coleção e  $|\{d : t_i \in d\}|$  é o número de documentos que possuem um termo  $t_i$ . Com isso, o cálculo para *TF-IDF* pode ser dado como se mostra a Função 4.5.

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \quad (4.5)$$

Dessa forma, conclui-se que os documentos que repetem o maior número de vezes os termos da *query* de busca, consequentemente possuem um grande conteúdo em comum com a *query* e receberão, por conseguinte, um maior valor para o *TF-IDF*. Além disso, o método proporciona atribuir valores maiores de *TF-IDF* para as palavras que aparecem frequentemente em um pequeno grupo de documentos do que as palavras com pouca relevância de análise como artigos, pronomes e preposições [41] [24].

## 4.4 Métricas de Similaridade

Após a escolha do modelo para a representação dos documentos, a fim de se obter um valor numérico indicando a semelhança entre as patentes avaliadas, adota-se uma medida ou distância de similaridade para a avaliação.

### 4.4.1 Distância Euclidiana

A Distância Euclidiana (*Euclidian Distance*) ou Distância L2 (*L2 Distance*) é a métrica mais simples e frequentemente utilizada para se calcular o grau de similaridade entre dois objetos em um espaço dimensional [3]. Para isso, os pares de objetos a serem analisados são representados na forma de um conjunto de vetores e, a partir disso, aplica-se a função da Distância Euclidiana (Função 4.6) para obter o nível de semelhança. De forma geral, pode-se dizer que a função calcula o grau de similaridade por meio da raiz quadrada da diferença entre as coordenadas dos pares de objetos [46].

$$D_{Euc}(x, y) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2} \quad (4.6)$$

Na função apresentada (Função 4.6), as variáveis  $x$  e  $y$  são vetores  $n$ -dimensionais das patentes  $x$  e  $y$ , respectivamente.  $N$  indica o número de elementos dos vetores. E, além disso,  $D_{Euc}(x, y)$  deve seguir as seguintes propriedades:

- $D_{Euc}(x, y) \geq 0$ , isto é, a função deve retornar zero ou um valor positivo;
- $D_{Euc}(x, y) = 0$ , se somente  $x = y$ , ou seja, os objetos a serem comparados são exatamente iguais;

### 4.4.2 Distância dos Cossenos

A Distância dos Cossenos (*Cosine Distance*), assim como a Distância Euclidiana, também é uma métrica utilizada para se calcular o grau de similaridade entre dois objetos. Para isso, ambos os objetos a serem comparados devem estar representados na forma de vetores  $n$ -dimensionais. A Função 4.7 apresenta a função para o cálculo da Distância dos Cossenos.

$$D_{Cos}(x, y) = \cos\theta = \frac{X \cdot Y}{\|X\| \|Y\|} = \frac{\sum_{i=1}^N x_i y_i}{\sqrt{\sum_{i=1}^N x_i^2} * \sqrt{\sum_{i=1}^N y_i^2}} \quad (4.7)$$

Onde,  $x$  e  $y$  são vetores  $n$ -dimensionais de uma patente  $x$  e outra patente  $y$ . Na função,  $N$  indica o número de elementos dos vetores e  $D_{Cos}(x, y) \geq 0$ .

### 4.4.3 Earth Mover's Distance

*Earth Mover's Distance (EMD)*, também conhecido como *1st Wasserstein* ou *Monge-Kantorovich*, pode basicamente ser definido como um método para se calcular a distância entre duas distribuições finitas, ou seja, o custo mínimo para se transformar uma determinada distribuição em outra movendo a sua *massa de distribuição* [49].

Dessa forma, pode-se dizer que *EMD* é baseado em um caso especial do problema de transporte (*transportation problem*) no qual se deseja minimizar o custo de transporte e alocação de recursos para uma série de pontos de demanda (consumidores) a partir de um grupo de pontos de oferta (fornecedores) [44].

As distribuições utilizadas para o cálculo de *EMD* são formadas por um conjunto de *signatures* – variáveis de características da distribuição compostas por um *identificador* e um *valor de massa* – na forma  $\{(x_1, p_1), \dots, (x_m, p_m)\}$ .

Dadas duas distribuições  $P = \{(x_1, p_1), \dots, (x_m, p_m)\}$  e  $Q = \{(y_1, q_1), \dots, (y_m, q_m)\}$ , deseja-se encontrar o fluxo  $F = f_{ij}$  que minimiza o custo total (Função 4.8).

$$W(P, Q, F) = \sum_{i=1}^m \sum_{j=1}^n f_{ij} c_{ij} \quad (4.8)$$

Onde,  $c_{ij} = c(x_i, y_j)$  é o custo ou distância do *signature*  $x_i \in P$  a  $y_j \in Q$  e  $W(P, Q, F)$  representa o trabalho necessário para se mover a massa de um *signature* a outro [31]. Assim,  $f_{ij}$  deve estar sujeito às seguintes restrições:

$$f_{ij} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n \quad (1)$$

$$\sum_{j=1}^n f_{ij} \leq p_i, 1 \leq i \leq m \quad (2)$$

$$\sum_{i=1}^m f_{ij} \leq q_j, 1 \leq j \leq n \quad (3)$$

$$\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min \left( \sum_{i=1}^m p_i, \sum_{j=1}^n q_j \right) \quad (4)$$

(4.9)

Assim, encontrado o  $f_{ij}$  mínimo, pode-se calcular o *EMD* entre P e Q através da Função 4.10.

$$EMD(P, Q) = \min_{f_{ij}} \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} c_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (4.10)$$

Onde,  $\sum_{i=1}^m \sum_{j=1}^n f_{ij}$  é um fator de normalização que permite favorecer *signatures* com menor massa total.



# Capítulo 5

## Extração de Tópicos

Com o crescimento exponencial da quantidade de informações publicamente acessíveis, principalmente com o advento da internet, tarefas antes consideradas essencialmente simples de serem realizadas tornaram-se extremamente complexas. Assim, de uma maneira geral, pode-se dizer que a informação está cada vez mais oscilando entre um aspecto útil ao problemático, limitando drasticamente tarefas de extração do valioso conhecimento necessário [34].

Para procurar contornar ou amenizar o excesso de dados e obter apenas a informação útil de um texto, a princípio, podem-se encontrar as partes consideradas significativas do conteúdo textual dos documentos, enfocando-se basicamente nos conceitos (tópicos) aos quais esses pertencem. Dessa forma, a posteriori, pode-se iniciar a investigação da informação almejada [16].

Adotando-se uma perspectiva da área da linguística, o significado das palavras não depende unicamente da palavra em si, mas, do contexto a qual ela se insere. Por meio disso, tópicos, também conhecidos como cluster de palavras [33], buscam sintetizar os assuntos de que se tratam um determinado documento, sendo cada conceito representado por uma distribuição de palavras que juntas buscam formar um significado. Além disso, o procedimento de se desvendar tópicos pode ser considerado o primeiro passo para se obter uma visão global do assunto presente em um determinado conteúdo textual [12].

O modelo de tópicos fornece um método não supervisionado de extrair uma representação interpretável de uma coleção de documentos, sendo útil para aplicações que requerem uma análise inicial da estrutura de texto. Alguns exemplos disso são o pré-processamento em mineração e sumarização de texto, extração de informações e similaridade de documentos [40].

## 5.1 Latent Dirichlet Allocation

Desenvolvido por *Blei et al.* [1], o modelo de tópicos *Latent Dirichlet Allocation (LDA)* é uma classe de rede *Bayesiana* probabilística para um corpus de dados distintos, usado principalmente para modelar palavras de um texto na estrutura de *bag of words*, ignorando, dessa forma, a ordem em que aparecem as palavras em um documento. Esse método baseia-se essencialmente na proposição de que um documento  $D$  pode ser representado como uma mistura de tópicos  $K$ , sendo cada um desses tópicos representados como uma distribuição probabilística multinomial de palavras de um vocabulário  $W$  [23].

A Figura 5.1 mostra a representação gráfica do modelo probabilístico *LDA*. Cada um dos retângulos é considerado uma espécie de *prato* que expressam réplicas de seu conteúdo. As variáveis  $\alpha$  e  $\beta$  são parâmetros de corpus de *Dirichlet*, sendo amostradas apenas uma única vez no processo de geração do corpus textual. Da mesma forma,  $\theta_d$  são variáveis de documento, sendo também amostrados uma vez por documento. Já as variáveis  $z_{id}$  e  $w_{id}$  são variáveis de palavra e são escolhidas uma vez para cada palavra no documento.

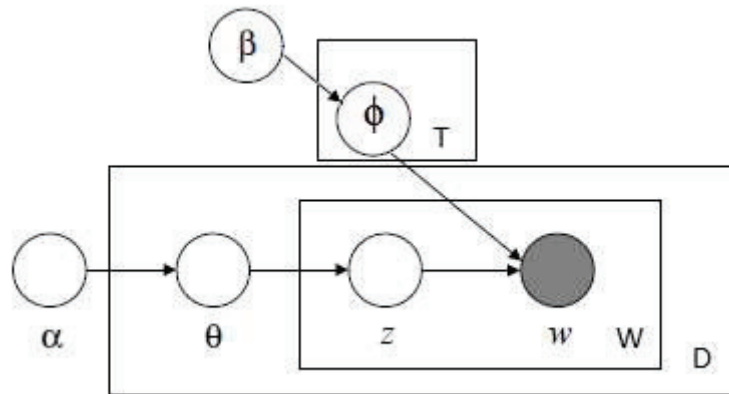


Figura 5.1: Modelo gráfico para *LDA*

No método *LDA*, para cada documento  $d = 1, \dots, D$  pertencente ao corpus, escolhe-se uma distribuição multinomial  $\theta_d = [\theta_{d1}, \dots, \theta_{dk}]^T$  condicionado a distribuição de *Dirichlet*  $\alpha = [\alpha_1, \dots, \alpha_k]^T$ . Assim, conforme a distribuição multinomial  $\theta_d$ , associa-se para cada palavra encontrada no documento  $d$  um tópico  $z_{id} = k$ , onde  $k = 1, \dots, K$ . A partir disso, o modelo *LDA* escolhe uma palavra  $w_{id} \in W$  de acordo com a distribuição multinomial  $\phi_k = [\phi_{k1}, \dots, \phi_{kV}]^T$  condicionado a distribuição de *Dirichlet*  $\beta = [\beta_{k1}, \dots, \beta_{kV}]^T$ .

Resumidamente, pode-se dizer que os parâmetros multinomiais  $\theta_d$  e  $\phi_k$  possuem prioridades de *Dirichlet*. E, dessa maneira, pode-se concluir que a distribuição  $\phi_k$  representa quais são as palavras mais importantes para um tópico  $k$  e o parâmetro  $\theta_d$  procura representar quais os prováveis tópicos que caracterizam um determinado documento  $d$  [40].

Assim, com a obtenção de  $w = \{w_{id}\}$  por meio da inferência bayesiana (*Bayesian Inference*), calcula-se a distribuição posterior de tópicos latentes  $z = \{z_{id}\}$ ,  $\theta_d$  e os parâmetros de tópico  $\phi_k$ . E para isso, um modo de se calcular essas distribuições pode ser feito através do método *Collapsed Gibbs Sampling* em que  $\theta$  e  $\phi$  são marginalizados e estimados após a amostragem do parâmetro  $z$ .

Conforme apresentado na Figura 5.1, a Distribuição Conjunta (*Joint probability distribution*) de todos os parâmetros e variáveis pode ser expressa pela Função 5.1:

$$p(w, z, \theta, \phi | \alpha, \beta) = p(\theta | \alpha) p(z | \theta) p(\phi | \beta) p(w | z, \phi) \quad (5.1)$$

Onde,  $p(\theta | \alpha) p(z | \theta) p(\phi | \beta)$  e  $p(w | z, \phi)$  são dados conforme as funções 5.2, 5.3, 5.4 e 5.5, respectivamente.

$$p(\theta | \alpha) = \left( \prod_{d=1}^D \frac{\Gamma(\alpha_{\cdot})}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{dk}^{\alpha_k - 1} \right) \quad (5.2)$$

$$p(z | \theta) = \left( \prod_{d=1}^D \prod_{k=1}^K \theta_{dk}^{n_{dk}} \right) \quad (5.3)$$

$$p(\phi | \beta) = \left( \prod_{k=1}^K \frac{\Gamma(\beta_{k\cdot})}{\prod_{v=1}^W \Gamma(\beta_{kv})} \prod_{v=1}^W \phi_{kv}^{\beta_{kv} - 1} \right) \quad (5.4)$$

$$p(w | z, \phi) = \left( \prod_{k=1}^K \prod_{v=1}^W \phi_{kv}^{n_{dkv}} \right) \quad (5.5)$$

Assim,  $p(w, z, \theta, \phi | \alpha, \beta)$  é dado conforme a função a seguir:

$$p(w, z, \theta, \phi | \alpha, \beta) = \left( \prod_{d=1}^D \frac{\Gamma(\alpha_{\cdot})}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{dk}^{\alpha_k + n_{dk} - 1} \right) \left( \prod_{k=1}^K \frac{\Gamma(\beta_{k\cdot})}{\prod_{v=1}^W \Gamma(\beta_{kv})} \prod_{v=1}^W \phi_{kv}^{\beta_{kv} + n_{\cdot kv} - 1} \right)$$

Onde,  $\alpha_k$  se refere à prioridade de *Dirichlet* para o tópico  $k$ ,  $\beta_{kv}$  representa a prioridade para a  $v$ -ésima palavra no vocabulário  $W$  sobre o tópico  $k$ ,  $\theta_{dk}$  indica a probabilidade de se associar palavras de  $d$  para um tópico  $k$ ,  $\phi_{kv}$  representa a probabilidade de gerar a  $v$ -ésima palavra de um vocabulário sobre um tópico  $k$  e  $n_{dkv}$  é o número de palavras do documento  $d$  que são associadas ao tópico  $k$  e possuem valor  $v$ . Além disso, para facilitar a visualização da função anteriormente descrita, adota-se o uso do *ponto subscripto* que denota a somatória de valores. Assim,  $\alpha_{\cdot} = \sum_{k=1}^K \alpha_k$  e  $n_{\cdot kv} = \sum_{d=1}^D (n_{dkv})$ .

*LDA* possui inúmeras vantagens em relação a outros métodos existentes para a representação das informações como, por exemplo, o fato dos tópicos serem extraídos de forma não supervisionada e a capacidade de serem individualmente interpretáveis.

O uso de métodos não supervisionados torna-se de grande importância para a extração das informações, uma vez que não se pode sempre utilizar tópicos pré-definidos para representar o corpus de um documento. Esse fato se deve normalmente a inconstante mudança tecnológica, pois, cada vez mais novos tópicos vão emergindo para as diversas áreas do conhecimento, não refletindo muitas vezes o real conteúdo dinâmico envolvido quando se utiliza o esquema de tópicos pré-definidos [42].

## 5.2 Collapsed Gibbs Sampling

Uma das maneiras utilizadas para o processo de se extrair tópicos de uma base de dados seria utilizar a distribuição posterior de associações de palavras para os tópicos e obter estimativas para  $\theta_d$  e  $\phi_k$ .

No entanto, calcular uma distribuição complexa de probabilidades não é uma tarefa simples de ser implementada. Assim, para se contornar esse problema, utiliza-se de *Markov Chain Monte Carlo*. Esse método procura construir uma cadeia de *Markov (Markov Chain)* específica que converge para uma distribuição alvo e, assim, as amostras são obtidas por meio dessa distribuição. Durante cada estado de transição da cadeia, utiliza-se de o método de *Gibbs sampling* [12] para a associação de valores das variáveis amostradas, isto é, obtêm-se os novos valores das variáveis de suas distribuições sendo essas condicionadas aos valores e outros parâmetros envolvidos [13].

*Collapsed Gibbs Sampling*, também conhecido apenas como *Gibbs Sampling*, consiste em um conjunto de técnicas iterativas com a finalidade de amostrar valores de distribuições complexas conforme a probabilidade condicional dada a todas as outras variáveis [23].

Para o modelo *LDA*, necessita-se estimar os parâmetros  $\theta_d$  e  $\phi_k$  conforme a distribuição de tópicos  $z_{id} = k$  [47] e, dessa forma, calcula-se a distribuição posterior  $P(z_{i,d} = k | z_{-i,d}, w_{i,d}, \alpha, \beta)$  [12] [13] conforme apresentado na Função 5.6.

$$P(z_{i,d} = k | z_{-i,d}, w_{i,d}, \alpha, \beta) = \frac{1}{Norm} \bullet \frac{(n_{-i,k}^{w_i} + \beta)(n_{-i,k}^{(d_i)} + \alpha)}{n_{-i,k}^{(\cdot)} + W\beta} \quad (5.6)$$

onde, *Norm* é a constante de normalização observada na Função 5.7.

$$Norm = \sum_{k=1}^K \frac{(n_{-i,k}^{w_i} + \beta)(n_{-i,k}^{(d_i)} + \alpha)}{n_{-i,k}^{(\cdot)} + W\beta} \quad (5.7)$$

$z_{-i,d}$  é a associação de todos  $z_{j,d} (j \neq i)$ ,  $n_{-i,k}^{w_i}$  é o número de mesmas palavras associadas ao tópico  $k$ ,  $n_{-i,k}^{(\cdot)}$  é o número de palavras associadas ao tópico  $k$ ,  $n_{-i,k}^{(d_i)}$  é o número de palavras associadas ao tópico  $k$  no documento  $d_i$ . Além disso, considera-se que todas essas variáveis apresentadas não incluem a associação corrente  $\phi$ .



Resumidamente, para cada processo iterativo do método de *Gibbs Sampling*, obtém-se um valor de  $z_{i,d}$  para cada palavra  $i$  em um documento  $j$  (Função 5.6) [40]. Com isso, após a *Markov Chain* alcançar o estado de convergência da distribuição, os atuais valores de  $z$  são gravados e, assim, pode-se realizar uma estimativa para os parâmetros  $\theta_d$  e  $\phi_k$  para uma determinada amostra. Então, as seguintes funções (Função 5.8 e Função 5.9) são utilizadas:

$$\hat{\phi}_w^{(z=k)} = \frac{n_k^{(w)} + \beta}{n_k^{(\cdot)} + W\beta} \quad (5.8)$$

$$\hat{\theta}_{z=k}^{(d)} = \frac{n_k^{(d)} + \alpha}{n^{(d)} + K\alpha} \quad (5.9)$$

Onde,  $n_k^{(w)}$  é o número de palavras  $w$  associadas ao tópico  $k$ ,  $n_k^{(\cdot)}$  é o número de palavras associadas ao tópico  $k$ ,  $n_k^{(d)}$  é o número de palavras associadas ao tópico  $k$  no documento  $d$  e  $n^{(d)}$  é o número de palavras associadas aos tópicos no documento  $d$ .

Uma característica interessante do *Collapsed Gibbs Sampling* é a sua adaptabilidade quanto à adição de novos documentos a corpora. Pois, sempre que se acrescenta um novo documento, não se torna obrigatório o processamento total da *Markov Chain* para se deduzir as novas variáveis latentes. Isso é feito apenas expandindo o estado anterior da cadeia com respeito ao novo documento, propiciando em uma obtenção mais ágil para a nova distribuição estacionária.



# Capítulo 6

## Sistema

Muitas vezes tornam-se inviáveis as análises manuais de cada uma das informações contidas em um conjunto de documentos, proporcionando uma tarefa completamente exaustiva e repetitiva de ser executada.

Nesse capítulo apresenta-se a arquitetura do sistema desenvolvido, mostrando o uso e as vantagens do método *LDA* em comparação aos métodos mais comuns existentes de similaridade de patentes. Também, procura-se demonstrar a aplicação do sistema para análises de redes de citações de patentes.

### 6.1 Objetivo

A necessidade cada vez maior de ferramentas para realizar extração de informações de forma automática, sem a necessidade de uma supervisão sistemática pelo ser humano, implica na descoberta de novos meios para desempenhar tarefas de alto custo temporal e, assim, poder obter as informações almejadas.

Seguindo esse conceito de agilidade para o processamento de informações, operações como a análise e a extração de dados perante a uma base de dados muito extensa tornam-se praticamente dependentes de meios computacionais para a obtenção de resultados.

Da mesma forma, devido à necessidade de análises semânticas de documentos de patentes, desenvolveu-se um sistema computacional para auxiliar no processo de análise de conteúdo de patentes utilizando métodos de similaridade textual.

O sistema computacional denominado *Odyseýs* possui como ênfase a aplicação do método *LDA with Collapsed Gibbs Sampling* para realizar a análise de similaridade entre documentos, utilizando o conceito de tópicos para construir também uma rede de citações de patentes. Essa rede resultante possui inúmeras aplicações, conforme mencionado no Capítulo 3.7, além da capacidade de proporcionar a análise da patente em avaliação em relação a outras patentes semelhantes pertencentes à rede de citações de uma determinada

tecnologia. Também, torna-se possível para o profissional especialista na área de análise de patentes compreender os campos emergentes e as prospecções para uma determinada tecnologia.

## 6.2 Implementação

O sistema de análise de patentes *Odysseýs* foi desenvolvido utilizando-se a linguagem de programação *Java*. Pois, além de fornecer a capacidade de portabilidade entre diferentes sistemas, disponibiliza vários pacotes para a criação de redes de citação e permite a adição de novas bibliotecas algorítmicas.

Assim, para a construção das redes de citações utilizadas no projeto, empregou-se a biblioteca de código aberto (*Open Source*) *JUNG (Java Universal Network/Graph Framework)*<sup>1</sup>. Essa interface fornece inúmeras técnicas para a visualização, a análise e a modelagem de dados em uma estrutura de grafos, permitindo também o suporte a representações de vários tipos de entidades e relacionamentos. Outro fator para a escolha de *JUNG* foi à disponibilidade de vários algoritmos de indicadores e métricas para a avaliação das redes de citações.

No processo de análise de similaridade entre patentes, o sistema emprega o método *LDA with Gibbs Sampling* para extrair os tópicos dos corpora de patentes e realizar a representação de documentos, utilizando-se para isso, pacotes pertencentes à biblioteca *Weka*<sup>2</sup> e ao projeto *knowceans.org*<sup>3</sup>. Por fim, o cálculo da distância de similaridade é obtido através do algoritmo *Earth Mover's Distance (EMD)*<sup>4 5</sup>.

Para se comparar e testar a eficiência do *LDA* em relação a outras técnicas de similaridade textuais, utiliza-se da biblioteca *LingPipe*<sup>6</sup> para a representação *TF-IDF* e a obtenção da distância dos cossenos. E, também, para a representação binária e a métrica da distância euclidiana utiliza-se da biblioteca *SimMetrics*<sup>7</sup>.

Por fim, para o tratamento e filtragem das informações contidas nas patentes são utilizadas técnicas de *Stemming* e eliminação de palavras irrelevantes [38] através de algoritmos pré-implementados na biblioteca *Weka*.

---

<sup>1</sup>JUNG - Java Universal Network/Graph Framework. Disponível em: <http://jung.sourceforge.net/>

<sup>2</sup>Weka Machine Learning Project. Disponível em: [www.cs.waikato.ac.nz/ml](http://www.cs.waikato.ac.nz/ml)

<sup>3</sup>Knowceans.org. Disponível em: <http://www.arbylon.net/projects/>

<sup>4</sup>Code for the Earth Mover's Distance. Disponível em: <http://ai.stanford.edu/~rubner/emd/default.htm>

<sup>5</sup>Fast Earth Mover's Distance (EMD) Code. Disponível em: <http://www.cs.huji.ac.il/~ofirpele/FastEMD/code/>

<sup>6</sup>LingPipe. Disponível em: <http://alias-i.com/lingpipe/>

<sup>7</sup>SimMetrics. Disponível em: <http://staffwww.dcs.shef.ac.uk/people/S>

## 6.3 Metodologia

O sistema funciona por meio de análises textuais entre uma patente a ser avaliada, ou mesmo um pré-modelo de patente provisória, e um grupo de documentos pertencentes a uma mesma área tecnológica ou, até mesmo, de campos distintos.

Para o processo de geração de redes de citações de patentes, utilizam-se como padrão as patentes originárias da base de dados *USPTO*, uma vez que os documentos são disponibilizados abertamente ao público via *internet* e possuem uma estrutura padrão para a avaliação e extração de informações.

Dessa forma, o *software* pode realizar buscas de patentes de forma não supervisionada, retornando um conjunto de documentos correlacionados presentes na base de dados do *USPTO* por meio de uma consulta – *query de busca* – designada pelo usuário. Para esse procedimento, os termos de pesquisa podem ser fornecidos conforme campos já pré-definidos pelo sistema ou utilizar uma *query* personalizada conforme o padrão descrito para buscas de patentes na base *USPTO*.

Assim, o sistema computacional inicia o procedimento de extração de informações para gerar uma rede de patentes descrita em forma de um grafo de relacionamento (dígrafo). Essa rede permite mostrar as correlações entre os atores e o fluxo de conhecimento para a tecnologia investigada, além das tendências de áreas de pesquisa e uma visão geral de progressão tecnológica. A característica de citar uma determinada patente significa que a essa possui algum tipo de informação considerada essencial para o desenvolvimento tecnológico das patentes que a citam.

O *software* permite ao pesquisador ganhar escala e agilidade na investigação, pois, cada *query de busca* pode resultar em um número indeterminado de patentes, podendo cada uma das patentes citar e ser citada, respectivamente, por várias outras patentes anteriores e posteriores adicionadas base de dados *USPTO*. Assim, percebe-se que a geração manual de tais redes de citações implica em um trabalho bastante desgastante e demorado, uma vez que a rede resultante pode ser muito densa e possuir várias conexões entre os seus nós.

Finalmente, obtida a rede de citações, torna-se então plausível a aplicação de métricas [43] e indicadores de rede [39] [32] para uma melhor avaliação de perspectiva da tecnologia em investigação.

A Figura 6.1 apresenta um exemplo de geração de rede de citações pelo *Odysseýs* para a tecnologia etanol de patentes com a classificação internacional da patente (*IPC*) *C07H*. As patentes principais retornadas possuem a palavra-chave *ethanol* nos campos *title*, *abstract* e *claims*; e são representadas por vértices vermelhos. Também na mesma imagem, são mostradas as patentes que citam e são citadas pelas patentes principais, sendo essas representadas por vértices azuis. Os números ao lado de cada autor indicam

o número das patentes em relação à base de dados *USPTO*. A *query de busca* utilizada para a geração da rede é apresentada a seguir:

*TTL/Ethanol and ABST/Ethanol and ACLM/Ethanol and ICL/C07H\$*

Onde, segundo o padrão *USPTO*, *TTL* indica o campo *title*, *ABST* o campo *abstract*, *ACLM* o campo *claims* e *ICL* representa a *IPC* para o documento.

Conforme a Organização Mundial da Propriedade Intelectual (*OMPI* ou, em inglês, *WIPO*)<sup>8</sup>, a *IPC C07H* é descrita como:

- **Seção C: Química e Metalurgia**

- **C07: Química Orgânica** (componentes como os óxidos, sulfetos ou Sulfeto de carbonila, cianogênio, Fosgênio, ácido hidrocianídrico ou sais de C01; produtos obtidos de mudança de base de silicatos em camadas por midança de íon com componentes orgânicos tais como amônia, fosfonio ou componentes de sulfonio ou por intercalação de componentes orgânicos C01B 33/44; componentes macromoleculares C08; corantes C09; produtos de fermentação C12; fermentação ou processos de uso de enzimas para sintetizar um componente químico desejado ou composição ou para separar isômeros ópticos de uma mistura racêmica C12P; produção de componentes orgânicos por eletrólise ou eletroforese C25B 3/00, C25B 7/00)

- \* **C07H: Açúcares; Derivados; Nucleósidos; Nucleotídeos; Ácidos Nucléicos** (derivados de ácidos aldônicos ou sacáricos C07C, C07D; ácidos aldônicos, ácidos sacáricos C07C 59/105, C07C 59/285; cianohidrinas C07C 255/16; glicais C07D; componentes de constituição desconhecida C07G; polissacarídeos, derivados C08B; DNA ou RNA sobre engenharia genética, vetores, como, por exemplo, plasmídeos, ou seu isolamento, preparação ou purificação C12N 15/00; indústria de açúcar C13)

---

<sup>8</sup>WIPO – World Intellectual Property Organization. Disponível em: <http://www.wipo.int/>

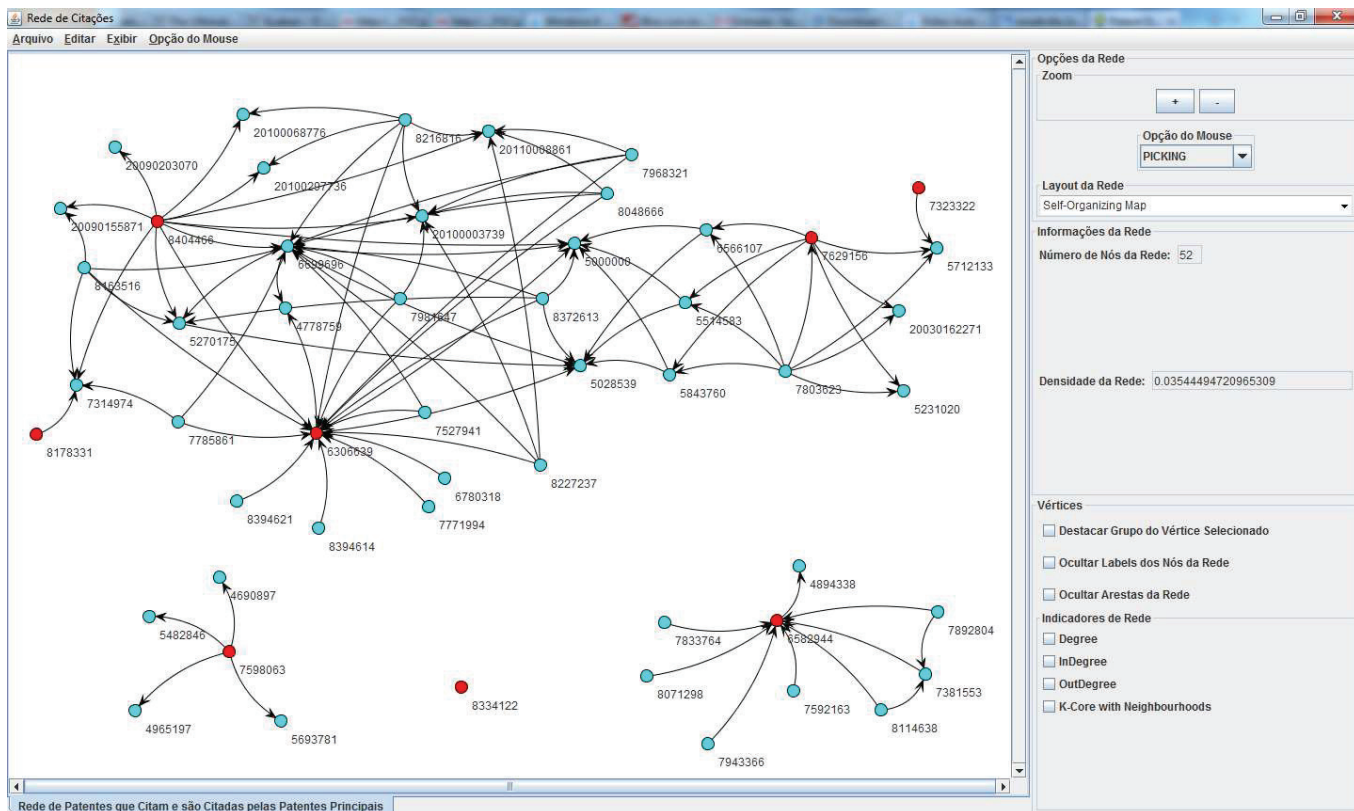


Figura 6.1: Rede de citações para a tecnologia *ethanol* conforme  $IPC=C07H$

Além da busca pela existência de patentes já concedidas de tecnologias ou produtos similares no banco de dados patentários, o exame do conteúdo dos documentos resultantes da pesquisa requer, normalmente, um grande trabalho manual por parte dos profissionais do ramo de análise de patentes.

Partindo-se desse fato, torna-se interessante à aplicação de técnicas de similaridade de patentes com o intuito de agilizar o processo de análise das informações. Com isso, pode-se realizar uma espécie de pré-seleção dos documentos e organizá-los conforme um grau de semelhança dos seus conteúdos em relação à tecnologia avaliada, reduzindo, assim, a quantidade de patentes a serem individualmente verificadas.

A partir de então, buscaram-se algoritmos de verificação de similaridade de texto que se adequem à análise de patentes, uma vez que uma patente pode variar muito a sua extensão textual. Assim, o sistema *Odysséys* procura avaliar a eficiência de se utilizar o método *LDA* para se encontrar as patentes mais semelhantes àquela patente dada pelo usuário. A escolha do método se deve ao fato de que, conforme estudos anteriores da literatura, a relevância dos resultados obtidos para o processo de categorização de documentos pode ser melhorada utilizando-se técnicas baseadas em tópicos [19]. Além disso, geralmente,

quando as unidades de textos são individualmente comparadas, a combinação de determinadas palavras e frases são suficientes para se detectar a similaridade entre documentos [15]. No entanto, quando os segmentos textuais para a análise possuem comprimento muito variado, alguns métodos de similaridade podem não identificar a semelhança existente entre alguns documentos.

Diferentemente dos métodos usuais, o procedimento de dispor documentos em forma de tópicos não necessita de um tamanho textual fixo para análise de textos. Além disso, a identificação de um documento por meio de tópicos permite a compreensão de uma rede de citações de maneira mais consistente pelo usuário, pois, consegue identificar a ideia do conteúdo presente em um grupo de patentes e suas correlações entre os autores.

Aproveitando-se dessa característica de agrupamento de patentes por tópicos, adicionou-se ao sistema a funcionalidade de se visualizar o documento e o grupo de análise em forma de rede de citações. Para isso, o sistema encontra as patentes similares de um grupo e cria arestas de citações do documento analisado em relação a essas patentes. Uma exemplificação mais detalhada para esse procedimento pode ser observada no Capítulo 6.4.

Para a avaliação da similaridade por meio da utilização de tópicos de documentos, o *software* utiliza-se do método *LDA with Gibbs Sampling*. A opção de uso do método se deve ao fato de que o modelo *LDA* e o algoritmo de inferência *Gibbs Sampling* são amplamente utilizados na literatura para a geração de tópicos em análise de textos [40] [47] e, dessa forma, aplicáveis para documentos patentários.

Para o contexto de análise de similaridade empregado pelo sistema desenvolvido, adota-se sempre como conjunto base de treinamento para o método *LDA* todos os corpora textuais presentes no experimento de análise de patentes. Assim, por exemplo, em um experimento composto por um grupo de 30 patentes e uma patente de análise usada como comparação ao grupo de patentes, ambos o conjunto e o documento de análise são utilizados para o processo de aprendizagem do algoritmo *LDA*.

As escolhas dos parâmetros de *dirichlet*  $\alpha$  e  $\beta$  podem ter implicações importantes nos resultados produzidos pelo modelo, uma vez que se aumentando o valor de  $\beta$ , diminui-se o número de tópicos usados para descrever a base de dados [12]. Segundo estudos realizados, os valores  $\alpha = 0.01$  e  $\beta = 0.5$  produzem bons resultados experimentais [16] e, com isso, optou-se pela adoção desses mesmos valores para os parâmetros a serem utilizados no sistema.

Realizada a etapa de associação de tópicos para cada uma das patentes, o passo seguinte é calcular a similaridade de cada um dos documentos do grupo de análise em relação à patente em verificação. Para isso, o sistema retorna por meio de *LDA* uma lista de tópicos pertencentes a cada uma das patentes com suas distribuições multinomiais  $\theta_d$ .

Assim, para o processo do cálculo de similaridade, todas as distribuições  $\theta_d$  do grupo de análise são comparadas ao  $\theta_d$  da patente em avaliação por meio do algoritmo



*Earth Mover's Distance* (Capítulo 4.4.3).

No contexto do sistema desenvolvido, cada patente é considerada uma distribuição e os signatures são representados pela distribuição multinomial  $\theta_d$  obtida pelo método *LDA*. Com isso, dada uma patente  $P = \{(x_1, p_1), \dots, (x_m, p_m)\}$  e uma outra patente  $Q = \{(y_1, q_1), \dots, (y_n, q_n)\}$ , as distância ou custo  $c_{ij} = c(x_i, y_j)$  de um *signature*  $x_i \in P$  em relação a um *signature*  $y_j \in Q$  são dadas conforme as Funções 6.1 e 6.2.

Para  $m > 0$ ,  $n > 0$  e  $i = j$ ,

$$c_{ij} = |p_i - q_j| \quad (6.1)$$

Para  $m > 0$ ,  $n > 0$  e  $i \neq j$ ,

$$c_{ij} = 1 \quad (6.2)$$

Finalmente, para se verificar a eficiência do método e a sua aplicabilidade em documentos de patentes, os resultados de similaridade obtidos pelo método *LDA+EMD* são comparados aos dos métodos Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos. No capítulo a seguir (Capítulo 6.4) são apresentados alguns testes com grupos de patentes e no Capítulo 6.5 a avaliação geral dos resultados obtidos.

## 6.4 Testes e Avaliação

A fim de se verificar a usabilidade do método *LDA* para documentos de patentes, realizaram-se testes utilizando patentes originárias da base de dados *USPTO*.

Para a primeira análise (Capítulo 6.4.1), empregaram-se três grupos de patentes provenientes de áreas de conhecimento distintas, além de uma patente base de análise pertencente a um dos grupos para se processar a análise de similaridade das patentes.

Em seguida, a fim de se examinar os resultados do método de forma mais específica, estabeleceram-se mais três testes com conjuntos de patentes. Porém, as patentes escolhidas para esse propósito possuem a característica de pertencerem a um mesmo campo tecnológico e serem muito similares entre si conforme a avaliação de um examinador de patentes.

Em Anexo 4 dispõem-se das tabelas e listas de tópicos obtidas a partir do método *LDA with Gibbs Sampling* para os experimentos realizados nos capítulos 6.4.1 e 6.4.2 – Análise de grupos tecnológicos distintos de patentes e Análise de grupos tecnológicos similares de patentes.

### 6.4.1 Análise de grupos tecnológicos distintos de patentes

Todas as patentes empregadas no procedimento de testes pertencem à base de dados *USPTO* e os assuntos que especificam os grupos tecnológicos envolvidos para a análise foram escolhidos de forma aleatória. No entanto, para a patente principal de análise, escolheu-se um documento obrigatoriamente pertencente a um dos grupos com o intuito de se avaliar os resultados obtidos. A escolha foi feita de forma completamente aleatória entre as patentes pertencentes ao grupo.

A partir de então, 2 experimentos foram realizados a fim de se verificar o uso e eficiência do método *LDA+EMD* em relação aos demais métodos (Espaço Vetorial Binário + Distância Euclidiana e *TF-IDF* + Distância dos Cossenos) para a análise de similaridade de patentes.

Para o primeiro experimento empregam-se patentes pertencentes a grupos tecnológicos distintos, utilizando-se todo o conteúdo disponível dos documentos para análise, inclusive informações específicas da patente (Capítulo 3.1). Por outro lado, para o segundo experimento, ainda utilizando-se os mesmos grupos de documentos como no experimento anterior, a patente de análise é usada de forma parcial, isto é, apenas certas partes do conteúdo da patente são empregadas para análise (*title*, *abstract*, *claims* e *description*).

Assim, escolheu-se aleatoriamente como documento principal de análise para os experimentos a patente de número *USPTO* 4876196. O seu conteúdo caracteriza um método para a produção de etanol por meio da fermentação de açúcares.

Para compor o primeiro grupo de análise, além da patente de número 4876196, escolheram-se também outros documentos com conteúdo envolvendo o tema etanol e açúcar, conforme mostra a Tabela 6.1.

Já para o segundo e o terceiro grupos utilizam também patentes de campos completamente distintos. O segundo grupo é formado por patentes sobre sistemas de jogos portáteis, enquanto que o terceiro conjunto apresenta patentes de tecnologias relacionadas a laranjeiras. A Tabela 6.1 apresenta o grupo, o número *USPTO* e o título das patentes escolhidas para a realização dos testes.

<b>Grupo</b>	<b>N° USPTO da patente</b>	<b>Título</b>
1	4876196	Method of continuously producing ethanol from sugar-containing substrates
1	4326036	Production of ethanol from sugar cane
1	4560659	Ethanol production from fermentation of sugar cane
1	4738930	Apparatus for continuously recovering ethanol from fermentable sugar solutions
2	5428528	Portable interactive game system between master/slave units containing detachable memories wherein a master unit downloads a master program to slave units respective detachable memories
2	6039574	Time monitoring portable game system
2	6416410	Data compression/decompression based on pattern and symbol run length encoding for use in a portable handheld video game system
2	6478583	Time monitoring portable game system
2	6500070	Combined game system of portable and video game machines
2	7316618	Steering wheel controller for use with embedded portable game system
2	7371163	3D portable game system
3	PP04161	Orange tree
3	PP06047	Variety of navel orange tree
3	PP07651	Navel orange tree named 'Rohde Summer Navel'
3	PP07700	Orange tree 'Beck Early Navel'
3	PP08212	'Chislett Summer Navel' orange tree
3	PP08238	Orange tree named 'Sweet Martin'
3	PP11246	Navel orange tree named 'Wiffen Summer Navel'
3	PP18774	Orange tree named 'M7'
3	PP19575	Orange tree named 'Alvarina'

Tabela 6.1: Patentes utilizadas para a execução de testes de similaridade

### 6.4.1.1 Experimento 1

Obtidas todas as patentes necessárias da base de dados *USPTO*, conforme apresentadas na Tabela 6.1, iniciou-se o processamento do método *LDA+EMD* para a análise de similaridade de patentes em relação ao documento base 4876196. Além disso, também foram testados os métodos de Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos com o intuito de comparação dos resultados dos métodos em relação ao *LDA*. A Tabela 6.2 mostra os resultados de similaridade obtidos para os métodos avaliados.

Grupo	Nº USPTO da patente	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
1	4876196	99,9762	100	100
1	4326036	87,3273	95,011406	34,9728
1	4560659	76,2738	95,18469	32,9674
1	4738930	71,239	94,689575	29,3144
2	5428528	59,6165	89,943085	7,0148
2	6039574	41,0027	93,511375	11,795
2	6416410	55,303	91,88554	9,9098
2	6478583	47,933	93,81358	12,4934
2	6500070	49,5798	88,272385	10,4904
2	7316618	35,4835	93,24559	4,7962
2	7371163	41,9381	91,861176	8,1337
3	PP04161	0,6145	93,663734	4,1367
3	PP06047	0,6211	93,88927	6,0575
3	PP07651	0,4214	93,96202	6,2545
3	PP07700	0,4587	93,97782	6,7322
3	PP08212	0,4762	92,91633	11,6983
3	PP08238	0,2877	94,07054	9,1945
3	PP11246	1,4609	94,08937	8,8773
3	PP18774	0,3774	93,95004	7,8252
3	PP19575	1,1695	93,77094	7,1547

Tabela 6.2: Grau de similaridade da patente 4876196 em relação aos documentos dos grupos de análise

Nessa tabela os resultados constituem os valores percentuais de similaridade que uma determinada patente possui em relação à patente principal 4876196, sendo esses calculados pelo método empregado na análise.

De maneira geral, por meio dos resultados obtidos pelos três métodos empregados para o teste de similaridade, pode-se afirmar que o método *LDA* prioriza o grupo tecnológico para a análise de similaridade dos documentos. Em muitos casos essa é uma característica

positiva, pois, quando uma patente pertence a um mesmo conjunto de patentes que o documento de análise, as chances de similaridade entre as duas patentes avaliadas são consideravelmente maiores devido à estrutura de tópicos que os documentos possuem em comum [12].

Nesse primeiro experimento, em que a patente base de avaliação é utilizada totalmente para a análise, basicamente todos os métodos obtiveram êxito em identificar as patentes pertencentes ao mesmo grupo do documento 4876196, considerando-as semelhantes à patente de avaliação.

Em Anexo 3 encontram-se as tabelas individuais de resultados para cada um dos métodos empregados no experimento, estando essas tabelas ordenadas em ordem decrescente de similaridade das patentes em relação ao documento 4876196. Também, em Anexo 4 estão os tópicos e as listas de tópicos retornados pelo método *LDA with Gibbs Sampling* para as patentes utilizadas no experimento.

#### 6.4.1.2 Experimento 2

A fim de simular experiências com modelos parciais de patentes, isto é, um documento com apenas partes do conteúdo de uma patente, empregou-se como documento principal de análise para o experimento 2 a mesma patente de número 4876196 utilizada no experimento 1 (Capítulo 6.4.1.1). Nesse caso foram usados apenas os textos dos campos *title*, *abstract*, *claims* e *descriptions* da patente *USPTO* 4876196 para se avaliar a similaridade de patentes em relação aos grupos apresentados anteriormente na Tabela 6.4.1. Por outro lado, para efeito de testes do modelo, ainda mantiveram-se intactos todos os demais documentos dos grupos de patentes para a realização da experiência. Os resultados obtidos pelos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF+Distância dos Cossenos* são apresentados na Tabela 6.3.

Grupo	Nº USPTO da patente	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
1	4876196	38,8139	92,96484	64,3583
1	4326036	35,3475	92,13364	28,5735
1	4560659	37,5935	92,15224	24,745
1	4738930	34,2908	91,14969	19,7982
2	5428528	18,0807	87,289116	2,2497
2	6039574	18,7151	92,48126	5,4162
2	6416410	19,5805	90,430214	3,4057
2	6478583	23,5409	92,59906	5,4625
2	6500070	24,3578	87,03087	4,7867
2	7316618	16,3367	89,45783	1,255
2	7371163	23,1022	90,4315	3,6806
3	PP04161	22,2429	88,48225	2,6787
3	PP06047	33,2007	89,14898	2,5143
3	PP07651	16,4518	90,53381	3,5932
3	PP07700	16,0205	89,900444	1,7068
3	PP08212	29,0465	90,91636	3,7005
3	PP08238	21,56	92,22324	3,3432
3	PP11246	33,1254	91,01645	5,7187
3	PP18774	21,355	90,94282	2,0732
3	PP19575	26,5089	90,751076	2,3941

Tabela 6.3: Grau de similaridade do modelo parcial de patente 4876196 em relação aos documentos dos grupos de análise

Nos resultados obtidos para o experimento, conforme mostram a tabela anterior, os métodos *LDA+EMD* e *TF-IDF+ Distância dos Cossenos*, da mesma forma que no experimento do Capítulo 6.4.1.1, conseguem manter praticamente o nível de similaridade dos documentos conforme o tema tecnológico dos subgrupos analisados. Em especial, ambos os métodos conseguiram retornar todas as patentes do conjunto relacionado ao tema etanol como sendo semelhantes em relação a patente de análise 4876196.

Por outro lado, observa-se claramente que o método Espaço Vetorial Binário+Distância Euclidiana apresenta problemas de similaridade para os documentos e, dessa forma, não se torna conveniente de ser empregado. Em Anexo 3 encontram-se as tabelas individuais dos métodos utilizados no experimento, organizadas em ordem decrescente de similaridade. Já em Anexo 4 dispõem-se das tabelas e listas de tópicos do método *LDA* das patentes empregadas.

### 6.4.2 Análise de grupos tecnológicos similares de patentes

Para se avaliar a aplicabilidade do método *LDA+EMD* para a similaridade de patentes provenientes de um mesmo campo tecnológico, assim como nos testes realizados anteriormente (Capítulo 6.4.1), os documentos adotados para a análise de patentes foram também obtidos da base de dados *USPTO*.

Para cada um dos testes propostos, utilizou-se de uma patente base de análise e um conjunto de 10 patentes com tecnologias relacionadas ao documento de análise. Dessa forma, com o intuito de garantir a semelhança tecnológica entre os documentos avaliados, o conjunto de patentes possui a característica peculiar de sempre citar a patente base, pois, essa fornece algum tipo de informação essencial para desenvolvimento tecnológico das patentes que a citam (Capítulo 3).

Assim, com a necessidade de se realizar uma análise mais precisa, minimizando possíveis problemas com dados irrelevantes como, por exemplo, os encontrados no primeiro nível da patente e informações a respeito dos autores (Capítulo 3.1), foram usados apenas os conteúdos dos campos *title*, *abstract*, *claims* e *description* para a análise dos documentos.

Novamente como no Capítulo 6.4.1, além do *LDA+EMD*, os métodos para a avaliação da similaridade foram as técnicas do Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos. No final do processo de cada um dos testes propostos, os resultados são sempre comparados aos de um especialista em patentes com experiência nas áreas tecnológicas que abrangem os documentos.

A avaliação dos resultados por meio do trabalho de apenas um especialista em patentes se deve a disponibilidade de profissionais dispostos a realizar a operação de análise que demanda uma grande quantia de tempo. Segundo as informações do profissional que realizou a tarefa de análise, apenas para os testes realizados para essa dissertação, o trabalho manual despendido ocupou de no mínimo três meses de trabalho com total dedicação de estudo e análise.

Para a obtenção manual dos resultados dos experimentos, o especialista em patentes, inicialmente, realiza uma análise minuciosa de todas as patentes do grupo tecnológico. Em seguida, essas patentes são classificadas em ordem crescente de similaridade em relação à patente base de análise (Anexo 2).

Ao fim, após os resultados ordenados em *Ranking*, calcula-se o coeficiente de correlação de postos de *Spearman* (*Spearman's rank correlation coefficient* ou *Spearman's rho*) para se verificar a eficiência dos métodos, em especial o *LDA*, em relação ao procedimento manual realizado pelo analista de patentes.

A Correlação de *Spearman*, também conhecido como Correlação de postos de *Spearman*, pode ser aplicada quando as variáveis são dispostas em forma ordinal, ou seja, em processo de *Ranking*. Não havendo vínculo entre as variáveis, o valor do coeficiente de *Spearman* ( $\rho$ ) pode ser dado conforme a Função 6.3:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (6.3)$$

Onde,  $n$  representa o número de pares de variáveis  $x_i$  e  $y_i$ .  $d_i$  é a diferença entre cada posto de valores, isto é,  $d_i = x_i - y_i$ .

Para os experimentos realizados com os grupos de patentes,  $x_i$  indica o valor da ordem de similaridade encontrada pelo analista para a patente  $i$  e  $y_i$  representa o valor da ordem de similaridade encontrada pelo método avaliado (nesse caso, *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana ou *TF-IDF*+Distância dos Cossenos). A Figura 6.2 representa a interpretação para o coeficiente de *Spearman* ( $\rho$ ).

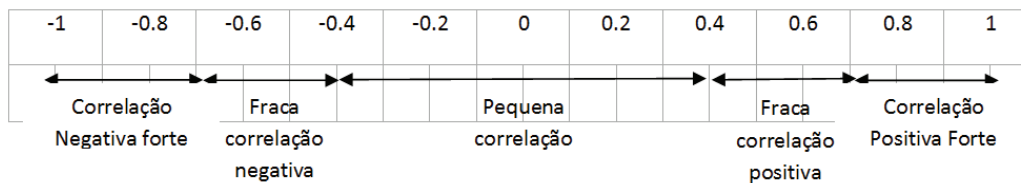


Figura 6.2: Interpretação do coeficiente de correlação de postos de *Spearman*

A seguir, os capítulos 6.4.2.1, 6.4.2.2 e 6.4.2.3 apresentam, respectivamente, os experimentos 1, 2 e 3 para a comparação dos resultados de análise de similaridade dos métodos computacionais em relação à análise manual feita pela analista de patentes.

Em Anexo 3 são apresentadas as tabelas de *Ranking* de similaridade resultante dos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos, além da classificação manualmente obtida pelo analista de patentes em cada um dos 3 experimentos realizados.

### 6.4.2.1 Experimento 1

Para o primeiro experimento, a patente base de avaliação é o documento de número *USPTO 4940835* (*Glyphosate-resistant plants*) que é relacionado a um método de clonagem de plantas por meio da compressão de um gene do polipeptídeo *EPSPS* (*5-enolpyruvylshikimate-3-phosphate synthase*), conferindo a planta certo grau de resistência ao herbicida glifosato.

A Tabela 6.4 apresenta os números das patentes segundo a base *USPTO* e seus respectivos títulos para o primeiro experimento realizado.



<b>Nº USPTO da patente</b>	<b>Título</b>
4940835	Glyphosate-resistant plants
6268550	Methods and a maize acetyl CoA carboxylase gene for altering the oil content of plants
6271016	Anthranilate synthase gene and method of use thereof for conferring tryptophan overproduction
6362396	Chimeric gene for the transformation of plants
6329574	High lysine fertile transgenic corn plants
6399861	Methods and compositions for the production of stably transformed, fertile monocot plants and cells thereof
6326527	Method for altering the nutritional content of plant seed
6331665	Insect resistant fertile transgenic corn plants
6395966	Fertile transgenic maize plants containing a gene encoding the pat protein
6306636	Nucleic acid segments encoding wheat acetyl-CoA carboxylase
6281411	Transgenic monocots plants with increased glycine-betaine content

Tabela 6.4: Experimento 1: Grupo tecnológico utilizado para os testes de similaridade

Logo a seguir, a Tabela 6.5 mostra a ordem de similaridade em ordem decrescente das patentes para o experimento, sendo essa obtida manualmente pelo analista em relação ao documento 4940835. Também nessa mesma tabela, apresenta-se uma breve descrição do conteúdo de cada um dos documentos segundo o examinador de patentes.

Ordem de similaridade	Nº USPTO da patente	Considerações
1	6268550	Planta GM; usa mesmo vetor, mas para tipo de controle genético diferente (óleo), ou seja, é muito semelhante, com aplicação de controle metabólico que é diferente.
2	6271016	Método de obtenção de tolerância
3	6362396	Gene quimérico diferente, mas processo semelhante de clonagem e vetor
4	6329574	Mesmo vetor (Streptomices), mas com aplicação em outra cultura e com outra enzima de resistência
5	6399861	vetor de Streptomices para milho
6	6326527	Plantas com conteúdo de amino-ácidos diversos; usa apenas vetores de genes
7	6331665	Plantas de milho resistente
8	6395966	Plantas de milho resistente
9	6306636	Genes do acetilCoA transferase
10	6281411	Resistência à seca em milho

Tabela 6.5: Experimento 1: Ordem decrescente de similaridade para a patente 4940835 segundo o analista de patentes

Realizado os experimentos de análise de similaridade, a próxima tabela (Tabela 6.6) mostra os resultados percentuais de similaridade obtidos pela utilização dos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos.

Nº USPTO da patente	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
6268550	94,9854	90,811966	19,1132
6271016	92,6304	92,159004	27,4192
6362396	96,7659	90,638084	25,0956
6329574	92,2692	93,50152	27,6931
6399861	86,693	93,749435	25,1944
6326527	92,4082	91,05985	25,6486
6331665	90,2505	93,70611	22,0186
6395966	87,4585	91,365776	12,448
6306636	91,6847	94,091324	29,5458
6281411	90,3339	92,77839	24,5911

Tabela 6.6: Experimento 1: Percentagem de similaridade das patentes em relação a patente de análise 4940835

Assim, calculando-se o coeficiente de correlação de postos de *Spearman* para a tabela de *Ranking* de similaridade do experimento (ver Anexo 3), obtém-se os valores de  $\rho$  conforme consta na Tabela 6.7:

Coeficiente de Spearman( $\rho$ )			
	<i>LDA+EMD</i>	Espaço Vetorial Binário +Distância Euclidiana	<i>TF-IDF</i> +Distância dos Cossenos
Experimento 1	$\rho = 0,6606$	$\rho = -0,503$	$\rho = 0,0303$

Tabela 6.7: Experimento 1: Valores do coeficiente de Spearman para os métodos aplicados

Por meio da análise dos valores de  $\rho$  para o experimento, observa-se que o algoritmo *LDA+EMD* obteve resultados com fraca correlação positiva, indicando que o mesmo apresenta resultados próximos aos do analista de patentes. Da mesma forma, conclui-se que o método Espaço Vetorial Binário+Distância Euclidiana apresenta resultados contrários aos do analista (fraca correlação negativa). Porém, para a aplicação do método *TF-IDF*+Distância dos Cossenos nada se pode inferir a respeito dos resultados, uma vez que apresenta pequena correlação em relação aos resultados do analista.

#### 6.4.2.2 Experimento 2

No segundo experimento, avaliou-se um grupo de documentos patentários pertencentes à mesma tecnologia da patente de número *USPTO 5107065* (*Anti-sense regulation of gene expression in plant cells*). A patente em análise 5107065 representa a tecnologia seminal para a transgenia, pois é a técnica de regulação da expressão de genes que possam ter sido alterados por engenharia genética (Anexo 2). A Tabela 6.8 mostra, além dos dados do documento em avaliação, o título das patentes utilizadas para o segundo experimento.

Nº USPTO da patente	Título
5107065	Anti-sense regulation of gene expression in plant cells
7754697	Control of gene expression
7723503	Desaturase genes, enzymes encoded thereby, and uses thereof
7759463	RNA interference pathway genes as tools for targeted genetic interference
7645925	Tobacco products with increased nicotine
7683237	Maize seed with synergistically enhanced lysine content
7700834	Nicotiana nucleic acid molecules and uses thereof
7700851	Tobacco nicotine demethylase genomic clone and uses thereof
7754869	Production of syringyl lignin in gymnosperms
7754942	Maize starch containing elevated amounts of actual amylose
7705203	Benzoate inductible promoters

Tabela 6.8: Experimento 2: Grupo tecnológico utilizado para os testes de similaridade

Da mesma forma que no experimento 1 (Capítulo 6.4.2.1), na Tabela 6.9 encontram-se os resultados alcançados manualmente pelo especialista em análise de patentes.

Ordem de similaridade	Nº USPTO da patente	Considerações
1	7754697	Patente muito similar, à principal, pois faz parte do pacote tecnológico da transgenia.
2	7723503	
3	7759463	Tecnologia de apoio e validação à principal; faz parte do pacote tecnológico da transgenia.
4	7645925	Cita a principal apenas como referência, pois é método de transgenia mais específico.
5	7683237	Cita a principal apenas como referência, pois é método de transgenia mais específico.
6	7700834	Patente de cunho metodológico.
7	7700851	Uso da tecnologia principal para transformar tabaco.
8	7754869	Controle de lignina em gimnospermas.
9	7754942	Tecnologia de resultado: milho com maior teor de amido.
10	7705203	Tecnologia parece abrir nova trajetória.

Tabela 6.9: Experimento 2: Ordem decrescente de similaridade para a patente 5107065 segundo o analista de patentes

Nesse experimento, conforme observações feitas pelo analista de patentes, observa-se que os documentos 7700851, 7754869 e 7754942 (ordem de classificação 7, 8 e 9) apre-

sentam similaridade muito grande entre si, sendo produtos de uso e aplicação da patente principal 5107065. Assim, para evitar a incoerência de valores empregados para o método *EMD*, adotaram-se para essas 3 patentes o valor de Ranking igual a 7 conforme é mostrado em Anexo 3. Os resultados obtidos pela utilização dos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos são mostrados na Tabela 6.10.

Nº USPTO da patente	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
7754697	34,5217	91,368126	17,9164
7723503	23,5154	92,82988	13,5048
7759463	35,2612	92,22149	15,9116
7645925	80,6589	92,909065	24,4005
7683237	62,1114	90,93185	9,5039
7700834	66,2092	91,36071	19,9691
7700851	55,8331	91,318405	19,3259
7754869	77,8123	92,128746	17,2146
7754942	7,2320	93,73094	15,523
7705203	53,1095	92,43569	23,8635

Tabela 6.10: Experimento 2: Percentagem de similaridade das patentes em relação a patente de análise 5107065

Na Tabela 6.11 são apresentados os valores do coeficiente de correlação de postos de *Spearman* para cada um dos 3 métodos testados no experimento. Em Anexo 3 encontram-se as tabelas de *Ranking* de similaridade para o cálculo de  $\rho$ .

Coeficiente de Spearman( $\rho$ )			
	<i>LDA+EMD</i>	Espaço Vetorial Binário +Distância Euclidiana	<i>TF-IDF</i> +Distância dos Cossenos
Experimento 2	$\rho = -0,2147$	$\rho = -0,0061$	$\rho = -0,2761$

Tabela 6.11: Experimento 2: Valores do coeficiente de Spearman para os métodos aplicados

Observando os valores de  $\rho$  encontrados para cada um dos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos, pode-se afirmar que ambos apresentaram os mesmos resultados (pequena correlação) e nada se pode afirmar sobre a eficiência dos três métodos.

### 6.4.2.3 Experimento 3

Por fim, para o terceiro experimento são utilizados documentos relacionados à patente de número 5352605 (*Chimeric genes for transforming plant cells using viral promoters*). Na Tabela 6.12 encontram-se as informações das patentes empregadas para o teste e na Tabela 6.13 os resultados obtidos pelo analista de patentes.

Nº USPTO da patente	Título
5352605	Chimeric genes for transforming plant cells using viral promoters
7183110	Antibody immunoreactive with a 5-enolpyruvylshikimate-3-phosphate synthase
7202083	Plant promoters and plant terminators
7189570	Putrescine-n-methyltransferase promoter
7192771	Plant promoter sequence
7161064	Method for producing stably transformed duckweed using microprojectile bombardment
7135626	Soybean seeds and plants exhibiting natural herbicide resistance
7087809	Natural herbicide resistance in wheat
6777546	Methods and substances for preventing and treating autoimmune disease
7135282	Viral particles with exogenous internal epitopes
7074987	Genetically-controlled herbicide resistance in cotton plants in the absence of genetic engineering

Tabela 6.12: Experimento 3: Grupo tecnológico utilizado para os testes de similaridade

Ordem de similaridade	Nº USPTO da patente	Considerações
1	7183110	Patente muito similar à principal, pois faz parte do pacote tecnológico da transgenia.
2	7202083	Patente muito similar à principal, pois faz parte do pacote tecnológico da transgenia.
3	7189570	Patente muito similar à principal, pois faz parte do pacote tecnológico da transgenia.
4	7192771	Patente muito similar à principal, pois faz parte do pacote tecnológico da transgenia.
5	7161064	Tecnologia de apoio e validação à principal. Faz parte do pacote tecnológico da transgenia.
6	7135626	Tecnologia de apoio e validação à principal. Faz parte do pacote tecnológico da transgenia.
7	7087809	Tecnologia de apoio e validação à principal. Faz parte do pacote tecnológico da transgenia.
8	6777546	Aplicação ímpar de modificação de plantas; deve citar a principal apenas porque esta última é referência para tudo o que for planta modificada geneticamente.
9	7135282	Nexos mais distantes com a principal, pois os vírus são também utilizados como vetores de transgenia.
10	7074987	Método que usa tecnologia nova e que pretende dispensar a tecnologia principal.

Tabela 6.13: Experimento 3: Ordem decrescente de similaridade para a patente 5352605 segundo o analista de patentes

Assim como no experimento 2 (Capítulo 6.4.2.2), as patentes classificadas pelo analista de patentes com *Ranking* 1, 2, 3, 4, 5 e 6 de similaridade em relação a patente base de análise 5352605 são muito semelhantes em termos tecnológicos e podem receber ordem de classificação diferentes conforme a opinião do profissional que realizar a tarefa. Assim, para contornar esse problema, o método *EMD* utiliza valores de Ranking igual a 1 para essas 6 patentes conforme podem ser observadas em Anexo 3. Os resultados obtidos pelos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos são mostrados na tabela a seguir (Tabela 6.14).

N° USPTO da patente	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
7183110	94,832	93,31648	20,359
7202083	91,135	92,76183	15,1138
7189570	91,545	93,64175	17,3834
7192771	89,734	92,86345	10,5188
7161064	92,433	93,28946	17,3805
7135626	96,456	90,418434	14,9256
7087809	96,335	90,595856	12,7283
6777546	36,433	93,57669	13,6431
7135282	95,052	92,70391	23,1535
7074987	92,577	88,84655	8,6738

Tabela 6.14: Experimento 3: Percentagem de similaridade das patentes em relação a patente de análise 5352605

Após a obtenção dos resultados em *Ranking* (Anexo 3), calculam-se os valores de  $\rho$  para cada um dos 3 métodos utilizados no experimento. A tabela abaixo (Tabela 6.15) apresentam os valores dos coeficientes de *Spearman*.

Coeficiente de Spearman( $\rho$ )			
	<i>LDA+EMD</i>	Espaço Vetorial Binário +Distância Euclidiana	<i>TF-IDF</i> +Distância dos Cossenos
Experimento 3	$\rho = -0,1229$	$\rho = 0,4165$	$\rho = 0,2731$

Tabela 6.15: Experimento 3: Valores do coeficiente de Spearman para os métodos aplicados

Da mesma maneira que os resultados obtidos no experimento 2 (Capítulo 6.4.2.2), a partir dos valores de  $\rho$  conclui-se que os algoritmos empregados no experimento (*LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos) possuem resultados muito semelhantes entre si, indicando, dessa forma, que todos os métodos avaliados não possuem uma posição bem clara em relação aos resultados obtidos pelo analista de patentes.

## 6.5 Discussão dos Resultados

Através dos resultados obtidos para os 2 grupos de análise – Grupos tecnológicos distintos e Grupos tecnológicos similares de patentes – empiricamente percebe-se a possibilidade de utilização do método *LDA* para a análise de similaridade de patentes em relação aos demais métodos avaliados. O método fornece, de maneira geral, resultados semelhantes ou



mesmo melhores em relação aos métodos do Espaço Vetorial Binário e *TF-IDF* conforme constatado no Capítulo 6.4.

Logo a seguir, capítulos 6.5.1 e 6.5.2, são discutidos respectivamente os resultados das experiências realizadas para os grupos tecnológicos distintos e similares de patentes.

### 6.5.1 Avaliação para grupos tecnológicos distintos de patentes

Observando os resultados obtidos pelos três métodos para a avaliação de grupos distintos de patentes (Capítulo 6.4.1 e Anexo 3), pode-se concluir que o método *LDA+EMD* procura priorizar grupos de patentes para a análise de similaridade. Essa característica se torna positiva para a operação de similaridade de documentos, pois, para uma patente que pertence ao mesmo conjunto que o documento de análise, as chances de similaridade entre as patentes de mesmo grupo são consideradas maiores. A razão disso se deve normalmente a estrutura de tópicos que ambos os documentos patentários possuem em comum [12].

Para o primeiro experimento, no qual a patente base de avaliação escolhida aleatoriamente de um dos grupos é utilizada totalmente para a análise, conclui-se que basicamente todos os métodos obtiveram êxito em identificar as patentes pertencentes ao mesmo grupo do documento 4876196, considerando-as então semelhantes a patente de avaliação.

Porém, no segundo experimento, em que apenas partes da patente de análise são utilizadas (Tabela 6.3), o método do Espaço Vetorial Binário+Distância Euclidiana claramente apresenta problemas de classificação e, dessa forma, não se torna conveniente de ser empregado para experimentos com conteúdos variados de patentes. Em Anexo 4 encontram-se as tabelas e listas de tópicos do método *LDA* das patentes obtidas para os dois experimentos realizados no Capítulo 6.4.1.

### 6.5.2 Avaliação para grupos tecnológicos similares de patentes

Por meio dos resultados dos experimentos que utilizam grupo de patentes similares para avaliação (Capítulo 6.4.2), pode-se calcular o coeficiente de Spearman com o intuito de se verificar a eficiência dos métodos em relação ao procedimento manual realizado pelo analista de patentes. A Tabela 6.5.2.1 apresenta os valores de  $\rho$  para os métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos dos 3 experimentos realizados.

Coeficiente de Spearman( $\rho$ )			
	<i>LDA+EMD</i>	Espaço Vetorial Binário +Distância Euclidiana	<i>TF-IDF</i> +Distância dos Cossenos
Experimento 1	$\rho = 0,6606$	$\rho = -0,503$	$\rho = 0,0303$
Experimento 2	$\rho = -0,2147$	$\rho = -0,0061$	$\rho = -0,2761$
Experimento 3	$\rho = -0,1229$	$\rho = 0,4165$	$\rho = 0,2731$

Tabela 6.16: Valores do coeficiente de correlação de postos de *Spearman* para os métodos aplicados

A partir dos valores de  $\rho$  obtidos para os experimentos, observa-se que os resultados do método *LDA+EMD* apresentam valores de correlação variando de uma pequena correlação a uma fraca correlação positiva (ver Figura 6.2).

Embora o ideal para se medir a eficiência do método seja encontrar valores de  $\rho$  pertencentes a uma forte correlação positiva, nem sempre isso se torna possível. Para o primeiro experimento (Capítulo 6.4.2.1), o algoritmo *LDA+EMD* obteve resultados próximos aos do analista de patentes, diferindo dos resultados encontrados no segundo (Capítulo 6.4.2.2) e terceiro (Capítulo 6.4.2.3) experimentos, em que nem sempre os métodos computacionais empregados nos experimentos podem afirmar alguma coisa sobre a ordenação dada pelo analista. Isso se deve, conforme observações feitas pelo analista de patentes, ao fato de que nem sempre é possível classificar patentes segundo uma ordem de similaridade e variações de ordenação se tornam inevitáveis conforme o profissional que realiza tal tarefa (Anexo 2).

Assim, em um experimento realizado à parte para a patente de número *USPTO* 5968830 e um grupo formado pelas patentes 7777104, 7777103, 7759553, 7759552, 7759551, 7732671, 7728204, 7728202 e 7728201 (ver Anexo 2), observa-se que as 4 primeiras patentes do grupo (todas pertencentes à empresa Monsanto<sup>9</sup> e abrangendo o assunto cultivares de soja desenvolvidas com base em técnicas de transgenia por *Agrobacterium*) são igualmente similares, sendo consideradas de difícil classificação em relação ao documento 5968830.

Caso semelhante pode ser observado para o experimento 2 (Capítulo 6.4.2.2) em que, segundo o analista de patentes, os documentos 7700851, 7754869 e 7754942 apresentam um grau de semelhança muito grande entre si, sendo patentes de aplicação da patente principal 5107065.

Também, em experimento 3 (Capítulo 6.4.2.3), essa mesma característica de similaridade pode ser observada. As patentes classificadas pelo analista como 1, 2, 3, 4, 5 e 6 são muito similares em termos tecnológicos e podem receber *Rankings* diferentes conforme a opinião do profissional que realizar a tarefa de *Ranking*.

Por meio dos resultados dos 3 experimentos, conclui-se que a utilização do método

<sup>9</sup>Monsanto Company. Disponível em: <http://www.monsanto.com.br/>

*LDA+EMD* para a análise de similaridade de patentes oferece resultados compatíveis aos métodos do Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos. Em casos como o do primeiro experimento, em que a ordem de similaridade se encontra mais evidente para o analista de patentes, o método *LDA* oferece resultados melhores em relação às demais técnicas testadas. Além disso, conforme observado no Capítulo 6.4.1, o método procura sempre desempenhar o processo de similaridade dando prioridade aos tópicos aos quais pertencem as patentes.



# Capítulo 7

## Trabalhos Relacionados

Atualmente o modelo de tópicos é bastante utilizado para o procedimento de representação e a classificação de coleções de documentos [51] [2], não possuindo ênfase específica para o processo de similaridade de documentos. Além disso, por meio da investigação de técnicas de similaridade de textos, observa-se que não existem muitos estudos especificamente focados em documentos patentários. A maioria dos sistemas prioriza o processo de recuperação e classificação de patentes, oferecendo pouco ou nenhum destaque ao relacionamento existente entre patentes similares.

Desenvolvido por *Larkey* [29], um dos primeiros aplicativos para a recuperação de informações patentárias possuía como ênfase procedimentos de recuperação e classificação de patentes. O *software*, basicamente, retorna um conjunto de documentos no padrão *USPTO* pertencentes a uma base dados de 401 coleções de patentes conforme os termos de busca estabelecidos pelo usuário. Para isso, definido o termo de busca, o sistema oferece a possibilidade de *query expansion* para o usuário, sendo possível acrescentar novos termos relacionados a *query de busca* inicial. A partir de então, o sistema utiliza o algoritmo *TF-IDF* para encontrar os conjuntos de documentos relacionados a *query de busca*. Todos os grupos de patentes possuem a característica de estarem previamente agrupados na base de dados conforme a classe *USPTO* do documento.

Ainda relacionada à recuperação de patentes, *Inoue et al* [20] desenvolveram um sistema que utiliza técnicas de filtragem de documentos baseadas em modelos probabilísticos para retornar patentes ao usuário. Nesse sistema, o usuário necessita cadastrar uma espécie de *perfil de usuário* para que, a cada adição de uma nova patente à base de dados, o sistema calcule o grau de similaridade entre a patente e o perfil cadastrado do usuário. Para isso, *Inoue* utiliza o modelo probabilístico *Iwayama's formulation* quando existem apenas poucos documentos a serem avaliados ou, caso contrário, o método de *cluster-based search*.

Outro exemplo de implementação é o *PRIME* [17] que também possui como obje-

tivo recuperar patentes de uma base de dados. No entanto, esse sistema difere ao de *Larkey* pelo fato de que a busca é efetuada em banco de dados distintos e automaticamente se realizam traduções dos documentos retornados para o idioma escolhido pelo usuário (*Cross-language information retrieval - CLIR*). O *PRIME* utiliza as bases de dados *USPTO* e *JPO*, mais especificamente, os dados provenientes de *US Patent Application* e *Patent abstract of Japan*. Dessa forma, para a obtenção das patentes, inicialmente a *query de busca* é traduzida para o idioma desejado (neste caso, o inglês ou o japonês) por meio do método de tradução de *query* proposto por *Fuji* e *Ishikawa*. Além disso, emprega-se *Nova dictionary* para evitar ambiguidade de palavras e auxiliar na tradução de termos técnicos. A partir de então, a busca é realizada nos dois bancos de dados e os documentos são retornados conforme o grau de relevância do método probabilístico baseado em *TF-IDF* proposto pelos autores. Assim, os resultados obtidos de ambas as bases de dados são automaticamente traduzidos para o idioma escolhido pelo usuário, empregando-se as mesmas técnicas da fase de tradução de *query*. Por fim, com o intuito de uma melhor organização das informações, o sistema utiliza o método *Hierarchical Bayesian Clustering (HBC)* para os documentos e, com isso, determinam-se os grupos de patentes mais representativos em relação a *query de busca* inicialmente dado pelo usuário.

Também, a partir dos resultados satisfatórios alcançados pelo *PRIME*, os mesmos autores desenvolveram uma nova extensão para o sistema, englobando o idioma coreano [35]. Testes mostraram que, para esse novo módulo, o processo de recuperação de patentes necessita ser melhor explorado, porém, percebe-se que com os resultados alcançados o sistema já se torna plausível para aplicação.

Por outro lado, baseando-se em técnicas de clusterização de texto, *Lee et al* [30] desenvolveram um sistema para buscar patentes similares que se encontram em idiomas diferentes. Nesse caso, os autores utilizaram patentes em chinês e inglês nas áreas de semicondutores e óptica. Nesse processo, os documentos são mapeados na forma de espaço vetorial baseado em *Latent Semantic Indexing (LSI)* e, posteriormente, aplicam-se os métodos *Hierarchical Agglomerative Clustering* e *Self-organizing Maps* para avaliar a similaridade das patentes.

Também, possuindo como foco a análise de similaridade, *Wu et al* [50] propõem um método que utiliza links de citações diretas e indiretas de patentes contidas em uma rede de citações. Nessa pesquisa, inicialmente para todas as patentes, calcula-se o grau de similaridade do documento em relação a patente que recebe uma citação direta da mesma através da aplicação do algoritmo *Cosine Similarity*. Em seguida, calcula-se a matriz de similaridade indireta e, por fim, a matriz de similaridade composta. Para o algoritmo, os autores empregam de dois critérios para validação: o critério empírico de que a similaridade entre duas patentes diminui à medida que se aumenta o ano de concessão e o critério de classificação *UPC (Universal Product Code)*.

# Capítulo 8

## Conclusão

Esse trabalho teve como abordagem a utilização do método *Latent Dirichlet Allocation (LDA)* conjuntamente com o algoritmo *Earth Mover's Distance (EMD)* para a análise de similaridade de patentes. Para isso, desenvolveu-se um sistema computacional denominado *Odysséys* que, além de buscar e classificar em processo de *Ranking* um conjunto de patentes em relação ao documento de análise, procura construir uma rede de citações de patentes. A rede de citações possui como objetivo compreender o relacionamento existente entre as patentes e, também, avaliar a prospecção da tecnologia contida nos documentos por meio do fluxo de conhecimento entre as patentes.

Para o efeito de verificação da usabilidade do método *LDA* no processo de análise de documentos patentários, as técnicas da representação vetorial binária utilizando a Distância Euclidiana e *TF-IDF* empregando a Distância dos Cossenos também foram testados e comparados.

Na fase de testes, realizaram-se inúmeros experimentos com grupos similares de documentos à patente de análise. Com isso, os resultados obtidos pela utilização dos métodos (*LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos) foram colocados em processo de *Ranking* e, então, comparados aos de um especialista em análise de patentes.

De maneira geral, por meio da avaliação dos experimentos, pode-se concluir que o método *LDA* obteve resultados empíricos compatíveis ao campo tecnológico aos quais pertencem as patentes avaliadas (Capítulo 6.4.1.1). Além disso, observa-se também que o método proposto permitiu uma classificação de similaridade eficiente mesmo em circunstâncias em que somente se utilizam determinadas partes do documento principal para a análise (Capítulo 6.4.1.2).

No entanto, conforme observado ao se estudar detalhadamente o conteúdo das patentes (Anexo 2), nem sempre o processo de análise atribuindo níveis de similaridade às patentes se torna algo simples de ser efetuado. Patentes podem ser classificadas como igualmente

similares e também receber uma classificação diferente conforme o profissional que realizar a avaliação dos documentos. Esse fato pode ser também observado no (Capítulo 6.4.1.1), em que para todos os métodos avaliados (*LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos), nenhum obteve resultados os quais sejam possíveis de inferir a eficiência dos métodos em relação a ordenação dada pelo analista de patentes.

Assim, para trabalhos futuros, podem-se utilizar outras formas adicionais para melhorar a avaliação do grau de similaridade entre as patentes como, por exemplo, o uso de links de citações e o emprego de *IPC* ou a classe *USPTO* presente nas patentes [50].

Alguns trabalhos utilizando o sistema *Odysseys* para a análise de redes de citações são apresentados nos artigos [5] e [6]. Nessas pesquisas voltadas ao programa de Bioenergia no Brasil (*BIOEN*)<sup>1</sup>, as redes de citações de patentes *USPTO* de determinadas tecnologias, neste caso etanol e plantas transgênicas, são usadas para avaliar a prospecção e trajetórias do mercado de inovação.

---

<sup>1</sup>BIOEN/FAPESP. Disponível em: <http://bioenfapesp.org/>



# Referências Bibliográficas

- [1] Ng A.Y. Jordan M.I. Blei, D.M. Latent dirichlet allocation. *In The Journal of Machine Learning Research*, 3:993–1022, 2003.
- [2] Löffström J. Perttu S. Valtonen K. Buntine, W. Topic-specific scoring of documents for relevant retrieval. *In 22nd International Conference on Machine Learning*, pages 34–41, 2005.
- [3] Aygun E. Cataltepe, Z. An improvement of centroid-based classification algorithm for text classification. *In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 952–956, 2007.
- [4] Decreto da Convenção de Berna para a proteção das Obras Literárias e Artísticas, 2013. Disponível em: [www.cultura.gov.br/site/wp-content/uploads/2007/10/decreto-75699.pdf](http://www.cultura.gov.br/site/wp-content/uploads/2007/10/decreto-75699.pdf).
- [5] Silveira J.M. Masago F. K. Dal Poz, M. E. Bioenergy brazilian program (bioen) innovation networks. *In Triple Helix Conference VIII*, pages 1–16, 2010.
- [6] Silveira J.M. Masago F. K. Dal Poz, M. E. Bioenergy brazilian program (bioen) innovation networks. *In 15th ICABR Conference on Sustainability and the Bioeconomy*, pages 1–26, 2011.
- [7] Núcleo de Estudos Internacionais do Largo São Francisco, 2013. Material de curso da WIPO/OMPI. Disponível em: <http://www.nei-arcadas.org/>.
- [8] Curso de Propriedade Intelectual e Busca em Bases de Patentes, 2009. Curso a Distância oferecido pela INOVA-UNICAMP. Campinas, SP.
- [9] Rees M. L. Townshend B. Feldman, R. P. The effect of industry standard setting on patent licensing and enforcement. *In IEEE Communications Magazine*, 38:112–116, 2000.
- [10] Strehl A. Ghosh, J. Similarity-based text clustering: A comparative study. *In Grouping Multidimensional Data*, Springer Berlin Heidelberg, pages 73–97, 2006.

- [11] L. Graham. Act quickly to avoid losing patents. *In IEEE Software*, 16(2):33–35, 1999.
- [12] Steyvers M. Griffiths, T. L. Finding scientific topics. *In Proceedings of the National Academy of Sciences*, 101(1):5228–5235, 2004.
- [13] T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. *Stanford University. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.3760>*., pages 1925–1945, 2002.
- [14] Jaff A. Trajtenberg M. Hall, B. Market value and patent citations: A first look. *In National Bureau of Economic Research Working*, (7741), 2000.
- [15] Klavans J. L. Eskin E. Hatzivassiloglou, V. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. *In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, 1999.
- [16] Kindermann J. Lauth C. Paaß G. Monzon J. S. Heinrich, G. Investigating word correlation at different scopes - a latent topic approach. *In Workshop Learning and Extending Lexical Ontologies at International Conference on Machine Learning*, pages 16–22, 2005.
- [17] Fukui M. Fujii A. Ishikawa T. Higuchi, S. Prime: A system for multi-lingual patent retrieval. *In Proceedings of MT Summit VIII*, pages 163–167, 2001.
- [18] Wang A. Hung, S. A small world in the patent citation network. *In IEEE International Conference on Industrial Engineering and Engineering Management*, pages 1–5, 2008.
- [19] Ambekar A. A. Sureka A. Indukuri, K. V. Similarity analysis of patent claims using natural language processing techniques. *In Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, 4:169–175, 2007.
- [20] Matsumoto K. Hoashi K. Hashimoto K. Inoue, N. Patent retrieval system using document filtering techniques. *In Proceedings of the Workshop on Patent Retrieval, ACM SIGIR*, 2000.
- [21] Inkpen D. Islam, A. Semantic text similarity using corpus-based word similarity and string similarity. *In ACM Transactions on Knowledge Discovery from Data*, 2:1–25, 2008.

- [22] Lerner J. Jaffe, A. B. Patent prescription: A radical cure for the ailing u.s. patent system. *In IEEE Spectrum Magazine*, 42(12):38–43, 2004.
- [23] Wanlong L. Jing, S. Topic discovery based on lda model with fast gibbs sampling. *In International Conference on Artificial Intelligence and Computational Intelligence (AICI 2009)*, 3:91–95, 2009.
- [24] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *In Journal of Documentation*, 28:11–21, 1993.
- [25] Risov M Kasravi, K. Patent mining - discovery of business value from patent repositories. *In Proceedings of the Fortieth Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 1–10, 2007.
- [26] Risov M. Kasravi, K. Multivariate patent similarity detection. *In 42nd Hawaii International Conference on System Sciences*, pages 1–8, 2008.
- [27] M. M. Klee. Where did the u.s. laws come from? *In IEEE Engineering in Medicine and Biology Magazine*, 17:135–139, 1998.
- [28] M. M. Klee. Patents – the u.s. patent that reached around the world. *In IEEE Engineering in Medicine and Biology Magazine*, 25(3):76, 2006.
- [29] L. S. Larkey. A patent search and classification system. *In Proceedings of DL-99, 4th ACM Conference on Digital Libraries (Berkeley, USA)*, pages 179–187, 1999.
- [30] Yang H. Li Y. Lee, C. Development of a multilingual text mining approach for knowledge discovery in patents. *In Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 2265–2269, 2009.
- [31] Bickel P. Levina, E. The earth mover’s distance is the mallows distance: Some insights from statistics. *In Proceedings of the Eighth IEEE International Conference on Computer Vision*, 2:251–256, 2001.
- [32] Rafols I. Leydesdorff, L. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citation. *In Journal of Informetrics*, 5:87–100, 2010.
- [33] Yamanishi K. Li, H. Topic analysis using a finite mixture model. *In Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 13:35–44, 2000.
- [34] Stella F. Faini M. Magatti, D. A software system for topic extraction and document classification. *In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 1:283–286, 2009.

- [35] Higuchi S. Fujii A. Ishikawa T. Makita, M. A system for japanese/english/korean multilingual patent retrieval. *In Proceedings of Machine Translation Summit IX*, pages 475–478, 2003.
- [36] Dumais S. Meek C. Metzler, D. Similarity measures for short segments of text. *In Proceedings of the 29th European Conference on Information Retrieval Research (ECIR 2007)*, 4425:16–27, 2007.
- [37] R. R. Michaud. Provisional patents solve inventors' problems. *In Bioengineering Conference. Proceedings of the IEEE 30th Annual Northeast*, 30:245–246, 2004.
- [38] Corley C. Strapparava C. Mihalcea, R. Corpus-based and knowledge-based measures of text semantic similarity. *In Proceedings of the 21st National Conference on Artificial Intelligence*, 1:775–780, 2006.
- [39] Mrvar A. Batagelj V. Nooy, W. *Exploratory Social Network Analysis with Pajek*. Cambridge: Cambridge University Press, 1st edition, 2005.
- [40] Newman D. Ihler A. Asuncion A. Smyth P. Welling M. Porteous, I. Fast collapsed gibbs sampling for latent dirichlet allocation. *In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 569–577, 2008.
- [41] J. Ramos. Using tf-idf to determine word relevance in document queries. *In First International Conference on Machine Learning*, pages 1–4, 2003.
- [42] Chemudugunta C. Griffiths T. Smyth P. Steyvers M. Rosen-Zvi, M. Learning author topic models from text corpora. *In ACM Transactions on Information Systems*, 10(1):1–38, 2010.
- [43] Hocayen-da-silva A. J. Ferreira Júnior I. Rossoni, L. Aspectos estruturais da co-operação entre pesquisadores no campo de administração pública e gestão social: Análise das redes entre instituições no brasil. *In Revista Brasileira de Administração Pública*, 42:1041–1067, 2008.
- [44] Tomasi-C.-Guibas L.J. Rubner, Y. A metric for distributions with applications to image databases. *In Proceedings of the 1998 IEEE International Conference on Computer Vision*, pages 59–66, 1998.
- [45] Buckley-C. Salton, G. Term weighting approaches in automatic text retrieval. *In Information Processing and Management*, 24:513–523, 1988.

- [46] Veni-R. Taghva, K. Effects of similarity metrics on document clustering. *In Proceedings of the 2010 Seventh International Conference on Information Technology: New Generations*, pages 222–226, 2010.
- [47] Newman-D. Welling M. Teh, Y. W. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *In Advances in Neural Information Processing Systems*, 19:1353–1360, 2007.
- [48] M. Trajtenberg. A penny for your quotes: Patent citations and the value of innovations. *In The RAND Journal of Economics*, 21(1):172–187, 1990.
- [49] Peng Y. Wan, X. The earth mover’s distance as a semantic measure for document similarity. *In International Conference on Information and Knowledge Management - CIKM*, pages 301–302, 2005.
- [50] Chen H.-Lee K. Liu Y. Wu, H. A method for assessing patent similarity using direct and indirect citation links. *In IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 149–152, 2010.
- [51] Y. Xu. Apply text minig in analysis of patent document. *In Computer-Aided Industrial Design and Conceptual Design. IEEE 10th International Conference*, pages 2350–2352, 2009.



# Glossário

**ACORDO TRIPS** – Acordo Relativo aos Aspectos do Direito da Propriedade Intelectual Relacionados com o Comércio. Veja ADPIC.

**ADPIC** – Acordo Relativo aos Aspectos do Direito da Propriedade Intelectual Relacionados com o Comércio. Também conhecido como acordo TRIPs, refere-se a um tratado internacional relacionado com direitos da propriedade intelectual e comércio multilateral assinado no ano de 1994 que encerrou a rodada Uruguai do *Acordo Geral de Tarifas e Troca (GATT)*. Órgão administrado pela Organização Mundial do Comércio.

**ALBIE'S FOODS VS SMUCKER CO** – Disputa patentária na corte americana entre as empresas *Albie's foods* e *Smucker* para o processo de preparo do popular sanduíche de manteiga de amendoim e geléia.

**ALBIE'S FOODS** – Empresa de tortas de carne fundada em 1987 em Detroit, *EUA*. Trabalha com a venda de vários produtos como tortas, pães recheados, *calzones* e sanduíches de manteiga de soja e geléia de uva.

**ALGORITMO** – Caracterizado como uma sequência de passos necessários para resolver um problema de forma procedural a partir de padrões e regras.

**BAG OF WORDS** – Modelo de representação simplificada em que textos de um documento são representados como uma coleção não ordenada de palavras, desconsiderando a gramática e a ordem das palavras do texto.

**BIBLIOTECA ALGORÍTMICA** – Coleção de subprogramas que possuem código e dados auxiliares, provendo serviços a programas independentes. Permite o compartilhamento e a alteração de código e dados de forma modular.

**BIBLIOTECA DE CÓDIGO ABERTO** – Também conhecida como biblioteca *Open Source*, assim como a biblioteca algorítmica, é uma coleção de subprogramas que possuem código e dados auxiliares. Essa biblioteca é de utilização e contribuição

livre, seja no desenvolvimento, correção de erros ou documentação, contanto que a condição de liberdade seja mantida.

**BITMAP** – Também conhecido como *Windows Bitmap (BMP)*, é um formato de gráficos por mapa de bits (*raster*). Armazena imagens em pequenos quadrados chamados de pixels. Quanto maior o número de *pixels*, maior a qualidade da imagem.

**CADEIAS DE MARKOV** – Nome dado por *Andrey Markov*, no modelo matemático de transições de estados entre um conjunto finito de estados, o modelo baseia-se no fato de que os estados anteriores são irrelevantes para a predição dos estados seguintes, desde que o estado atual seja conhecido.

**CLASSIFICADOR** – Técnicas computacionais para o processo de discriminação entre classes ou categorias utilizadas para entender os dados existentes e prever como os mesmos irão se comportar.

**CLUSTERING** – Técnica de *Data Mining* para se realizar agrupamentos de dados conforme o grau de semelhança entre os mesmos. O critério de semelhança depende do problema em averiguação e do algoritmo empregado.

**CODE OF FEDERAL REGULATIONS (CRF)** – Codificação das regras gerais e permanentes publicados no registro federal pelo departamento executivo e agências do governo federal. É dividido em 50 *títulos* que representam vastas áreas sujeitas ao regulamento federal.

**COLLAPSED GIBBS SAMPLING** – Também conhecido como *Gibbs Sampling*, consiste em um conjunto de técnicas iterativas para amostrar valores de distribuições complexas conforme a probabilidade condicional dada a todas as outras variáveis.

**COMMON LAW** – Termo derivado do inglês *direito comum*, é o direito que se desenvolveu por meio das decisões dos tribunais e não mediante atos legislativos ou executivos. Nesse sistema, os juízes possuem a autoridade para criar o direito, estabelecendo um precedente.

**CONVENÇÃO DE BERNA** – Também conhecida como Convenção da União de Berna, estabelece uma união internacional para a proteção dos direitos de autor em obras literárias e artísticas. Incorpora o princípio do tratamento nacional, onde os autores pertencentes a um dos países da união usufruirão em outras nações, para as suas obras, dos direitos das suas respectivas leis. O tratado ocorreu na cidade de Berna, Suíça, em 1886.



**CONVENÇÃO DE PARIS (CUP)** – Acordo internacional relativo à Propriedade Intelectual, assinado em 1883 em Paris, para a Proteção da Propriedade Industrial.

**CORPUS-BASED SIMILARITY METHODS** – Ver o termo Métodos Baseados em Corpus.

**CORRELAÇÃO DE POSTOS DE SPEARMAN** – Também conhecido em inglês como *Spearman's rank correlation coefficient* ou *Spearman's rho*, é uma medida de correlação não-paramétrica que avalia uma função monótona arbitrária, sem fazer suposições sobre a distribuição de frequências das variáveis.

**COURT OF APPEALS FOR THE FEDERAL CIRCUIT** – Suprema corte federal dos Estados Unidos, localiza-se na cidade de Washington. Foi criado pelo Congresso com a aprovação da lei de 1982, que juntou o tribunal de Alfândegas e patentes.

**CULTIVAR** – Conforme a legislação brasileira cultivar é a variedade de qualquer gênero ou espécie vegetal que seja claramente distinguível de outras conhecidas. São protegidas pela Lei nº 9.456, de 25 de abril de 1997, regulamentada pelo Decreto nº 2.366, de 5 de novembro de 1997. O Ministério da Agricultura e Abastecimento fica encarregado de efetuar os registros por meio do Serviço Nacional de Proteção de Cultivares (SNPC).

**DESENHO INDUSTRIAL** – Definido como qualquer forma ou característica de um objeto apresentando um novo aspecto visual e, dessa forma, servindo como um tipo de fabricação industrial. A proteção do desenho industrial oferece direitos exclusivos para o seu titular, excluindo terceiros de produzir, comercializar ou importar produtos que possuam ou contenham cópias do modelo protegido.

**DIREITO AUTORAL** – Direito que protege trabalhos nas áreas da literatura, teatro, música, dança, filmes, pinturas, esculturas e outros trabalhos visuais de arte como programas de computador (*softwares*). O direito autoral protege a expressão de ideias e reserva para seus autores o direito exclusivo de reprodução da obra.

**DIREITO DE PROPRIEDADE INTELECTUAL (DPI)** – É o direito de exploração da propriedade Intelectual, isto é, exploração do bem imaterial, intangível, fruto da criatividade humana. O direito de propriedade intelectual propõe modalidades de proteção separadas em três categorias: Direito Autoral, Propriedade Industrial e *Proteção Sui Generis*.

**DIREITOS CONEXOS** – Direitos reservados às obras intelectuais previamente criadas, referindo-se à difusão criativa destas obras. Incluem elementos criativos da sua

personalidade (como no caso dos artistas intérpretes e executantes) ou através da tecnologia (produtores, emissoras de televisão).

**DISTRIBUIÇÃO CONJUNTA** – Em inglês *Joint probability distribution*, ou também conhecida como Distribuição de Probabilidade Conjunta, na estatística se refere a uma distribuição de probabilidades em que um conjunto de variáveis aleatórias pertence a uma faixa específica ou conjunto discreto de valores especificados.

**DISTRIBUIÇÃO PROBABILÍSTICA MULTINOMIAL** – Na teoria das probabilidades, a distribuição multinomial é uma generalização da distribuição binomial em que para  $n$  ensaios independentes e  $k$  categorias, a distribuição resulta na probabilidade de qualquer combinação específica de números de sucesso para as diferentes categorias.

**DISTÂNCIA DOS COSSENOS** – Em inglês *Cosine Distance*, é uma métrica utilizada para se calcular o grau de similaridade entre dois objetos. Para isso, ambos os objetos a serem comparados devem estar representados na forma de vetores  $n$ -dimensionais.

**DISTÂNCIA EUCLIDIANA** – Em inglês *Euclidian Distance* ou Distância L2 (*L2 Distance*) é a métrica frequentemente utilizada para se calcular o grau de similaridade entre dois objetos em um espaço dimensional. Para isso, os pares de objetos a serem analisados são representados na forma de um conjunto de vetores e, a partir disso, aplica-se a função da Distância Euclidiana para obter o nível de semelhança.

**EARTH MOVER'S DISTANCE (EMD)** – Também conhecido como *1st Wasserstein* ou *Monge-Kantorovich*, é um método para se calcular a distância entre duas distribuições finitas, ou seja, o custo mínimo para se transformar uma determinada distribuição em outra movendo a sua *massa de distribuição*.

**ESP@CENET** – Ou *Espacenet*, é um serviço on-line gratuito para busca de patentes. Foi desenvolvido pelo escritório de patentes europeu (*European Patent Office - EPO*) juntamente com os estados-membros da organização de patentes européia. A maioria dos membros possui um serviço *Espacenet* em seu próprio idioma e acesso ao banco de dados mundial da *EPO*.

**ESPAÇO VETORIAL BINÁRIO** – Conceito da álgebra linear em que todos os elementos que compõem o conjunto são representados na forma de vetores binários.

**FATOR DE NORMALIZAÇÃO** – Fator ou variável matemática usada para ajustes em distribuições probabilísticas.

**GIGABYTE (GB)** – Unidade de medida para o armazenamento de informação digital, sendo equivalente a  $10^9$  *bytes*.

**GRAFO** – Modelo de representação em que um conjunto de pontos (vértices) são ligados por um conjunto de linhas (arestas) direcionais ou bidirecionais.

**GRAHAM BELL** – *Alexander Graham Bell* é considerado historicamente como o inventor e fundador da empresa telefônica *Bell*. Em 2002, o italiano *Antonio Meucci* foi reconhecido pelo Congresso americano como o inventor oficial do telefone.

**INDICAÇÃO GEOGRÁFICA** – Meio de indicar a procedência de produção ou o local de extração de um determinado produto ou serviço conforme o nome da região, da cidade ou do país de origem que implicou na fama ou reconhecimento desse pelo público.

**INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL (INPI)** – Fundado em 1970 e vinculado ao *Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC)*, o *INPI* é responsável pelo aperfeiçoamento, disseminação e gestão do sistema brasileiro de concessão e direitos de propriedade intelectual para a indústria. Seus principais serviços são os registros de marcas, desenhos industriais, indicações geográficas, programas de computador e topografias de circuitos, as concessões de patentes e as averbações de contratos de franquia e das distintas modalidades de transferência de tecnologia.

**JANELA DE PREDIÇÃO** – Termo em inglês *prediction window*, é a lei que permite que os inventores ou titulares da invenção possam revelar a invenção ao público após 18 meses da data de depósito da patente na base *USPTO* e antes da data de publicação da mesma.

**JAVA UNIVERSAL NETWORK/GRAPH FRAMEWORK (JUNG)** – É uma biblioteca de programação escrita na linguagem *Java* que fornece processos para a geração, a análise e visualização de dados representando-os na estrutura de grafo ou rede.

**KILOBYTE** – Unidade de medida para o armazenamento de informação digital, sendo equivalente a  $10^3$  *bytes*.

**KNOWCEANS.ORG** – Biblioteca de programação na linguagem *Java* que apresenta algoritmos computacionais para cálculos de similaridade, extração de dados, métodos estatísticos e probabilísticos para redes de conhecimento.

**LATENT DIRICHLET ALOCATION (LDA)** – O modelo de tópicos *Latent Dirichlet Allocation* é uma classe de rede *Bayesiana* probabilística para um corpus de dados distintos, usado principalmente para modelar palavras de um texto na estrutura de *bag of words*. Baseia-se essencialmente na proposição de que um documento pode ser representado como uma mistura de tópicos, sendo esses tópicos representados como uma distribuição probabilística multinomial de palavras de um vocabulário.

**LINGPIPE** – Biblioteca de programação em linguagem *Java* que possui um conjunto de algoritmos computacionais usados para processamento de texto usando linguística computacional.

**LINGUAGEM DE PROGRAMAÇÃO JAVA** – Linguagem de programação orientada a objeto desenvolvida pela empresa *Sun Microsystems*.

**LONGEST SUBSEQUENCE MATCHING** – Também conhecido como *Longest common subsequence (LCS) matching*, é uma técnica normalmente usada para medir a similaridade entre sequências de *DNA* nos estudos de biologia molecular. O método calcula o maior comprimento de todas as *substrings* entre duas *strings*, onde as *substrings* podem ser não-contíguas e aparecer na mesma ordem em que aparecem em outra *string*.

**MARCA** – Conjunto de sinais visuais que fornece a individualidade para uma empresa em comparação aos seus concorrentes, possuindo um caráter *distintivo* e não enganoso dos produtos e serviços. O termo *distintivo* refere-se à necessidade da marca não ser de caráter descritivo, pois, uma vez que a marca seja um sinal comum para uma área de produtos e serviços da empresa, os seus concorrentes não poderão utilizá-la para descrever os seus produtos.

**MARKOV CHAIN MONTE CARLO** – Esse método procura construir uma cadeia de *Markov* específica que converge para uma determinada distribuição alvo e, assim, as amostras são obtidas por meio dessa distribuição.

**MARKOV CHAIN** – Ver o termo Cadeias de *Markov*.

**MDIC** – Ministério do Desenvolvimento, Indústria e Comércio Exterior. Órgão integrante da estrutura da administração pública federal direta. Responsável por formular, executar e avaliar políticas públicas para a promoção da competitividade, do comércio exterior, do investimento e da inovação nas empresas e do bem-estar do consumidor.

- MEGABYTE** – Unidade de medida para o armazenamento de informação digital, sendo equivalente a  $10^6$  *bytes*.
- MINERAÇÃO DE TEXTO** – Também conhecido como *Text Mining*, derivado da mineração de dados, consiste em extrair informação de dados textuais não estruturados ou semiestruturados.
- MONSANTO** – Empresa multinacional no desenvolvimento de tecnologias que contribuem para aliar produção de alimentos com preservação ambiental. Possui base na agricultura e biotecnologia, sendo líder mundial na produção do herbicida *glifosato* e sementes geneticamente modificadas (transgênicos).
- MPEP** – Manual publicado para prover aos examinadores de patentes do *USPTO*, advogados, agentes e representantes das aplicações, uma referência de trabalho para as práticas e processos de patentes. O *MPEP* contém instruções e outros materiais de informação, resumindo os processos que os examinadores de patentes são outorgados a seguir em determinados casos no exame de patentes.
- MÉTODOS BASEADOS EM CARACTERÍSTICAS** – Termo derivado do inglês para *Feature-based methods*, buscam determinar o nível de similaridade textual empregando um conjunto de características pré-definidas para a representação dos textos.
- MÉTODOS BASEADOS EM CORPUS** – Métodos computacionais de análise de texto que levam em conta a similaridade entre palavras usando informações de grandes corpora textuais.
- MÉTODOS HÍBRIDOS** – Em inglês *Hybrid methods*, são métodos de similaridade que utilizam a junção de vários algoritmos para verificar a semelhança entre textos, mesclando tanto técnicas de análise semântica como métodos baseados em corpus.
- NÓ** – Também conhecido como nodo ou vértice, na Teoria dos Grafos são tratados como objetos inexpressivos e indivisíveis, embora possam ter uma estrutura adicional, dependendo da aplicação a partir da qual surge o grafo. Normalmente são representados visualmente na forma de um círculo.
- ODYSSEYS** – Sistema computacional desenvolvido para a dissertação de mestrado que verifica a similaridade entre uma determinada patente dada pelo usuário e um grupo de documentos, ordenando-os conforme o seu grau de semelhança em relação à patente em avaliação. O *software* também permite buscas de um conjunto de patentes correlacionadas na base de dados do *USPTO* a partir de uma consulta designada

pelo usuário, utilizando essas patentes para a geração de uma rede de fluxo de conhecimento (Rede de citações).

**ORGANIZAÇÃO MUNDIAL DA PROPRIEDADE INTELECTUAL (OMPI)** –

Organismo das nações unidas que cuida para que os países membros possam aplicar os acordos e tratados a fim de garantir a segurança jurídica e o uso estratégico pelos agentes em todos os países membros.

**ORGANIZAÇÃO MUNDIAL DO COMÉRCIO (OMC)** – Criada em 1995, a *OMC*

está sediada na cidade de Genebra, Suíça. Procura lidar com regulamentações ligadas ao comércio entre os seus países membros. Também, fornece uma estrutura para negociação, acordos comerciais e um processo de resolução de conflitos.

**OXIDO DE ETILENO** – Gás inflamável incolor ou líquido refrigerado com odor doce.

É um composto químico usado na produção de *etilenoglicol* e como um esterilizante para alimentos e materiais de uso médico.

**PACOTE** – Termo usado em orientação a objetos para agrupar um conjunto de classes semelhantes.

**PATENT COOPERATION TREATY (PCT)** – Termo derivado do inglês, conhecido como Tratado de Cooperação em Matéria de patentes, *PCT* é o tratado mais importante administrado pela *OMPI* que facilita o depósito de pedidos de patentes em diversos países.

**PATENTE** – Documento sobre uma propriedade de criação concedida pelo Estado aos autores que impede terceiros a produzir, utilizar, comercializar, importar e exportar a invenção descrita sem a devida autorização do titular do documento, válido por um período determinado conforme as legislações locais.

**PERÍODO DE GRAÇA** – Termo derivado do inglês para *Grace period*, lei que assegura que as divulgações feitas pelo inventor ou terceiros não serão consideradas, desde que sido realizadas até 12 meses antes da data de depósito ou da prioridade reivindicada.

**PESO** – Também conhecido como *custo*, é o valor da aresta para se passar entre um estado que se encontra em um vértice da ligação para o estado que se encontra em outro vértice.

**PFAFF VS. WELLS ELECTRONICS** – Caso judicial de patentes em que a empresa *Pfaff* processou a empresa *Wells Electronics* por infringir uma patente de componente do computador.

**PRIMEIRO A DEPOSITAR** – Também conhecido em inglês como *First to file*, é a aplicação do sistema em que uma patente será concedida a primeira pessoa que depositar uma solicitação de patente.

**PROBLEMA DE TRANSPORTE** – Derivado do termo em inglês *Transportation problem*, consiste em determinar o custo mínimo de transporte e alocação de recursos para uma série de pontos de demanda (consumidores) a partir de um grupo de pontos de oferta (fornecedores).

**PROPRIEDADE INDUSTRIAL** – Conjunto de direitos que compreende as patentes, os modelos de utilidade, os desenhos industriais, as marcas, as indicações de proveniência ou denominações de origem e a repressão da concorrência desleal.

**PROPRIEDADE INTELECTUAL (PI)** – Tipo de direito estabelecido, conforme as legislações locais, sobre a posse legal de alguma coisa gerada pela criação do espírito humano.

**PROTEÇÃO SUI GENERIS** – Também conhecida como *híbridos jurídicos*, é a forma de denominação das novas criações intelectuais que não se enquadram por completo em nenhuma das modalidades fixadas de Propriedade Intelectual, podendo a invenção possuir características tanto dos Direitos de Autor quanto da Propriedade Industrial.

**QUERY DE BUSCA** – Termos de pesquisa dada pelo usuário a fim de obter as informações necessárias (conjunto de documentos) sobre determinado assunto.

**RANKING** – Relacionamento de ordenação entre um conjunto de itens. Para quaisquer dois itens, o primeiro é ordenado como maior, menor ou igual ao segundo item.

**RECUPERAÇÃO DE INFORMAÇÃO** – Área da computação que trabalha com o armazenamento de documentos e a recuperação automática das informações contidas em tais documentos.

**REDE BAYESIANA PROBABILÍSTICA** – São grafos acíclicos dirigidos que representam dependências entre variáveis em um modelo probabilístico.

**REDES DE CITAÇÕES** – Modelo de grafo dirigido ou dígrafo para a representação de redes de fluxo de conhecimento.

**REDUZINDO A IDEIA À PRÁTICA** – Termo em inglês derivado de *Reducing the idea to practice*. É a concretização de uma ideia em relação à utilidade e aplicabilidade na vida real.



**REVOLUÇÃO INDUSTRIAL** – Período histórico que engloba conjunto de mudanças que aconteceram na Europa nos séculos XVIII e XIX com impacto no nível econômico e social. A principal característica dessa revolução foi a substituição do trabalho antes completamente artesanal pelo trabalho assalariado com o uso das máquinas.

**ROYALTY** – É uma forma de pagamento pela licença a um terceiro para explorar algo patenteado pelo licenciador.

**SHORT PATH LENGTH** – Também conhecido como menor comprimento de caminho, em Teoria dos grafos, é a menor distância entre dois vértices em um grafo.

**SIMILARIDADE TEXTUAL LÉXICA** – Termo em inglês *Text-based lexical similarity*, são métodos que resultam em um valor numérico indicando o grau de semelhança entre os textos analisados. Para a realização do cálculo de similaridade, utiliza-se do número de unidades léxicas que os textos avaliados possuem em comum.

**SIMILARIDE TEXTUAL SEMÂNTICA** – Termo derivado do inglês *Text-based semantical similarity*, assim como na similaridade textual léxica, resultam em um valor numérico indicando o grau de semelhança entre os textos analisados. Possuem a característica de procurar identificar a similaridade semântica entre os textos.

**SIMMETRICS** – Biblioteca de programação escrita em *Java* usada para cálculos de métricas de similaridade.

**SMALL WORLD** – Rede ou grafo com alto nível de *clustering* e com menor comprimento de caminho (*short path length*) entre dois vértices.

**SMUCKER CO** – *J.M. Smucker Company* é uma empresa Americana especializada na fabricação de produtos como geleia de frutas, coberturas de sorvete, bebidas e manteiga de amendoim natural. Sua sede está localizada em Orrville, Ohio.

**SOFTWARE** – Conjunto dos componentes que não fazem parte do equipamento físico propriamente dito e que incluem as instruções e programas (e os dados a eles associados) empregados durante a utilização do sistema.

**SPEARMAN'S RANK CORRELATION COEFICIENTE** – Ver o termo Correlação de Postos de *Spearman*.

**STEMMING** – No processamento de linguagem natural, *Stemming* é o processo para a redução de palavras à sua raiz ou base.

**STOP WORDS** – São palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em uma busca realizada.



**SUMARIZAÇÃO DE TEXTO** – Área de processamento de linguagem natural em que consiste em reduzir um determinado corpus textual de forma a retirar textos que agregam pouco valor à ideia principal contida no documento.

**TERABYTE** – Unidade de medida para o armazenamento de informação digital, sendo equivalente a  $10^{12}$  bytes.

**TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)** – Também conhecido como *TFIDF*, é um método de caráter estatístico baseado em vetor, sendo bastante empregado nas áreas de recuperação de informação e mineração de texto.

**TFT-LCD** – Sigla em inglês para *Thin film transistor liquid crystal display*, é uma variação da tela de *LCD* que usa *Thin-film transistor (TFT)*, tecnologia que melhora a qualidade de imagem.

**TRADE RELATED ASPECTS OF INTELLECTUAL PROPERTY RIGHTS (TRIPS)**  
Termo em inglês para Acordo Relativo aos Aspectos do Direito da Propriedade Intelectual Relacionados com o Comércio (*ADPIC*).

**TRANSGENIA** – Consiste no desenvolvimento de organismos geneticamente modificados por meio da introdução na carga genética ou genoma – *DNA (deoxyribonucleic acid)* ou *ADN (ácido desoxirribonucleico)* – de um organismo receptor, de genes alterados de indivíduos de mesmas ou diferentes espécies.

**TÉCNICA ITERATIVA** – É o processo conhecido na área de computação como a repetição de uma ou mais ações através de cálculos sucessivos.

**TÓPICOS** – Também conhecidos como cluster de palavras, buscam sintetizar os assuntos de que se tratam um determinado documento, sendo cada conceito representado por uma distribuição de palavras que juntas buscam formar um significado.

**UNION CARBIDE VS. SHELL** – Caso em que, segundo a Corte Suprema dos Estados Unidos, a empresa *Shell* infringiu três patentes pertencentes à empresa *Union Carbide*. As patentes tratam de um processo para a produção de *óxido de etileno* com um catalisador de prata que contém uma combinação de césio e outro metal alcalino (lítio, sódio, potássio ou rubídio).

**UNITED STATES PATENT AND TRADEMARK OFFICE (USPTO)** – Agência do Departamento de Comércio dos *EUA* responsável pela emissão de patentes e obtenção de registro de marcas.

**VECTOR-BASED DOCUMENT MODEL** – Também conhecido como *Vector space model* ou *term vector model*, é um modelo algébrico para representar documentos de texto (ou outros objetos) na forma de vetores de identificadores. Usado normalmente na filtragem e recuperação de informações, indexação e rankings de relevância.

**VETOR ESPACIAL DE CARACTERÍSTICAS** – Em inglês *Feature vector space*, modelo matemático usado para a representação de um objeto conforme as características que o descrevem na forma de vetor n-dimensional.

**WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS (WEKA)** – Biblioteca de programação escrita em *Java*, *WEKA* possui vários algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial, focando-se no tema aprendizagem de máquina.

**WORD CO-OCCURRENCE METHOD** – Termo em inglês para Co-ocorrência de palavras, também conhecidos como métodos de modelo de documento baseados em vetor (*Vector-based document model method*), são os métodos mais comuns aplicados no campo de similaridade de textos e baseiam-se na proposição de que documentos semelhantes possuem muitas palavras em comum em seus corpora.

**WORLD INTELLECTUAL PROPERTY ORGANIZATION (WIPO)** – Termo em inglês para *OMPI*. Ver o termo Organização Mundial da Propriedade Intelectual no glossário.

**WORLD TRADE ORGANIZATION (WTO)** – Termo em inglês para *OMC*. Ver o termo Organização Mundial da Propriedade Intelectual no glossário.

# Apêndice A

## Propriedade Intelectual no Brasil

A seguir apresenta-se o quadro geral referente à Propriedade Intelectual no Brasil:

### A.1 Quadro Atual

- Decreto Legislativo nº 1355/94 – TRIPs (Trade Related Aspects of Intellectual Property Rights)
- Lei nº 9.279/96 – Lei de Propriedade Industrial
- Lei nº 9.610/98 – Lei do Direito Autoral
- Lei nº 9.456/97 – Lei de Proteção de Cultivares
- Decreto nº 2.366/97 – Regulamenta a Lei de Proteção de Cultivares
- Lei nº 9.609/98 – Lei de Proteção de Programas de Computadores
- Lei nº 10.196, de 14 de fevereiro de 2001 – Altera e acrescenta dispositivos à Lei nº 9.279, de 14 de maio de 1996, que regulam direitos e obrigações relativas à propriedade industrial
- Decreto nº 2.553/98 – Premiação do Inventor
- Decreto nº 3.201/99 – Licença compulsória de Patentes
- Decreto nº 4.830/03 – Nova redação aos artigos 1º, 2º, 5º, 9º e 10º do Decreto nº 3.201, de 6 de outubro de 1999, que dispõe sobre a concessão, de ofício, de licença compulsória nos casos de emergência nacional e de interesse público de que tratou o artigo 71 da Lei nº 9.279, de 14 de maio de 1996

- Decreto nº 4.602/2003 – Constitui Comissão Interministerial que envolva a pesquisa, licenciamento, autorização, cultivo, manipulação, transporte, comercialização, consumo, armazenamento, liberação e descarte de Organismos Geneticamente Modificados - OGM
- Lei nº 10.973/2004 – Lei de Inovação
- Lei nº 11.105/2005 – Lei de Biossegurança
- Ato Normativo nº 117/93 – Institui o uso do dígito verificador na numeração dos processos de patentes
- Ato Normativo nº 126/96 – Regulamenta o procedimento de depósito previsto nos artigos 230 e 231 da Lei nº 9.279/96
- Ato Normativo nº 127/97 – Dispõe sobre a aplicação da Lei de Propriedade Industrial em relação às Patentes e Certificados de Adição de Invenção
- Ato Normativo nº 128/97 – Dispõe sobre a aplicação do Tratado de Cooperação em Matéria de Patentes
- Ato Normativo nº 130/97 – Dispõe sobre a instituição de formulários para apresentação de requerimentos e petições na área de Patentes, Certificados de Adição de Invenção e Registro de Desenho Industrial
- Ato Normativo nº 152/99 – Dispõe sobre a apresentação de auxílio voluntário para o exame técnico, em relação às patentes e certificados de invenção (Ato Normativo revogado pela Resolução DIRPA nº 118/05 de 15 de junho de 2005)

## Apêndice B

# Análise Manual de Similaridade de Patentes

A análise manual da similaridade de patentes dos experimentos apresentados no Capítulo 6.4.2 foi realizada por um especialista na área de patentes.

Para a obtenção de cada grupo e patente base de análise para uma determinada tecnologia, utilizou-se de documentos pertencentes a uma área tecnológica de total domínio por parte do profissional. Para isso, inúmeros testes foram feitos de forma manual escolhendo-se patentes que possuem a característica de sempre citar a patente principal de análise. A seguir são apresentadas todas as análises dos experimentos feitos pelo especialista em patentes, inclusive para os experimentos em que não foi possível classificar as patentes por uma ordem de similaridade.

Análise de similaridade de conteúdos técnicos de patentes

Termos da busca – USPTO:

ABST = “35S OR ubiquitin” OR Claim = “35S OR ubiquitin” AND IPC = C07H021 OR C12N.

## B.1 1<sup>a</sup> tentativa

Patente-chave: <b>4940835</b> Similares (citantes)	Ordem crescente de similaridade presumida
7,829,767	
7,829,766	
7,825,304	
7,825,303	
7,825,302	
7,825,301	
7,825,300	
7,820,888	
7,820,887	
7,803,997	

**4,940,835** – patente do vetor e gene de resistência ao glifosato. As outras patentes são de variedades de plantas geneticamente modificadas que também possuem tal resistência.

Cada patente, em especial 7829767, 7829766, 7825304, são formas de proteção às diferentes partes do mesmo invento, ou de variedades de plantas resultantes de cruzamentos diversos, mas sempre com a referida resistência. Assim, não é possível designar, para este grupo, graus de similitude diferentes, pois praticamente todas as patentes desta lista de citadoras são variações semelhantes de uma mesma tecnologia.

Portanto, conclui-se que a similaridade não procede.

## B.2 2ª tentativa

Patente-chave: <b>4940835</b> Similares (citantes)	Ordem crescente de similaridade presumida	Considerações
6,399,81	6	vetor de Streptomices para milho
6,395,966	3	Plantas de milho resistente
6,362,396	8	Gene quimérico diferente, mas processo semelhante de clonagem e vetor
6,331,665	4	Plantas de milho resistente
6,329,54	7	Mesmo vetor (Streptomices), mas com aplicação em outra cultura e com outra enzima de resistência
6,326,57	5	Plantas com conteúdo de amino-ácidos diversos; usa apenas vetores de genes
6,306,66	2	Genes do acetilCoA transferase
6,281,41	1	Resistência à seca em milho
6,271,06	9	Método de obtenção de tolerância
6,268,50	10	Planta GM; usa mesmo vetor, mas para tipo de controle genético diferente (óleo), ou seja, é muito semelhante, com aplicação de controle metabólico que é diferente

### B.3 3<sup>a</sup> tentativa

Patente-chave: <b>5968830</b> Similares (citantes)	Ordem crescente de similaridade presumida	Considerações
7,777,104		Cultivar de soja (26650228) da Monsanto Usou tecnologia de 5,968,830 (Agrobacterium) como vetor de genes
7,777,103		Cultivar de soja (6900358) da Monsanto Usou tecnologia de 5,968,830 (Agrobacterium) como vetor de genes
7,759,553		Cultivar de soja (5070152) da Monsanto Usou tecnologia de 5,968,830 (Agrobacterium) como vetor de genes
7,759,552		Cultivar de soja (7429331) da Monsanto Usou tecnologia de 5,968,830 (Agrobacterium) como vetor de genes
7,759,551		
7,732,671		
7,728,204		
7,728,202		
7,728,201		
7,728,200		

Este grupo de patentes demonstra que nem sempre faz sentido buscar graus diversos de similaridades entre patentes. As 4 primeiras citadoras (todas da *Monsanto*) são, igualmente similares, pois são 4 cultivares de soja que foram desenvolvidas com base na técnica de transgenia por *Agrobacterium* (patente 5,968,830).

### B.4 4<sup>a</sup> tentativa

Patente-chave: <b>5968830</b> Similares (citantes)	Ordem crescente de similaridade presumida	Considerações
7,820,888		Cotton variety MCS0701B2RF Idem ao grupo da tentativa 3



## B.5 5ª tentativa

Patente-chave: <b>5968830</b> Similares (citantes)	Ordem crescente de similaridade presumida	Considerações
7,759,463	8	RNA interference pathway genes as tools for targeted genetic interference Tecnologia de apoio e validação à principal; faz parte do pacote tecnológico da transgenia
7,754,942	2	Maize starch containing elevated amounts of actual amylase Tecnologia de resultado: milho com maior teor de amido
7,754,869	3	Production of syringyl lignin in gymnosperms Controle de lignina em gimnospermas
7,754,697	10	Control of gene expression Patente muito similar, à principal, pois faz parte do pacote tecnológico da transgenia
7,723,503	9	Desaturase genes, enzymes encoded thereby, and uses thereof
7,705,203	1	Benzoate inductible promoters Tecnologia parece abrir nova trajetória.
7,700,851	4	Tobacco nicotine demethylase genomic clone and uses thereof Uso da tecnologia principal para transformar tabaco
7,700,834	5	Nicotiana nucleic acid molecules and uses thereof Patente de cunho metodológico
7,683,237	6	Maize seed with synergistically enhanced lysine content Cita a principal apenas como referência, pois é método de transgenia mais específico
7,645,925	7	Tobacco products with increased nicotine Cita a principal apenas como referência, pois é método de transgenia mais específico

A patente 5,107,065 (Anti-sense regulation of gene expression in plant cells) representa tecnologia seminal para a transgenia, pois é a técnica de regulação da expressão de genes que possam ter sido alterados por engenharia genética.

As patentes de nível 2, 3 e 4 apresentam similaridade muito grande entre si, sendo, as 3, produtos do uso e aplicação da principal.

## B.6 6<sup>a</sup> tentativa

Patente-chave: <b>5968830</b> Similares (citantes)	Ordem crescente de similaridade presumida	Considerações
7,202,083	9	Plant promoters and plant terminators. Patente muito similar, à principal, pois faz parte do pacote tecnológico da transgenia
7,192,771	7	Plant promoter sequence Patente muito similar, à principal, pois faz parte do pacote tecnológico da transgenia
7,189,570	8	Putrescine-n-methyltransferase promoter Patente muito similar, à principal, pois faz parte do pacote tecnológico da transgenia
7,074,987	1	Genetically-controlled herbicide resistance in cotton plants in the absence of genetic engineering Método que usa tecnologia nova, e que pretende dispensar a tecnologia principal
7,183,110	10	Antibody immunoreactive with a 5-enolpyruvylshikimate-3-phosphate synthase Patente muito similar, à principal, pois faz parte do pacote tecnológico da transgenia
7,161,064	6	Method for producing stably transformed duckweed using microprojectile bombardment Tecnologia de apoio e validação à principal; faz parte do pacote tecnológico da transgenia
6,777,546	3	Methods and substances for preventing and treating autoimmune disease Aplicação ímpar de modificação de plantas; deve citar a principal apenas pq esta última é referência para tudo o que for planta modificada geneticamente.
7,135,626	5	Soybean seeds and plants exhibiting natural herbicide resistance. Tecnologia de apoio e validação à principal; faz parte do pacote tecnológico da transgenia
7,135,282	2	Viral particles with exogenous internal epitopes Nexos mais distantes com a principal, pois os vírus são também utilizados como vetores de transgenia.
7,087,809	4	Natural herbicide resistance in wheat Tecnologia de apoio e validação à principal; faz parte do pacote tecnológico da transgenia

**5,352,605** Chimeric genes for transforming plant cells using viral promoters Patentes de similitude 5, 6, 7, 8,9 e 10 são muito semelhantes, em termos tecnológicos.



# Apêndice C

## Tabelas de Similaridade

A seguir são apresentadas as tabelas individuais dos resultados obtidos pela utilização dos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos para os experimentos realizados no Capítulo 6.4.

As tabelas são organizadas em ordem decrescente de similaridade segundo os resultados obtidos pelo método utilizado para se avaliar a similaridade da patente *USPTO 4876196* em relação às demais patentes do grupo de análise.

### C.1 Análise de grupos tecnológicos distintos de patentes

#### C.1.1 Experimento 1

As Tabelas C.1, C.2 e C.3 apresentam, respectivamente, os resultados obtidos pela aplicação dos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos. As patentes utilizadas para o experimento pertencem a um dos três grupos tecnológicos distintos mencionados no Capítulo 6.4.1 e são comparadas à patente base de análise 4876196 a fim de se encontrar o nível de similaridade entre os documentos. Os resultados mostrados em cada tabela são ordenados em ordem percentual decrescente de similaridade.

<b>Grupo</b>	<b>Nº USPTO da patente</b>	<b>LDA + EMD</b>
1	4876196	99,9762
1	4326036	87,3273
1	4560659	76,2738
1	4738930	71,239
2	5428528	59,6165
2	6416410	55,303
2	6500070	49,5798
2	6478583	47,933
2	7371163	41,9381
2	6039574	41,0027
2	7316618	35,4835
3	PP11246	1,4609
3	PP19575	1,1695
3	PP06047	0,6211
3	PP04161	0,6145
3	PP08212	0,4762
3	PP07700	0,4587
3	PP07651	0,4214
3	PP18774	0,3774
3	PP08238	0,2877

Tabela C.1: Experimento 1: Grau de similaridade obtido pelo método *LDA+EMD* para os documentos de análise em relação a patente 4876196

<b>Grupo</b>	<b>Nº USPTO da patente</b>	<b>Espaço Vetorial Binário + Distância Euclidiana</b>
1	4876196	100
1	4560659	95,18469
1	4326036	95,011406
1	4738930	94,689575
3	PP11246	94,08937
3	PP08238	94,07054
3	PP07700	93,97782
3	PP07651	93,96202
3	PP18774	93,95004
3	PP06047	93,88927
2	6478583	93,81358
3	PP19575	93,77094
3	PP04161	93,663734
2	6039574	93,511375
2	7316618	93,24559
3	PP08212	92,91633
2	6416410	91,88554
2	7371163	91,861176
2	5428528	89,943085
2	6500070	88,272385

Tabela C.2: Experimento 1: Grau de similaridade obtido pelo método Espaço Vetorial Binário+Distância Euclidiana para os documentos de análise em relação a patente 4876196

Grupo	Nº USPTO da patente	TF-IDF + Distância dos Cossenos
1	4876196	100
1	4326036	34,9728
1	4560659	32,9674
1	4738930	29,3144
2	6478583	12,4934
2	6039574	11,795
3	PP08212	11,6983
2	6500070	10,4904
2	6416410	9,9098
3	PP08238	9,1945
3	PP11246	8,8773
2	7371163	8,1337
3	PP18774	7,8252
3	PP19575	7,1547
2	5428528	7,0148
3	PP07700	6,7322
3	PP07651	6,2545
3	PP06047	6,0575
2	7316618	4,7962
3	PP04161	4,1367

Tabela C.3: Experimento 1: Grau de similaridade obtido pelo método *TF-IDF*+Distância dos Cossenos para os documentos de análise em relação a patente 4876196



### C.1.2 Experimento 2

Seguindo o mesmo esquema realizado no experimento 1, as tabelas C.4, C.5 e C.6 também apresentam os resultados obtidos pela aplicação dos métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos para um conjunto de patentes que possuem áreas tecnológicas distintas. Porém, nesse experimento, apenas partes do documento original de cada patente foram empregados para a análise (modelo parcial de patentes), englobando, assim, o campo *title*, *abstract*, *claims* e *descriptions* das patentes. Dessa forma, para a patente base de análise 4876196 também foram utilizados somente esses campos para a análise de similaridade.

Grupo	Nº USPTO da patente	LDA + EMD
1	4876196	38,8139
1	4560659	37,5935
1	4326036	35,3475
1	4738930	34,2908
2	PP06047	33,2007
2	PP11246	33,1254
2	PP08212	29,0465
2	PP19575	26,5089
3	6500070	24,3578
3	6478583	23,5409
3	7371163	23,1022
2	PP04161	22,2429
2	PP08238	21,56
2	PP18774	21,355
3	6416410	19,5805
3	6039574	18,7151
3	5428528	18,0807
2	PP07651	16,4518
3	7316618	16,3367
2	PP07700	16,0205

Tabela C.4: Experimento 2: Grau de similaridade obtido pelo método *LDA+EMD* para os documentos de análise em relação ao modelo parcial de patente 4876196

<b>Grupo</b>	<b>Nº USPTO da patente</b>	<b>Espaço Vetorial Binário + Distância Euclidiana</b>
1	4876196	92,96484
3	6478583	92,59906
3	6039574	92,48126
2	PP08238	92,22324
1	4560659	92,15224
1	4326036	92,13364
1	4738930	91,14969
2	PP11246	91,01645
2	PP18774	90,94282
2	PP08212	90,91636
2	PP19575	90,751076
2	PP07651	90,53381
3	7371163	90,4315
3	6416410	90,430214
2	PP07700	89,900444
3	7316618	89,45783
2	PP06047	89,14898
2	PP04161	88,48225
3	5428528	87,289116
3	6500070	87,03087

Tabela C.5: Experimento 2: Grau de similaridade obtido pelo método Espaço Vetorial Binário+Distância Euclidiana para os documentos de análise em relação ao modelo parcial de patente 4876196

Grupo	Nº USPTO da patente	TF-IDF + Distância dos Cossenos
1	4876196	64,3583
1	4326036	28,5735
1	4560659	24,745
1	4738930	19,7982
2	PP11246	5,7187
3	6478583	5,4625
3	6039574	5,4162
3	6500070	4,7867
2	PP08212	3,7005
3	7371163	3,6806
2	PP07651	3,5932
3	6416410	3,4057
2	PP08238	3,3432
2	PP04161	2,6787
2	PP06047	2,5143
2	PP19575	2,3941
3	5428528	2,2497
2	PP18774	2,0732
2	PP07700	1,7068
3	7316618	1,255

Tabela C.6: Experimento 2: Grau de similaridade obtido pelo método *TF-IDF*+Distância dos Cossenos para os documentos de análise em relação ao modelo parcial de patente 4876196

## C.2 Análise de grupos tecnológicos similares de patentes

Cada uma das tabelas a seguir mostram os resultados alcançados pelo analista de patentes e os métodos *LDA+EMD*, Espaço Vetorial Binário+Distância Euclidiana e *TF-IDF*+Distância dos Cossenos, encontrando-se ordenados em forma de *Ranking* para os experimentos realizados no Capítulo 6.4.2.

### C.2.1 Experimento 1

Nº USPTO da patente	Analista de Patentes	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
6268550	1	2	9	9
6271016	2	3	6	3
6362396	3	1	10	6
6329574	4	5	4	2
6399861	5	10	2	5
6326527	6	4	8	4
6331665	7	8	3	8
6395966	8	9	7	10
6306636	9	6	1	1
6281411	10	7	5	7

Tabela C.7: Experimento 1: *Ranking* de similaridade de patentes em relação ao documento *USPTO 4940835*

### C.2.2 Experimento 2

Nº USPTO da patente	Analista de Patentes	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
7754697	1	8	7	5
7723503	2	9	3	9
7759463	3	7	5	7
7645925	4	1	2	1
7683237	5	4	10	10
7700834	6	3	8	3
7700851	7	5	9	4
7754869	7	2	6	6
7754942	7	10	1	8
7705203	10	6	4	2

Tabela C.8: Experimento 2: *Ranking* de similaridade de patentes em relação ao documento USPTO 5107065

### C.2.3 Experimento 3

Nº USPTO da patente	Analista de Patentes	LDA + EMD	Espaço Vetorial Binário + Distância Euclidiana	TF-IDF + Distância dos Cossenos
7183110	1	4	3	2
7202083	1	8	6	5
7189570	1	7	1	3
7192771	1	9	5	9
7161064	1	6	4	4
7135626	1	1	9	6
7087809	7	2	8	8
6777546	8	10	2	7
7135282	9	3	7	1
7074987	10	5	10	10

Tabela C.9: Experimento 3: *Ranking* de similaridade de patentes em relação ao documento USPTO 5352605



## Apêndice D

# Tabelas e Listas de Tópicos de *LDA with Gibbs Sampling*

A seguir são apresentadas as tabelas e listas de associação de tópicos dos experimentos realizados no Capítulo 6.4.1 que utiliza grupos tecnológicos distintos de patentes. Antes da geração automática de tópicos pelo método *LDA with Gibbs Sampling*, realizou-se um pré-tratamento das patentes, removendo *stopwords* e efetuando processamento de *stemming*.

---

**Experimento 1 – Patente 4876196**  
Tópicos obtidos por meio do método *LDA with Gibbs Sampling* para grupos distintos de patentes

---

<b>Tópico</b>	<b>Termos</b>
0	Assist, seed, dpal, pat, character, intern, font, typ, provinc, html, bagwil,reach, direct, clar, approxim, patimg, janu, anther, citrusd, exemplif
1	Orang, navel, tre, fruit, varies, color, inch, lat, fp, matur, approxim, washingt dp, cultivar, character, siz, green, tree, length, lan
2	Vessel, pl, ferm, ac, pressur, sugar, flow, yeast, loc, direct, separ, solut, pas, rot, mic, liquid, discharg, apparatus, chamber, horizont
3	Ferm, substr, zymomon, mobil, ethanol, stag, method, sugar, fermentor, produc, med, cel, rat, yeast, system, approxim, concentr, ferment, part, flow
4	Gam, left, port, align, sect, bool, pal, qu, hitoff, netacg, fig, html, object, fpt, bas, nph, mach, ses, href, elong
5	Parser, href, direct, disclos, com, ad, part, day, pr, ser, beck, provid, district, depend, increase, uspt, select, dur, patent, farm

6	Align, left, pt, width, center, valign, href, sect, border, alt, src, gif, netaicon, img, middl, tabl, uspt, http, gov, top
7	Pr, afric, not, attorne, util, field, pagen, shop, cap, dur, length, start, rip, pag, abstract, patfthdr, dpal, tabl, addit, docum
8	Can, ethanol, sugar, sucros, ferm, yeast, suspens, process, extract, particl, claim, stalk, comminute, separ, strain, saccharomyc, biomas, produc, concentr, consum
9	Provinc, valign, dist, trademark, dpn, fsrchn, obtain, larg, jl, point, sunrays, occur, bool, adv, srchn, understood, farm, templ, rapid, pr
10	Produc, sugar, ethanol, ferm, digest, fibr, process, can, stag, residu, lin, juic, liquid, enzym, pas, ferment, step, cultur, extract, separ
11	Ac, stock, bottom, disclos, shap, tanger, robertson, flavor, provinc, consum, mad, foreign, cur, import, evid, refer, accompan, gram, origin, seed
12	Miam, cap, assist, appl, inclus, pr, attract, sid, ad, detachm, docnumber, section, stock, improve, andrew, compact, angl, period, regl, ens
13	Car, fil, graft, resid, eas, appl, foreign, dpt, cap, remov, robers, origin, grey, fp, approxim, held, td, observ, fnph, width
14	Foreign, cel, search, textur, distribut, narrow, border, citrusd, specif, provid, consider, differ, gif, straight, prov, ens, gram, backlabel, field, seed
15	Usu, fffff, larger, document, volkamerian, lower, netaicon, valign, docum, lemon, environm, exces, fil, read, cut, pating, wid, applic, febru, piw
16	Fffff, design, cart, miam, rest, nieuwoudt, disclos, respect, dist, show, http, belief, graft, consider, common, indic, pal, pat, larger, great
17	Clas, strong, htm, exhibit, mas, parser, car, ref, mak, upper, winter, bagwil, intens, unit, shoppingcart, titl, improve, rough, larger, glandl
18	Templ, remov, vari, font, abstract, depends, parser, parent, act, oil, pronounc, flavor, seed, bitter, firm, manner, ebiz, desir, sphere, backlabel
19	Button, tim, display, mod, enter, devic, spac, timer, control, pres, gam, user, hous, inclus, screen, invent, select, squar, oper, mean
20	Walker, unit, href, abstract, indic, examiner, intens, effect, blossom, shap, width, obtus, paul, citrusd, head, dpn, prior, exterior, continu, degree
21	Wheel, transfer, rot, power, steer, unit, funct, controller, key, support, lin, port, seat, steel, mach, serv, cover, ses, act, lever
22	Xd, field, dat, redund, compres, fil, pattern, character, displa, decompress, process, symbol, encod, block, prefer, gam, sentinel, embodim, bys, background



23	Gam, unit, fig, program, player, stor, mach, mult, casses, child, parent, shown, transfer, communic, detach, show, system, oper, embodim, receiv
24	Oil, miam, clim, comparison, com, determin, middl, approxim, hom, larger, graft, docnumber, citrusd, district, solubl, draw, view, low, cur, textur
25	Gam, mach, port, left, align, stor, display, vide, inform, unit, program, oper, sect, connect, player, proces, dat, fig, pictur, refsrch
26	Effect, environm, common, lemon, remain, view, read, method, detachm, group, singl, valign, document, smaller, med, jun, improv, rtyp, plan, janu
27	Inch, varies, tre, navel, pl, excel, acis, test, ens, grov, orang, year, november, number, washingt, parent, equ, solis, effect, unit
28	Gam, system, dat, lcd, port, displa, object, fig, proces, dimens, player, devic, view, gener, siml, viewpoint, world, imag, stereoscop, display
29	District, pr, dpn, seed, august, pick, tabl, firm, graft, middl, spher, lighter, continu, mclar, thor, ref, str, ens, improv, pag

### Experimento 1 – Patente 4876196

Lista de Tópicos obtidos por meio do método  
LDA with Gibbs Sampling para as patentes

Nº USPTO da patente	Tópicos
4876196	3, 6, 10, 8, 2, 25, 22, 19, 23, 28, 20, 15, 0, 27, 11, 5, 16, 1, 7, 14, 9, 13, 12, 4, 24, 26, 29, 18, 21, 17
PP04161	6, 1, 18, 29, 27, 23, 11, 12, 4, 13, 7, 20, 16, 15, 17, 9, 26, 0, 14, 5, 24, 10, 2, 22, 25, 3, 28, 21, 8, 19
PP06047	6, 1, 9, 27, 2, 20, 29, 0, 26, 7, 5, 11, 18, 24, 10, 16, 13, 14, 15, 17, 4, 12, 3, 8, 19, 28, 21, 25, 22, 23
PP07651	6, 1, 16, 29, 14, 12, 5, 15, 24, 4, 9, 13, 11, 25, 26, 17, 7, 18, 2, 20, 27, 23, 0, 21, 8, 19, 3, 10, 28, 22
PP07700	6, 1, 5, 16, 27, 7, 17, 14, 4, 13, 18, 9, 24, 26, 29, 11, 15, 20, 0, 12, 19, 23, 22, 8, 25, 10, 28, 3, 2, 21
PP08212	1, 6, 27, 3, 25, 12, 17, 11, 0, 14, 26, 9, 24, 5, 19, 8, 10, 4, 16, 18, 29, 7, 15, 22, 20, 13, 2, 28, 21, 23
PP08238	1, 6, 27, 19, 4, 16, 29, 2, 18, 13, 5, 15, 14, 20, 17, 12, 0, 11, 9, 26, 7, 25, 24, 21, 10, 22, 23, 3, 28, 8
PP11246	6, 1, 3, 8, 25, 14, 29, 18, 27, 0, 24, 7, 26, 16, 17, 5, 9, 15, 20, 2, 11, 4, 21, 13,

	19, 22, 12, 10, 23, 28
PP18774	6, 1, 12, 16, 29, 13, 18, 4, 7, 27, 0, 2, 26, 15, 11, 5, 24, 20, 17, 14, 9, 10, 23, 28, 21, 8, 22, 3, 25, 19
PP19575	6, 1, 11, 25, 7, 8, 23, 27, 10, 26, 18, 14, 16, 5, 4, 28, 13, 9, 20, 0, 17, 12, 29, 24, 15, 22, 19, 3, 2, 21
4326036	10, 6, 3, 8, 2, 25, 28, 19, 22, 18, 17, 21, 0, 4, 11, 13, 9, 20, 12, 14, 5, 16, 23, 7, 24, 27, 26, 15, 29, 1
4560659	8, 6, 10, 3, 25, 19, 2, 22, 23, 28, 15, 9, 12, 1, 29, 17, 20, 26, 5, 14, 21, 13, 16, 7, 27, 4, 24, 18, 11, 0
4738930	2, 6, 3, 10, 25, 19, 23, 8, 28, 22, 21, 9, 12, 15, 26, 11, 27, 14, 16, 29, 4, 18, 20, 17, 7, 5, 0, 13, 24, 1
4876196	3, 6, 10, 8, 2, 25, 22, 19, 28, 23, 9, 27, 7, 11, 24, 16, 20, 5, 4, 14, 1, 15, 21, 12, 13, 0, 26, 17, 18, 29
5428528	23, 6, 25, 28, 19, 22, 21, 4, 10, 26, 2, 0, 13, 27, 9, 20, 12, 17, 3, 8, 14, 24, 29, 18, 5, 11, 7, 15, 1, 16
6039574	19, 6, 25, 23, 28, 22, 21, 1, 4, 5, 17, 16, 2, 18, 26, 7, 12, 13, 29, 15, 20, 8, 10, 11, 14, 24, 0, 3, 9, 27
6416410	22, 6, 25, 28, 19, 23, 4, 21, 3, 2, 1, 26, 15, 18, 10, 14, 24, 11, 17, 8, 13, 0, 20, 5, 7, 12, 27, 29, 9, 16
6478583	19, 6, 25, 28, 23, 22, 4, 1, 21, 2, 7, 9, 11, 3, 27, 14, 24, 15, 16, 5, 10, 26, 18, 17, 20, 0, 13, 8, 29, 12
6500070	25, 6, 4, 23, 28, 19, 22, 21, 3, 0, 27, 12, 7, 18, 2, 5, 17, 14, 10, 15, 26, 24, 11, 20, 29, 13, 1, 16, 8, 9
7316618	6, 21, 25, 23, 19, 28, 4, 22, 2, 11, 26, 27, 7, 24, 12, 13, 15, 17, 5, 0, 14, 29, 16, 20, 18, 8, 9, 3, 10, 1
7371163	28, 6, 25, 19, 23, 22, 4, 21, 20, 12, 27, 2, 16, 0, 5, 13, 3, 9, 11, 24, 18, 14, 15, 8, 7, 26, 10, 29, 17, 1

**Experimento 2 – Partes da Patente 4876196**

Tópicos obtidos por meio do método *LDA*

*with Gibbs Sampling* para grupos distintos de patentes

<b>Tópico</b>	<b>Termos</b>
0	Align, left, pt, width, center, valign, href, sect, border, alt, img, netaicon, src, gif, middl, tabl, http, uspt, gov, top
1	Lemon, com, seed, obtain, cap, provinc, thin, method, green, st, exces, dist, moder,

	denom, short, back, acis, gov, shown, hundr
2	Fsect, substant, tot, pr, ad, eas, lower, ses, pres, robers, refer, entir, hom, intens, yield, lemon, cit, insid, continu, structur
3	Qu, determine, entir, manner, year, order, red, shap, lighter, gif, draw, august, assigne form, simultan, signif, april, srchn, citrusd, resid
4	Substant, rati, com, direct, condit, cart, prov, specif, citrus, adv, improv, dpal, breed, bel, aspect, cut, up, applic, bottom, august
5	Vessel, pl, ferm, ac, pressur, sugar, flow, yeast, loc, direct, separ, solut, pas, rot, mic, liquid, discharg, chamber, apparatus, horizont
6	Es, top, backlabel, are, robers, tot, afric, bonanz, stock, suffici, altern, tim, stor, src, jl, juic, vari, denom, volum, depends
7	Mil, templ, miam, rough, detachm, exces, thor, agent, simultan, sampl, open, entir, text, consist, thereof, greater, sect, reach, htm, cros
8	Uspt, virtu, pal, split, south, cultur, andrew, continu, simultan, parser, ag, piw, text, consist, separ, dhitoff, centimeter, afric, improve, week
9	Cros, stor, continu, dpn, nieuwoudt, deem, fnetacg, prior, econom, pagen, inventor, dpt, assigne, simultan, abstract, temperatur, glandl, dens, intens, remain
10	Consist, thin, dur, quant, flat, bel, inclus, idke, assigne, control, interior, siz, www, prim, scion, cut, jun, character, mar, shoppingcart
11	Xd, field, dat, redund, compres, fil, pattern, displa, character, decompress, proces, symbol, encod, gam, prefer, block, sentinel, embodim, bys, background
12	Can, ethanol, sugar, sucros, ferm, yeast, suspens, proces, extract, particl, claim, stalk, separ, strain, comminut, saccharomyc, produc, concentr, biomas, water
13	Button, tim, display, mod, enter, devic, spac, timer, control, pres, gam, user, hous, inclus, screen, invent, select, squar, oper, mean
14	Maxim, examiner, delan, result, improv, bagwil, produc, accept, docnumber bel, boolean, firm, ordin, clas, district, homeurl, highes, bonanz, yellow, expres
15	Navel, orang, tre, fruit, varies, color, inch, lat, fp, matur, approxim, washingt, cultivar, dp, green, character, siz, tree, lan, length
16	Grown, disclos, short, random, minim, uspt, maxim, pronounc, ser, count, ens, environm, effect, quant, fsect, desir, imag, lin, proceed, font
17	Produc, sugar, ethanol, ferm, digest, fibr, proces, can, stag, residu, lin, liquid, juic, enzym, pas, ferment, step, cultur, extract, separ
18	Evid, netahtml, fpatft, fnph, boolean, addit, firm, periph, plan, character, rel,

	order, entir, mild, fresh, pl, prim, method, attain, achief
19	Loc, addit, st, fin, short, august, split, netahtml, south, pating, compar, alt, rati, febru, issu, produc, centr, appl, start, showshoppingcart
20	Illustr, patfthdr, mas, occur, altern, found, lemon, design, andrew, fig, degree, direct, determin, thor, solis, sunrays, fiber, method, conduc, es
21	Wheel, rot, power, transfer, steer, unit, funct, controller, key, port, support, lin, mach, seat, steel, serv, ses, cover, act, lever
22	Retain, cont, lower, depend, bitter, genus, service, lin, patent, pt, middl, prim, acquir, consum, refer, random, attain, spain, thor, prior
23	Gam, mach, stor, port, unit, program, left, align, fig, player, displa, vide, oper, inform, sect, connect, proces, dat, inclus, shown
24	Comb, background, singl, moder, addit, determin, ebiz, propag, manner, origin, rapid, temperatur, rtyp, seed, character, weight, solis, trademark, equ, form
25	Common, intern, accompan, embed, markes, apr, netahtml, siz, left, stock, ident, robers, gif, requir, shap, assist, show, surfac, equ, count
26	Ferm, substr, stag, zymomon, mobil, ethanol, method, cel, sugar, med, produc, fermentor, rat, system, concentr, yeast, continu, part, separ, approxim
27	District, definit, origin, cel, parent, background, jun, apr, patent, hundr, equ, reg, maxim, upper, cas, october, border, miam, offic, abstract
28	Gam, system, dat, lcd, port, displa, object, fig, proces, dimens, player, devic, view, gener, siml, viewpoint, imag, world, display, stereoscop
29	Background, ad, thicker, addtoshoppingcart, held, mas, inform, botan, med, consist, targes, maxim, separ, afric, fin, cap, pronounc, hom, fnetacg, fam

**Experimento 2 – Partes da Patente 4876196**

Lista de Tópicos obtidos por meio do método *LDA*  
with Gibbs Sampling para as patentes

Nº USPTO da patente	Tópicos
4876196	26, 17, 12, 5, 13, 28, 21, 11, 18, 23, 27, 19, 29, 22, 9, 1, 20, 7, 25, 16, 14, 6, 2, 10, 8, 0, 3, 15, 4, 24
PP04161	0, 15, 19, 25, 7, 6, 23, 20, 4, 18, 10, 17, 9, 22, 14, 1, 8, 27, 3, 2, 29, 24, 16, 5, 26, 21, 11, 28, 12, 13
PP06047	0, 15, 5, 29, 1, 3, 8, 2, 27, 4, 25, 9, 17, 20, 22, 6, 7, 10, 24, 18, 16, 14, 12, 19, 28, 26, 23, 13, 21, 11

PP07651	0, 15, 27, 10, 20, 3, 6, 14, 19, 25, 7, 18, 16, 22, 24, 23, 4, 29, 1, 9, 8, 26, 5, 12, 2, 13, 21, 28, 11, 17
PP07700	0, 15, 2, 14, 18, 20, 24, 9, 7, 1, 29, 8, 19, 6, 4, 25, 3, 22, 16, 27, 10, 13, 28, 26, 23, 11, 12, 5, 17, 21
PP08212	15, 0, 26, 1, 2, 25, 23, 12, 4, 18, 19, 13, 8, 9, 24, 16, 22, 20, 11, 7, 27, 3, 14, 10, 28, 6, 17, 29, 21, 5
PP08238	15, 0, 13, 4, 10, 24, 25, 16, 8, 5, 3, 20, 19, 18, 23, 6, 1, 9, 29, 27, 2, 7, 22, 11, 17, 26, 14, 21, 12, 28
PP11246	0, 15, 26, 12, 22, 16, 9, 20, 25, 23, 7, 27, 4, 18, 10, 24, 3, 29, 8, 2, 14, 6, 19, 1, 5, 11, 13, 21, 17, 28
PP18774	0, 15, 16, 20, 1, 29, 7, 10, 4, 2, 22, 25, 5, 14, 8, 18, 27, 24, 19, 3, 6, 9, 17, 28, 26, 21, 11, 12, 23, 13
PP19575	0, 15, 23, 28, 25, 4, 12, 24, 18, 17, 6, 22, 10, 14, 2, 27, 3, 16, 11, 1, 9, 29, 19, 20, 13, 7, 8, 21, 26, 5
4326036	17, 0, 26, 12, 5, 23, 28, 21, 13, 24, 11, 10, 18, 7, 3, 20, 14, 1, 29, 19, 27, 16, 4, 2, 25, 9, 6, 22, 8, 15
4560659	12, 0, 17, 26, 23, 5, 13, 11, 28, 6, 25, 29, 1, 19, 8, 4, 27, 7, 21, 24, 15, 20, 16, 14, 9, 10, 2, 3, 18, 22
4738930	5, 0, 17, 26, 23, 13, 12, 28, 11, 21, 24, 7, 4, 19, 14, 1, 20, 25, 2, 18, 22, 8, 6, 29, 10, 9, 27, 3, 16, 15
4876196	26, 0, 17, 12, 5, 11, 23, 13, 28, 22, 6, 10, 9, 29, 18, 3, 4, 7, 1, 14, 2, 20, 25, 16, 8, 15, 27, 24, 19, 21
5428528	23, 0, 11, 13, 28, 21, 9, 5, 19, 12, 27, 14, 24, 10, 29, 1, 3, 16, 2, 6, 17, 4, 7, 18, 25, 8, 20, 22, 26, 15
6039574	13, 0, 23, 28, 11, 21, 15, 6, 5, 7, 25, 29, 20, 8, 3, 2, 9, 24, 12, 27, 17, 10, 26, 18, 4, 14, 1, 16, 22, 19
6416410	11, 0, 23, 28, 13, 21, 15, 27, 19, 8, 5, 29, 24, 17, 25, 20, 9, 14, 18, 26, 6, 12, 16, 7, 22, 4, 3, 2, 1, 10
6478583	13, 0, 23, 28, 11, 15, 5, 21, 26, 22, 10, 24, 18, 14, 1, 19, 16, 7, 2, 9, 27, 20, 3, 6, 8, 4, 17, 25, 29, 12
6500070	23, 0, 28, 13, 11, 26, 21, 17, 2, 8, 22, 16, 10, 25, 29, 5, 27, 20, 4, 18, 14, 9, 24, 7, 3, 1, 6, 19, 15, 12
7316618	0, 21, 23, 28, 13, 11, 5, 10, 14, 6, 19, 20, 2, 7, 16, 27, 1, 29, 8, 25, 24, 4, 9, 3, 18, 22, 12, 15, 17, 26

7371163	28, 0, 23, 13, 11, 21, 17, 5, 9, 25, 22, 27, 26, 14, 1, 8, 29, 4, 18, 20, 19, 16, 2, 3, 6, 24, 12, 10, 7, 15
---------	--

Da mesma forma que a anterior, com os testes realizados no Capítulo 6.4.2 utilizando o método *LDA with Gibbs Sampling*, obtiveram-se as seguintes tabelas e listas de tópicos para cada um dos três experimentos de grupos de patentes semelhantes.

<b>Experimento 1</b>	
Tópicos obtidos por meio do método <i>LDA with Gibbs Sampling</i>	
Tópico	Termos
0	Untransl, termin, vasil, matur, duplic, unit, plasm, glycin, grown, lack, potat, lat, report, consider, ens, constitute, test, element, quantif, nontransform
1	Round, origin, uptake, grow, herbic, order, cons, ribulos, im, herb, part, particl, codon, cycl, fam, bevan, viridochromogen, plastid, domin, antibiot
2	Herbic, art, port, antisens, trait, lawt, pair, term, germin, streptomyc, advantag, fragm, circl, el, segm, arabidops, methotrec, grain, repeat, toler
3	Metal, reporter, tumefacien, cons, extens, embodiment, deriv, rn, accumul, suit, broad, arrang, emplo, continu, propag, fre, residu, temper, barle, herb
4	Forml, serv, broad, undifferenti, start, polypept, anthocyanin, transform, extens, fragm, streptomyc, cre, norm, herbic, sorghum, cycl, respect, like, mos, typ
5	Cel, cal, constitute, chlorophyl, replac, frequ, southern, chemic, list, plant, firef, protoplast, left, evalu, transit, rapid, detail, import, method, maiz
6	Arabidops, morpholog, fragm, cons, matur, phosphinothricin, larg, distinct, field, fluoresc, stalker, prev, correl, mad, soybean, fus, fri, induc, uid, propers
7	Thalian, technolog, rol, fragm, corn, reg, rec, leav, ori, induc, member, origin, apparatus, wac, import, embryo, herbic, secons, deriv, belief
8	Abil, cal, herbic, common, arabidops, cycl, plasmid, expos, consider, bouchez, therefrom, encompas, mrn, proper, pe, ribulos, eukaryot, pl, oil, modif
9	Epsp, glyphos, claim, chimer, promoter, reg, transit, polypept, chloroplast, resist, pept, sequ, camv, virus, cod, vect, phosph, ctp, mut, bacter
10	Inhibitor, propos, broad, lys, complement, understood, precur, exogen, agent, random, vari, alfalfa, food, self, morpholog, desir, start, accumul, exposure, subsequ
11	Fam, creat, photosynthes, construc, harvest, routin, comb, identif, defin, cost, usag, epsp, monocotyledon, confirm, wheat, adh, eas, subunit, val, transl
12	Plant, cel, dn, gen, expres, transform, sequ, encod, inclus, pres, method,

	select promoter, invent, acids, maize, tissue, comprised, resist, transgene
13	Gene, introduction, medium, culture, production, recipe, protein, corn, contemplation, fertile, insect desire, may, zein, proposal, improve, marker, resist, transgene, transform
14	Cap, pollen, Arabidopsis, moss, peptide, util, wheat, brought, production, tungsten, film, soil, rough, design, threonine, character, broad, methodologist, abstract, tumefaciens
15	Fus, consist, synthetase, part, prior, segreg, chloroplast, contact, cauliflower, replicate, agriculture, peptide, uptake, permis, understood, steifel, en, surprise, transport, practice
16	Maize, phosphinothricin, bar, fertile, patent, acetyl, integr, transferase, transgene, yield, cross, event, design, plant, seed, atcc, deposit, claim, genotype, backcross
17	Herbicide, toler, nutri, arginine, transit, var, fact, maxim, glutamine, suscept, varies, data, agent, pr, reflect, methion, reg, mutagenesis, neces, chabaut
18	Tryptophan, anthranil, synthase, analog, free, inhibit, casses, amine, lin, methyltryptophan, feedback, substant, segm, toler, amount, molecule, transit, resist, overproduction, standard
19	Optim, pathwa, real, condit, val, herbicide, duplic, tabl, fus, rapid, linker, mut, consider, maize, leaf, maxim, defici, discuss, particl, oxygenase
20	Glyphos, abstract, entire, length, report, assay, tr, alter, sunflower, morphological synthesis, rapid, rat, immedi, cons, format, contempl, catechol, vector, borer
21	Synthetase, expect, petun, prim, prior, maxim, untransl, rat, accumul, glycine, inherit, serv, wheat, interact, segreg, number, lys, apic, quantif, sweets
22	Seed, protein, preselect, sequ, zein, stor, amine, weight, es, acids, perc, substant rn dn, prefer, amount, polypept, alph, dies, anim
23	Co, acetyl, carboxylase, oil, casses, herb, toler, cont, alter, herbicide, accas, expres, molecule, amount, act, mut, funct, fat, transit, chloroplast
24	Unit, accumul, larger, inhibitor, untransl, adjust, stain, var, phas, glutam, accord, stor, refer, port, vir, expans, soybean, cost, sect, rapid
25	Recombin, corn, ze, may, endotoxin, fertile, preselect, thuringiensis, bacillus, callus, transgene, insect, procedur, methion, tox, reporter, pest, lys, transmis, genotyp
26	Accas, seq, polypept, nucle, acids, reg, co, acetyl, segm, herb, sequ, segment, carboxylase, cod, yeast, amplif, rn, strand, primer, vect
27	Segm, preselect, water, stress, monocot, osmoprotect, casses, sugar, mannitol, avail, drought, physiol, deficit, toler, osmot, catalys, perform, untransform, mol, salt
28	Polyadenyl, deriv, enh, fragm, precis, polypept, human, attach, order, wheat, untransl, optim, evalu, promoter, typhymur, pr, tumefaciens, affect, modific, refer
29	Releas, deposit, rol, loc, subunit, induc, nopal, tomat, klein, spectr, acetyl, transl,

regl, mol, agrobacter, leader, amin, con, disclos, rearrang

### Experimento 1

Lista de Tópicos obtidos por meio do método *LDA*  
with Gibbs Sampling para as patentes

Nº USPTO da patente	Tópicos
4940835	12, 9, 22, 18, 26, 13, 23, 0, 5, 27, 25, 2, 16, 11, 29, 14, 6, 7, 10, 19, 17, 24, 8, 28, 15, 3, 4, 1, 21, 20
6268550	12, 23, 22, 26, 18, 9, 16, 27, 7, 13, 11, 10, 29, 3, 5, 6, 0, 14, 28, 21, 1, 17, 20, 4, 8, 15, 2, 24, 25, 19
6271016	12, 18, 22, 13, 27, 23, 9, 25, 26, 16, 2, 11, 6, 1, 29, 20, 5, 8, 7, 10, 17, 0, 28, 21, 19, 14, 3, 15, 24, 4
6281411	12, 27, 13, 22, 18, 23, 25, 26, 16, 9, 2, 15, 29, 11, 1, 21, 17, 20, 7, 3, 28, 0, 8, 4, 5, 19, 10, 6, 14, 24
6306636	12, 26, 22, 23, 13, 9, 18, 8, 27, 10, 16, 17, 19, 20, 0, 14, 15, 29, 25, 21, 11, 24, 3, 7, 28, 4, 1, 5, 6, 2
6326527	12, 22, 13, 16, 18, 26, 27, 9, 23, 25, 17, 10, 0, 28, 29, 2, 19, 15, 21, 5, 7, 14, 6, 3, 24, 11, 8, 20, 4, 1
6329574	12, 13, 22, 16, 25, 18, 9, 26, 27, 23, 7, 21, 1, 4, 3, 17, 29, 28, 15, 10, 0, 24, 19, 2, 5, 8, 6, 11, 20, 14
6331665	12, 25, 13, 22, 16, 9, 18, 27, 26, 6, 11, 23, 20, 19, 17, 3, 24, 7, 8, 28, 5, 4, 14, 15, 29, 1, 0, 10, 21, 2
6362396	12, 9, 22, 23, 26, 18, 13, 24, 16, 19, 7, 11, 1, 8, 10, 2, 29, 3, 17, 14, 28, 27, 20, 21, 4, 15, 6, 5, 0, 25
6395966	16, 12, 22, 13, 26, 25, 23, 18, 9, 0, 29, 19, 27, 28, 15, 21, 2, 4, 11, 8, 14, 10, 24, 7, 20, 6, 1, 5, 3, 17
6399861	12, 13, 22, 16, 26, 27, 23, 9, 18, 25, 15, 17, 19, 0, 8, 21, 29, 7, 3, 6, 10, 5, 20, 11, 24, 28, 4, 2, 1, 14

### Experimento 2

Tópicos obtidos por meio do método *LDA*  
with Gibbs Sampling

Tópico	Termos
0	Eukaryot, purif, hydroxylas, phys, precipit, fir, mees, dn, expect, succes,



	dicotyledon, belong, assa, tomat, distribut, draw, an, parallel, sum, pres
1	Convent, exist, natl, untransl, wheat, sequo, clas, tradit, pas, read, digest, entires, multipl, illustr, proxim, intracelll, respons, norm, antibiot, inst
2	Starch, amylos, corn, branch, amylopectin, enzym, actu, industr, seed, alph, sb, refer, cod, chain, isoform, grain, mut, suppres, biol, cont
3	Mammal, edit, advant, morpholog, leader, catalys, receiv, var, polypept, alvorad, wish, rapeseed, equivalent, lag, sourc, length, pack, chang, transform, regener
4	Chang, kanamycin, restrict, activ, giv, agrobacter, test, trait, bear, acces, exogen, transfect, public, repeat, electropor, vers, rn, way, norm, subunit
5	Mak, antisens, op, transferas, practic, col, employ, pack, famili, proc, cover, exist, component, in, prev, synthas, nat, messenger, activ, varies
6	Rna, pathwa, rd, protein, antibod, polypept, compound, identif, mut, interfer, method, dsrn, act, screen, component, fus, genet, bind, elegan, molecl
7	Subunit, uniqu, cal, revers, biosynthes, numer, structur, electrophores, embryo, illustr, react, grow, spring, minim, dr, fre, codon, fragment, neg sod
8	Desaturas, acis, fat, composit, puf, delt, polyunsatur, omeg, substr, convers, host, pres, nutrit, enzym, isol, oil, anim, produc, carbon, ident
9	Melt, enzym, light, stor, fig, trens, virus, protect, test, starch, wish, fragm, allel, du, coat, eas, octop, growth, engineer, metabol
10	Construc, propers, cod, mammal, messenger, develop, famili, shock, releas, wheat, increas, accompl, gold, phosphatas, correspons, inactiv, acad, reduc, prev, loc
11	Dist, detail, exist, develop, propers, rat, caus, natur, acis, respons, fruit, cuphe, prior, subject, prov, grapefruit, mar, minim, funct, comples
12	Transgen, lys, maiz, zein, seed, dn, plant, suppres, biosynthes, reduc, cont, increas, orient, inclus, synerg, cordap, mrn, genom, sens, patent
13	Targes, repres, organ, diagram, structur, synthas, plasmid, molecl, anim, foreign, dispers, eukaryot, gen, cop, pcmv, multipl, tissu, vir, genet, virus
14	Indigen, ant, messenger, dsdn, polygalacturonas, conveni, fruit, border, transcrib, rip, tomat, dicotyledon, fewer, pcgn, repetit, join, employ, plur, heat, trait
15	Matur, reductas, combin, control, initi, definit, reduc, smal, progen, host, biosynthes, camv, transcrib, featur, solis, natur, downstream, pear, interest, applic
16	Nat, propers, pin, harbor, repetit, fat, border, ori, unit, monocotyledon, consider, distribut, polymeras, el, plur, gal, segment, ant, substr, family
17	Entires, employ, repres, percent, spray, ant, defin, replic, modific, deposit,

	predict, phosph, polyadenyl, center, aden, recipi, foreign, high, plur, bin
18	Ratio, encompass, soil, control, tomat, biol, tumefacien, microinject, mol, rtm, assa, agrobacter, draw, level, polyadenyl, light, restrict, transcrib, embryo, process
19	Tobacc, plant, nicot, demethylas, desir, seq, nicotian, acis, expres, inclus, nucle, sequ, method, aspect, featur, ident, reduc, induc, molecl, rn
20	Tobacc, nicot, qprt, transferas, phosphoribosyl, quinol, nic, ntqpt, transgen, level, exogen, increas, root, enzym, transform, cigares, cigar, cont, car, biosynthes
21	Genotyp, linker, entir, promoter, cit, tomat, join, mammal, targes, mos, understand, cons, reductas, ribosom, pept, short, pal, gener, prov, deles
22	Benzo, promoter, induc, embodiment, protein, gad, hybrid, refer, interest, pres, acis, fact, seq, respons, br, term, transcript, system, compris, control
23	Oxidas, conserv, trait, transfect, exogen, monocotyledon, plur, duplec, standard, practic, us, chang, downstream, color, fat, understood, carboxylas, differ, exemplif, analog
24	Recogn, asparagus, gal, read, recomb, background, endogen, hydroxylas, cros, put, plasmid, precipit, addit, injur, prefer, differ, exogen, morpholog, pecan, popl
25	Ssu, monocotyledon, valu, acad, fruit, breed, ant, expect, sens, found, messenger, water, integr, screen, lack, succes, exampl, treat, purif, ec
26	Lignin, gymnosperm, syringyl, casses, angiosperm, sweetgum, lolol, pin, gsl, insers, fus, illustr, lan, loc, biosynthes, genom, asl, prob, omt, flank
27	Sum, monocotyledon, alcohol, cod, repetit, predict, integr, exist, deriv, plur, test, norm, long, join, bons, chos, fre, correspons, phys, green
28	Sequ, gen, plant, cel, expres, nucle, acis, invent, promoter, produc, dn, inclus, reg, pres, compris, encod, rn, method, molecl, transcript
29	Understand, employ, codon, pack, support, claim, convent, barle, natl, repetit, flavon, sambrook, tumefacien, color, mil, chart, sugar, port, funct, suit

### Experimento 2

Lista de Tópicos obtidos por meio do método *LDA*  
with Gibbs Sampling para as patentes

N° USPTO da patente	Tópicos
5107065	28, 14, 26, 20, 6, 13, 19, 22, 12, 27, 23, 24, 2, 8, 16, 0, 11, 10, 3, 9, 5, 7, 4, 25, 29, 18, 15, 21, 1, 17
7645925	28, 20, 19, 12, 6, 22, 13, 14, 8, 2, 26, 17, 9, 1, 15, 21, 11, 4, 5, 7, 0, 10, 29, 27, 23, 16, 24, 25, 18, 3

7683237	12, 28, 19, 2, 13, 22, 8, 6, 20, 17, 23, 18, 24, 26, 29, 15, 9, 14, 16, 10, 5, 21, 27, 4, 0, 25, 11, 3, 7, 1
7700834	19, 28, 22, 6, 13, 2, 12, 8, 20, 26, 23, 25, 15, 1, 16, 7, 18, 10, 3, 9, 14, 29, 5, 0, 11, 27, 17, 24, 4, 21
7700851	19, 28, 6, 22, 13, 12, 2, 8, 20, 3, 25, 15, 1, 4, 26, 10, 16, 18, 21, 27, 23, 9, 11, 24, 14, 17, 7, 0, 5, 29
7705203	22, 28, 19, 6, 13, 12, 2, 5, 26, 4, 29, 8, 16, 0, 18, 7, 21, 1, 11, 14, 24, 9, 15, 23, 25, 27, 3, 10, 17, 20
7723503	28, 8, 2, 19, 22, 13, 6, 12, 10, 4, 29, 25, 27, 23, 9, 16, 17, 15, 18, 1, 3, 0, 7, 24, 11, 21, 5, 14, 20, 26
7754697	28, 13, 22, 19, 12, 6, 8, 5, 25, 7, 23, 18, 15, 21, 11, 27, 1, 4, 2, 14, 29, 17, 10, 9, 20, 24, 3, 26, 0, 16
7754869	28, 26, 22, 12, 6, 19, 20, 2, 13, 8, 7, 10, 15, 27, 29, 4, 0, 16, 21, 18, 5, 14, 1, 11, 23, 24, 3, 17, 9, 25
7754942	28, 2, 12, 13, 19, 22, 8, 6, 11, 7, 29, 15, 17, 20, 25, 18, 27, 1, 23, 14, 9, 10, 3, 16, 5, 24, 21, 4, 0, 26
7759463	6, 28, 19, 13, 22, 12, 2, 11, 25, 8, 27, 10, 16, 20, 1, 3, 15, 5, 9, 23, 21, 0, 17, 7, 4, 26, 29, 14, 24, 18

### Experimento 3

Tópicos obtidos por meio do método *LDA*

*with Gibbs Sampling*

Tópico	Termos
0	Ctb, autoantig, protein, fus, diseas, autoimmun, invent, vaccin, antibod, ed, transgen, potat, construc, oral, chimera, pat, cod, compris, induc, autoantigen
1	Hsp, molecl, promot, standard, borer, progres, researches, report, result, mendu, subst, antisens, structur, embodiment, prototroph, tg, neg, altschl, simultan, manner
2	Henikoff, ribozym, therewith, major, broad, progres, grown, luciferas, enter, agent, visu, adenyl, brassic, van, wordlength, component, lambda, transport, wood, inact
3	Stud, relationship, coordin, convers, tfast, scal, publ, upsteam, tobpmt, soybean, thought, situ, transloc, cap, nematocis, velt, yield, extens, hohn, prk
4	Intergen, list, cit, categor, flac, depens, metallothionein, shorter, vari, visu, pattern, rout, nitrilas, nlm, cashmor, blast, hsp, valle, precipit, convers
5	Visu, chem., impart, enter, ausubel, higgina, nacl, reduc, coordin, promoter, secres, rna, cashmor, aver, respect, spur, mendu, nitrilas, embryogenes, describe

6	Herb, plant, resist, genet, glyphos, seed, wheat, cotton, soybean, control, gen, natur, engineer, pres, occur, acis, invent, result, produc, select
7	Segm, background, book, sensit, abandon, metallothionein, entir, init, convers, formam, farm, search, sedim, psc, cuml, assembl, promoter, accompl, coeffici, vector
8	Tree, algorithm, inact, yamamot, cuml, categor, math, concers, transfect, tobrb, researches, minor, softwar, match, deposit, simplif, modific, sedim, catalys, camv
9	Continu, polypept, subsequ, wordlength, event, col, lepidopter, propagl, render, technolog, prokaryot, quant, softwar, fast, limis, inspect, dendogram, ed, viroid, cop
10	Duckweed, transform, tissu, nucle, callus, select, method, step, agrobacter, pept, cultur, interest, st, microprojectil, med, frons, prefer, targes, altern, transfer
11	Simplif, satisfactor, mad, upsteam, attack, altschl, progen, rna, hpt, neomycin, speed, green, except, initi, sak, databas, norm, prototroph, sal, sphere
12	Blast, extern, threshold, cover, disregard, celll, adv, repeat, column, reg, glucuronidas, plasm, elim, eukaryot, fern, complem, agent, alignment, microorgan, hit
13	Sal, lepidopter, ssc, enter, alg, ed, counterclock, pathogen, continu, neighbor, simpl, search, cytoplasm, render, incub, nightshad, pathwa, gap, bind, strateg
14	Vir, pept, parti, embodiment, virus, epitop, acis, exogen, protein, refer, coat, nucle, capsid, insers, term, pres, rn, seq, bes, sid
15	Penalt, categor, start, glucurodinas, continu, attack, primros, transloc, alg, softwar, convers, researches, genus, breed, counterclock, pileup, inspect, immedi, strawber, bisphosph
16	Sedim, relationship, enolpyruvylphosphoshik, auxotroph, denhardt, intermedi, flower, pileup, simplif, prokaryot, practic, sequ, word, recogn, pot, categor, prk, var, short, threshold
17	Progres, alignment, sensit, evid, hit, counterclock, solanac, upsteam, maniat, softwar, ribozym, gold, supr, messenger, crystal, alfalf, corn, hour, tumor, conifer
18	Nucle, pres, dn, termin, membran, seq, gen, sod, clon, dna, hybrid, prepar, host, funct, shown, phag, solut, vect, degree, obtain
19	Plant, gen, sequ, cel, promoter, invent, dn, expres, transform, pres, protein, method, compris, produc, inclus, acis, nucle, prefer, construc, transcript
20	Immedi, tree, practic, nematod, var, conkl, homolog, cat, auxili, rout, amin, standard, casses, succeed, entit, sulf, rm, except, embodiment, individu
21	Tobacc, nicot, nic, tsn, ci, reduc, act, el, produc, regl, level, amount, ppm, interest, element, qptas, pmtas, ppb, embodiment, transgen
22	Sugarbees, pos, parent, hypocotyl, green, parameter, matric, metallothionein, wordlength, pair, counterclock, blosum, lambd, pathwa, altschl, nopal, relationship, aver, vitr, drawn

23	Altschl, pearson, metallothionein, ncb, nacl, pe, schilperoort, weight, hypocotyl, default, mam, repeat, strength, pacyc, bisphosph, ultim, nematod, doolittl, quant, thereof
24	Hamper, conduc, accompl, affect, berg, particl, growth, pileup, hit, rapid, wordlength, belief, sens, insecticis, sedim, repres, linker, top, lys, por
25	Por, practic, schem, transloc, cit, hereinafter, cluster, softwar, degre, initi, drought, perc, fram, residu, longer, metallothionein, scor, fast, string, inspect
26	Hit, tobrb, except, solanac, compound, yield, contribut, disclosur, strateg, indig, phag, pmon, wisconsin, hamper, chlorophyl, wood, lack, short, drawn, progen
27	Weight, coeffici, mos, ap, blastn, oat, glyphosph, attack, potat, technolog, program, concern, funct, proper, ultim, computer, book, hsv, liter, broad
28	Epsp, glyphos, seq, clas, gen, enzym, sequ, acis, amin, pmon, col, toler, pep, strain, prob, minut, ctp, sit, isol, chloroplast
29	Perc, feng, shorter, ntqpt, categor, alignm, pe, publ, ausubel, nitrilas, simpl, cold, lif, celll, softwar, polymeras, threshold, unit, rna, cytoplasm

### Experimento 3

Lista de Tópicos obtidos por meio do método *LDA*  
with *Gibbs Sampling* para as patentes

Nº USPTO da patente	Tópicos
5352605	19, 14, 28, 18, 6, 13, 10, 26, 2, 4, 15, 9, 24, 17, 3, 1, 7, 27, 25, 12, 20, 23, 5, 16, 8, 11, 22, 29, 0, 21
6777546	0, 19, 14, 6, 18, 10, 21, 28, 24, 11, 7, 29, 22, 25, 13, 20, 4, 23, 17, 2, 1, 9, 5, 27, 26, 16, 15, 12, 3, 8
7074987	6, 19, 18, 14, 10, 28, 0, 21, 13, 20, 2, 15, 1, 8, 17, 9, 12, 26, 25, 3, 4, 11, 29, 27, 22, 7, 16, 24, 5, 23
7087809	6, 19, 28, 14, 18, 10, 21, 0, 12, 23, 29, 5, 2, 3, 22, 16, 7, 9, 13, 11, 25, 20, 1, 27, 24, 15, 26, 8, 4, 17
7135282	14, 19, 18, 6, 28, 0, 10, 21, 25, 7, 2, 22, 13, 8, 3, 16, 17, 9, 12, 5, 4, 29, 1, 27, 11, 15, 23, 20, 26, 24
7135626	6, 19, 28, 14, 18, 0, 10, 21, 23, 5, 20, 26, 27, 29, 7, 2, 8, 16, 9, 12, 25, 1, 17, 4, 13, 11, 22, 3, 15, 24
7161064	10, 19, 6, 14, 28, 18, 0, 21, 29, 27, 9, 13, 1, 16, 15, 12, 11, 25, 2, 22, 20, 24, 8, 5, 4, 7, 26, 3, 17, 23
7183110	28, 19, 6, 18, 14, 10, 27, 0, 22, 16, 11, 3, 17, 9,

	13, 26, 7, 8, 4, 23, 25, 29, 12, 20, 2, 5, 21, 24, 1, 15
7189570	19, 21, 10, 18, 6, 28, 14, 15, 8, 22, 17, 3, 13, 25, 23, 29, 12, 11, 20, 27, 2, 7, 26, 4, 24, 1, 16, 9, 5, 0
7192771	21, 19, 6, 10, 18, 14, 28, 0, 16, 20, 13, 4, 27, 2, 25, 5, 9, 11, 24, 29, 17, 1, 3, 22, 7, 8, 15, 26, 23, 12
7202083	18, 19, 6, 28, 14, 10, 0, 21, 17, 9, 5, 26, 22, 24, 12, 3, 7, 2, 16, 8, 1, 15, 20, 29, 13, 23, 11, 4, 27, 25