UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS SOCIAIS APLICADAS
CURSO DE GESTÃO DA INFORMAÇÃO

GUILHERME PATRICIO SILVEIRA

MINERAÇÃO DE TEXTOS APLICADA A BASES DE DADOS JURÍDICAS

CURITIBA
2013

GUILHERME PATRICIO SILVEIRA

MINERAÇÃO DE TEXTOS APLICADA A BASES DE DADOS JURÍDICAS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção de grau no Curso de Gestão da Informação, Departamento de Ciência e Gestão da Informação do Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Orientadora: Profª. Drª. Denise Fukumi Tsunoda.

CURITIBA

2013

AGRADECIMENTOS

A minha família e amigos pelo incentivo e apoio.

À professora Denise Fukumi Tsunoda pelos ensinamentos e orientações.

RESUMO

Pesquisa experimental em mineração de texto aplicada a uma base de documentos de jurisprudências relacionadas ao comércio eletrônico. Objetiva aplicar técnicas de mineração de texto na tentativa de prover descobertas relevantes nos documentos da base citada. Explora as características da mineração de dados e áreas do conhecimento relacionadas. Agrega valor científico a partir de pesquisas acerca das propriedades da Informação, da recuperação da informação, da mineração de dados e da mineração de texto, bem como a breve descrita do poder judiciário brasileiro. Descreve as funcionalidades do software PreText, utilizado no presente trabalho, e como o mesmo se aplica ao presente contexto. Descreve os resultados de forma detalhada utilizando métricas de valoração e contagem de termos, além de descrever interpretações por análise associativa. Obtém como resultados a construção de n-gramas, sendo esses as principais fontes para a análise.

Palavras Chave: Mineração de Texto. Mineração de Dados. Pesquisa Exploratória. Gestão da Informação.

ABSTRACT

Experimental Research on Text Mining applied to a document base made by case laws related to e-commerce. Intends to apply text mining as an attempt to provide relevant discoveries in the documents base of this paper. Explores data mining features an related knowledge areas. In order to add scientific value, explains the proprieties of information, information retrieve, Data Mining and Text Mining. Describe the Pretext Software features used in this paper, and how it applies on the context of the document base. By the associative analyses, shows the results of the text mining methods As a result obtains the construction of n-gram, these being the main source for analysis.

Key Words: Text Mining. Data Mining. Experimental Research. Information Management.

LISTA DE FIGURAS

FIGURA 1 - SEXTA LEI DE MOODY E WALSH 13
FIGURA 2 - FLUXOGRAMA DOS TRÊS PROCESSOS BÁSICOS DA RECUPERAÇÃO DA INFORMAÇÃO 22
FIGURA 3 - RECUPERAÇÃO DA INFORMAÇÃO – MODELO BOOLEANO, OPERADOR LÓGICO E 23
FIGURA 4 - RECUPERAÇÃO DA INFORMAÇÃO: MODELO BOOLEANO, OPERADOR LÓGICO OU 23
FIGURA 5 - FIGURA 5 – RECUPERAÇÃO DA INFORMAÇÃO: MODELO , OPERADOR LÓGICO NÃO 24
FIGURA 6 - RECUPERAÇÃO DA INFORMAÇÃO: MODELO BOOLEANO,OPERADOR LÓGICO NÃO, CASO T1 E NÃO T2 24
FIGURA 7 - EQUAÇÃO DO MODELO ESPAÇO VETORIAL 25
FIGURA 8 - SISTEMA ESPECIALISTA 27
FIGURA 9 - RECUPERAÇÃO DA INFORMAÇÃO: REDES NEURAIS 28
FIGURA 10 - ALGORITMOS GENÉRICOS 29
FIGURA 11 - NUVEM DE TAG 30
FIGURA 12 - KDD 32
FIGURA 13 - PROCESSO DE MINERAÇÃO DE TEXTO 35
FIGURA 14 - MODELO BÁSICO DE MINERAÇÃO 39
FIGURA 15 - ORGANOGRAMA BÁSICO DO PODER JUDICIÁRIO 41
FIGURA 16 - PROCESSO DE MINERAÇÃO DE TEXTO DA PRESENTE BASE 45
FIGURA 17 - PRETEXT II 47
FIGURA 18 - TELA DE CONFIGRAÇÃO DO PRETEXT III 48
FIGURA 19 – EXEMPLO DE *STEM*MING 49
FIGURA 20 – 1 GRAMA: SER 57

LISTA DE GRÁFICOS

GRÁFICO 1 - FREQUENCIA 1 GRAMA: RECURSO NEGADO (TFIDF) 57
GRÁFICO 2 – 1 GRAMA: GRÁFICO TERMO X FREQUÊNCIA – RECURSO NEGADO.....58
GRÁFICO 3 - FREQUENCIA 1-GRAMA: COMÉRCIO ELETRÔNICO: RECURSO PROVIDO (TFIDF) 59
GRÁFICO 4 – 1 GRAMA: GRÁFICO TERMO X FREQUÊNCIA – RECURSO NEGADO.....60
GRÁFICO 5 - FREQUÊNCIA 4 GRAMA: COMÉRCIO ELETRÔNICO RECURSO NEGADO (TFIDF) 61
GRÁFICO 6 - 4-GRAMA: COMÉRCIO ELETRÔNICO GRÁFICO TERMO X FREQUÊNCIA - RECURSO NEGADO 62
GRÁFICO 7 - FREQUENCIA 4 GRAMA: COMÉRCIO ELETRÔNICO RECUROS PROVIDO (TFIDF) 63
GRÁFICO 8 – 4-GRAMA: COMÉRCIO ELETRÔNICO GRÁFICO TERMO X FREQUÊNCIA – RECURSO PROVIDO 64
GRÁFICO 9 - FREQUENCIA 1 GRAMA: COMPRAS PELA INTERNET (TFIDF).....65
GRÁFICO 10 - 1-GRAMA: GRÁFICO TERMO x FREQUËNCIA - COMÉRCIO ELETRÔNICO.....65
GRÁFICO 11 - FREQUENCIA 1 GRAMA: E-COMMERCE PROVIMENTO NEGADO (TFIDF).....66
GRÁFICO 12 - 1-GRAMA: E-COMMERCE GRÁFICO TERMO x FREQUËNCIA - RECURSO NEGADO.....67
GRÁFICO 13 - FREQUENCIA 1 GRAMA: E-COMMERCE RECURSO PROVIDO (TFIDF).....68
GRÁFICO 14 - 1-GRAMA: E-COMMERCE GRÁFICO TERMO x FREQUËNCIA - RECURSO PROVIDO.....68
GRÁFICO 15 - FREQUENCIA 1 GRAMA GERAL: RECURSO NEGADO (TFID).....69
GRÁFICO 16 - 1-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO NEGADO.....70
GRÁFICO 17 - FREQUENCIA 1 GRAMA GERAL: RECUROS PROVIDO (TFIDF) .71

GRÁFICO 18 - 1-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO PROVIDO.....72
GRÁFICO 19- FREQUENCIA 4 GRAMA GERAL: RECURSO NEGADO (TFIDF) ...73
GRÁFICO 20 - 4-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO NEGADO.....74
GRÁFICO 21 - FREQUENCIA 4 GRAMA GERAL: RECURSO PROVIDO (TFIDF) .75
GRÁFICO 22 - 4-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO PROVIDO.....76

LISTA DE QUADROS

QUADRO 1 - FREQUENCIA DE TERMOS: COMÉRCIO ELETRÔNICO 52
QUADRO 2 - FREQUENCIA DE TERMOS: E-COMMERCE 53
QUADRO 3 - FREQUENCIA DE TERMOS: COMPRAS PELA INTERNET.....54
QUADRO 4 - FREQUENCIA DE TERMOS: COMPRAS PELA INTERNET.....55
QUADRO 5 - ANALISE 1 GRAMA: COMÉRCIO ELETRÔNICO 60
QUADRO 6 - ANÁLISE 1 GRAMA: TODOS OS DOCUMENTOS.....72
QUADRO 7 - ANALISE 4-GRAMA: RANKING DE JUÍZES 76

LISTA DE SIGLAS

FI – FORMA INTERMEDIÁRIA

JF – JUSTIÇA FEDERAL

KDD – DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

LABIC – LABORATÓRIO DE INTELIGÊNCIA COMPUTACIONAL

NASA – ADMINISTRAÇÃO NACIONAL DE AERONÁUTICA E ESPAÇO

OCR – RECONHECIMENTO ÓTICO DE CARACTERES

RI – RECUPERAÇÃO DA INFORMAÇÃO

SPT – SELEÇÕES DE PARTE DO TEXTO

STF – SUPREMO TRIBUNAL FEDERAL

STJ – SUPREMO TRIBUNAL DE JUSTIÇA

STM – SUPERIOR TRIBUNAL MILITAR

TB – TERMO DE BUSCA

TF – FREQUENCIA DO TERMO

TFIDF – FREQUENCIA DO TERMO – FREQUENCIA DO DOCUMENTO INVERSA

TJPR – TRIBUNAL DE JUSTIÇA DO PARANÁ

TRE – TRIBUNAL REGIONAL ELEITORAL

TSE – TRIBUNAL SUPERIOR ELEITORAL

URSS – UNIÃO DAS REPÚBLICAS SOCIALISTAS SOVIÉTICAS

SUMÁRIO

1 INTRODUÇÃO 11
1.1 PROBLEMA 12
1.2 JUSTIFICATIVA 14
1.3 OBJETIVOS 15
1.4 ESTRUTURA DO DOCUMENTO 16
2 FUNDAMENTAÇÃO TEÓRICA 17
2.1 DADO, INFORMAÇÃO, CONHECIMENTO E TECNOLOGIA 17
2.2 RECUPERAÇÃO DA INFORMAÇÃO 19
2.3 TIPOS DE RECUPERAÇÃO DA INFORMAÇÃO 21
2.3.1 Modelos Quantitativos 22
2.3.2 Modelos Dinâmicos 26
2.4 NUVEM DE *TAG* 30
2.5 MINERAÇÃO DE DADOS e KDD 31
2.6 MINERAÇÃO DE TEXTOS 34
2.6.1 Processo de descoberta de padrões em dados não estruturados 37
2.7 PODER JUDICIÁRIO 40
3 METODOLOGIA DA PESQUISA 42
3.1 BASES DE DADOS DE PROCESSOS JURÍDICOS: COMÉRCIO ELETRÔNICO 43
3.2 FERRAMENTA PARA A MINERAÇÃO DE TEXTO 44
3.3 PROCEDIMENTOS METODOLÓGICOS 44
3.3.1 Software PreText 46
4 DESCRIÇÃO DO EXPERIMENTO E ANÁLISE DE RESULTADOS 51
4.1 ANÁLISE POR FREQUÊNCIA DE PALAVRAS 51
4.2 ANÁLISE TF-IDF E N-GRAMA 56
4.2.1 Análise Pretext: Grupo Comércio eletrônico 56
4.2.2 Análise Pretext: Grupo Compras pela internet 64
4.2.3 Análise Pretext: Grupo E-commerce 66
4.2.4 Análise Pretext: Processamento geral 69
4.3 SÍNTESE DAS ANÁLISES 77
5 CONSIDERAÇÕES FINAIS 78
REFERÊNCIAS 81
ANEXOS 84
APÊNDICES 113

1 INTRODUÇÃO

Imersas em informações de diferentes fontes, organizações e pessoas atentam cada vez mais para a gestão da informação. O problema principal deixou de ser o acesso à informação, mas sim o uso da mesma, dado, não somente a facilidade na recuperação e acesso a informação, mas também a produção e disseminação crescente de dados e informações.

Nesta era dita informacional, a comunidade científica se sentiu estimulada a criar ferramentas, métodos e conceitos que pudessem contribuir com o processo de gestão informacional. A Descoberta de Conhecimento em Banco de Dados, ou ainda *Knowledge Discovery in Databases* (KDD) foi um termo cunhado em 1989 com o objetivo de representar o processo de busca e extração de conhecimento que, em uma das fases de seu nível mais operacional, inclui a aplicação de técnicas e algoritmos de *data mining* (mineração de dados) para manipular e encontrar, por exemplo, indícios de correlação ou implicação em bases de dados.

O KDD tem como propósito a descoberta de conhecimento interessante e inédito dentro de uma base de dados. Este processamento passou a ser utilizado em maior escala nas organizações à medida que se descobriu que utilizar as informações de maneira correta passou a ser um diferencial competitivo, e ainda, um importante insumo estratégico.

Dada à relevância da etapa de mineração de dados (MD), alguns autores adotam o termo data mining para designar as mesmas atividades e procedimentos que se descreve como KDD. No entanto, para fins desta pesquisa, a MD é uma das etapas do KDD, conforme explicitado na seção 2.5.

Estudos relacionados à gestão da informação ganham notoriedade não somente pela comunidade científica informacional, mas também por todas as organizações e pessoas que veem a informação como um importante ativo. Estudos relacionados à visualização, segurança, uso, e descobertas dentro da mineração são cada vez mais importantes aos olhos do mundo.

No presente trabalho, foca-se em um dos ramos da ciência da informação, a mineração de texto. Diferentemente da mineração de dados, onde toda a ação ocorre dentro de uma base de dados estruturada, na mineração de texto as descobertas são realizadas em textos de forma livre. Portanto, para organizações de

diferentes mercados, a mineração de texto se torna uma importante ferramenta para suas ações.

Nesta pesquisa a mineração de texto é aplicada a uma base de processos jurídicos relacionados à compra pela internet. Os processos são extraídos do portal do Tribunal de Justiça do Estado do Paraná, localizado no endereço eletrônico http://portal.tjpr.jus.br/jurisprudencia. E se referem a casos em segunda instância.

As possíveis descobertas quanto às decisões de juízes não necessariamente corresponderão a um demonstrativo de casos concluídos, uma vez que as partes possuem o direito de recorrer sob a decisão da Turma Recursal.

Destaca-se que são coletados apenas os processos judiciais eletrônicos, pois a mineração de texto é aplicada em meio digital.

A partir de então, realizam-se todas as etapas do KDD, com o propósito de descobrir informações relevantes que antes não eram conhecidas ou explícitas.

A mineração de texto, de modo geral, pode ser utilizada para todos os conjuntos de textos correlacionados entre si. Exemplos são encontrados em pesquisas de mercado, padrões de consumidores, descobertas em determinado histórico de documentos etc. Sendo assim, pode-se considerar que os estudos acerca do tema possuem uma infinidade de temas possíveis, sendo esse apresentado, servente de estímulo para futuros estudos em outras áreas.

1.1 PROBLEMA

No presente trabalho, estudam-se as bases de processos jurídicos presentes em meio digital no portal da justiça federal do Paraná.

O advento da internet possibilitou com que quantidades imensas de informação estivessem disponíveis a todos que possuem acesso a esse meio. Logo, percebeu-se que este excesso informacional por muitas vezes é prejudicial.

Como exemplo, pode-se notar que muitas organizações procuram receber a maior quantidade de informação possível, de diferentes fontes, com a premissa de que estarão à frente no mercado. O que ocorre é que, na maioria das vezes, as organizações estão na verdade inundadas em um mar de múltiplas informações. O excesso prejudica a compreensão e a tomada de decisão tanto quanto a falta de informação.

Pesquisadores da Universidade de Melbourne, Moody e Walsh (1999), desenvolveram leis informacionais com o propósito de estudar os fenômenos da comunicação e informação. É na sexta lei onde se pode observar a teoria de que mais informação não necessariamente resulta em mais resultados, mas sim, na desvalorização da mesma.

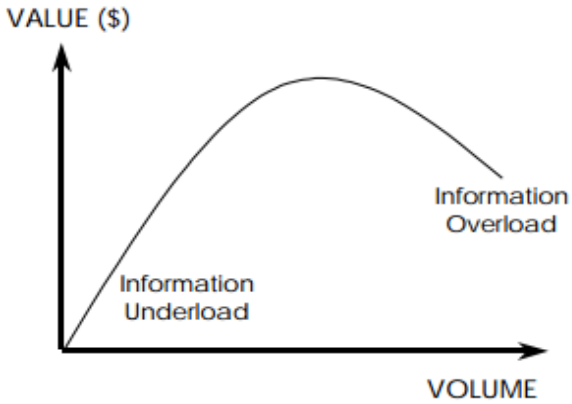

FIGURA 1 - SEXTA LEI DE MOODY E WALSH
FONTE: MOODY e WALSH (1999)

Segundo Moody e Walsh (1999), o excesso de informação prejudica pessoas e empresas à medida que o valor ou o conteúdo a ser extraído está segmentando em diferentes canais ou em excesso de quantidade.

Quando o assunto são os processos jurídicos se observa um cenário parecido com a sexta lei de Moody e Walsh (1999). É difícil realizar uma constatação a partir de uma base de processos jurídicos, pois a quantidade de informações apresentadas é grande.

Ressalta-se, ainda, o fato de que nem todos os processos de uma matéria estão disponíveis em ambientes físicos e eletrônicos, de fato, a digitalização de jurisprudências ainda se encontra em processo de evolução.

Assim como no mar de informações em que empresas vivem, os processos jurídicos presentes no Tribunal de Justiça do Paraná (TJPR), relacionados ao comércio eletrônico (base de dados do projeto) ainda não possuem estudos aprofundados de análises não quantitativas. Sendo assim, o cenário é de muitos arquivos armazenados, que dificultam a análise específica do tema.

Segundo dados do Instituto de Pesquisa Econômica Aplicada, publicados em 2012, 19% dos internautas brasileiros participam do comércio eletrônico. Em 2011, o setor movimentou R$ 18 bilhões. Em 2012, segundo a Braspag, o volume de

transações no ramo cresceu 46% em relação a 2011, estima-se que o setor detém a participação de 2% das vendas no varejo brasileiro. De fato, esses números tendem a crescer à medida que cada vez mais brasileiros possuem acesso a internet.

Por vezes, os consumidores se sentem lesados de tal forma a buscarem seus direitos perante a justiça. É crescente o número de processos relacionados à compra da internet, ao passo que o comércio eletrônico está também em uma crescente.

Assim, o trabalho abordará o seguinte problema: quais são os resultados da mineração de texto nas análises dos processos jurídicos registrados no estado do Paraná?

1.2 JUSTIFICATIVA

À luz da era da informação e da tecnologia da informação, os estudos que envolvem a gestão da informação ganham valor na medida em que a sociedade se beneficia das produções científicas do tema.

O presente estudo aplica a mineração de texto em bases de texto jurídicas, especificamente em processos judiciais relacionados a compra pela internet, com o propósito de prover descobertas em uma área que produz grandes quantidades de documentos textuais. Ilustrando a versatilidade do método aplicado a um segmento específico.

Uma das soluções para o problema de volume cada vez maior de informação foi o desenvolvimento da mineração. A técnica permitiu com que, a partir de métodos lógicos, fosse possível extrair de um determinado contingente de dados e descobertas que antes não eram previstas.

A mineração, seja de dados ou de texto, possui interdisciplinaridade com diversas áreas do conhecimento. Não é possível determinar com precisão todas as áreas de aplicação, pois todo dado registrado, independente de sua temática, pode ser tratado de modo a possibilitar a mineração.

Portanto, é imprescindível que o usuário do método tenha conhecimentos em gestão do conhecimento, recuperação da informação, gestão da informação, estatística, teoria da informação, mas ainda tão imprescindível quanto é o conhecimento do meio aos quais os dados são recuperados.

De uma visão holística, pode-se constatar que do mesmo modo que a mineração demanda a interpretação do meio informacional, a interpretação do meio informacional demanda a mineração.

O fato é que os múltiplos meios e canais informacionais criaram não uma rede, mas uma malha de informação. Informações com alto grau de valor (o valor dependerá do usuário) estão cada vez mais segregadas em diferentes meios. Um único “fio” dessa malha não permitirá com que o receptor diga o tamanho, a completude ou a forma de sua informação, mas sim o conjunto de todos eles.

Seguindo a analogia, é interessante conhecer todos os fios de sua malha, mas ainda mais interessante é saber quais deles mais caracterizam a mesma. Desse modo se norteia a mineração de dados e texto, e desse modo organizações e pessoas ganham em suas ações.

1.3 OBJETIVOS

Os objetivos deste trabalho estão divididos em objetivo geral e objetivos específicos. O objetivo geral consiste em aplicar a Mineração de Texto na tentativa de prover descobertas relevantes nos processos judiciais referentes ao comércio eletrônico.

Do objetivo geral derivam-se os específicos:

- realizar e descrever todo processo que envolve a mineração de texto;
- realizar uma pesquisa literária sobre o tema da mineração de texto indicando os principais estudos e teorias;
- demonstrar as descobertas a partir da base de texto minerada;
- demonstrar uma das possibilidades da mineração de texto;
- conhecer e realizar descoberta no âmbito judiciário paranaense.

1.4 ESTRUTURA DO DOCUMENTO

A primeira seção deste documento apresenta a introdução à pesquisa realizada, detalhando o problema e a justificativa bem como objetivos e delimitações do tema.

A fundamentação teórica, que explana os conceitos de dado, informação, conhecimento e tecnologia, necessários para a subsequência dos temas explanados, que consi*stem* na explicação do conceito e dos tipos de recuperação de informação. A mineração de dados, a base da mineração de texto, e o KDD, que ilustra processos estruturados que norteiam um projeto de mineração introduzem a mineração de texto, principal ferramenta para a realização do projeto.

Após a fundamentação teórica explana-se o poder judiciário, que engloba a temática do projeto, explicando a estrutura de uma dos poderes federais, bem como a metodologia da pesquisa.

A pesquisa realizada analisa 84 processos jurídicos que correspondem a casos que envolvem o comércio eletrônico registrados no estado do Paraná. Todos os documentos se encontram em mídia digital e podem ser encontrados no portal do Tribunal da Justiça do Paraná. Conclui-se o projeto com a análise da mineração de texto utilizando o software PreText. A terceira seção detalha os procedimentos metodológicos do presente estudo, bem como os resultados obtidos e a conclusão da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

A seguir, realiza-se um breve levantamento dos referencias teóricos que embasam a pesquisa.

2.1 DADO, INFORMAÇÃO, CONHECIMENTO E TECNOLOGIA

Na era da informação e da tecnologia da informação, segundo Carvalho e Tavares (2001) a produção de dados, informações e a troca de conhecimentos são atividades básicas para a sociedade. Em uma análise cronológica, nota-se que tecnologias para a comunicação, essa compreendida como um processo intermediário que permite a troca de informação entre dois elementos (COADIC, 1996), são desenvolvidas desde o surgimento da humanidade.

Uma pessoa em busca de determinada informação utiliza si*stem*as de pesquisa para atingir um ou mais de seus objetivos (BAEZA-YATES; RIBEIRO-NETO, 1999).

Os primeiros membros do gênero homo a habitarem o planeta, por exemplo, já buscavam formas para a coleta, armazenamento e transmissão da informação para a sobrevivência. Naquele momento surgia a informação como insumo estratégico, utilizada em grande escala, mesmo que por vezes de modo tácito, por grande parte das organizações atualmente.

As capacidades físicas e mentais humanas se desenvolveram, e juntamente com elas, milhares de ferramentas que auxiliam o ser em suas tarefas. Atualmente a humanidade se encontra em um estágio em que as tecnologias aprimoram a aquisição de conhecimento e informação, ao passo que o conhecimento e a informação aprimoram as tecnologias.

A inegável contribuição das tecnologias de informação para a criação de uma complexa rede de dados e informações foi de suma importância para que se atingisse o nível de sociedade globalmente integrada que se observa na atualidade. Ainda sim, os principais fatores para o desenvolvimento da humanidade não são as tecnologias, mas sim, as capacidades cognitivas do ser humano, traduzidos em dados e informações, combinados ao conhecimento. Orientados a compreender

melhor o impacto desses fatores, teóricos definiram os termos que englobam a comunicação.

A menor parte da informação, o dado, é compreendido como uma sequência de símbolos quantificados ou quantificáveis (SETZER, 1999). O tipo do dado varia conforme sua aplicação, portanto, é uma unidade de valor volátil.

De acordo com Setzer (1999), a informação não pode ser medida ou valorada por meios matemáticos. Ainda, a informação está presente em diferentes plataformas, podemos encontrar informação em linguagens corporais, imagens, ações, músicas, conjunto de dados ou sequência de palavras, entre outros. Isso, pois a definição da informação é a de uma forma abstrata que representa valor ou significância ao receptor em um processo de comunicação (SETZER, 1999).

A carga informacional (informação contida no meio percebida pelo ser) é subjetiva conforme o receptor, ou seja, o que pode ter valor para alguém, pode significar nada para outra pessoa. A forma abstrata pode ou não obter carga informacional.

Exemplificando o caso, pode-se analisar o discurso de um ditador chinês. Para os chineses, as palavras proferidas pelo ditador possuem alto grau de carga informacional. Já para um brasileiro, sem conhecimento do idioma, não se faz possível à interpretação das palavras do ditador, portanto, para o brasileiro, as mesmas não possuem valor, e não configuram uma informação.

O valor da informação é determinado exclusivamente pelo usuário ou receptor (MCGEE; PRUSAK, 1990).

Nesse sentido, a informação depende de uma vasta quantidade de variáveis presentes entre o receptor e emissor. A essa configuração, nomeia-se o processo de comunicação.

O processo de comunicação pode ser entendido como um determinado meio ambiente em que o emissor transmite sua mensagem ao receptor, e como o receptor reage à transmissão conforme as características do meio, do emissor e as de si próprio (CRUZ; SEGATTO, 2009).

Tendo a definição, pode-se perceber que a informação é extremamente subjetiva, pois cada indivíduo possui seus valores, crenças, costumes e pontos de vista, que irão valorar a informação não somente pelo seu conteúdo, mas pelo meio ao qual ela está inserida.

Muito se discute sobre o real papel do conhecimento e informação nos dias de hoje. Alguns consideram que o conhecimento deixou de ser uma ferramenta externa, e passou a integrar a economia clássica. E ainda, que o conhecimento não é mais apenas um componente - chave nas empresas, mas sim o insumo mais importante para o crescimento econômico (BEIJERSE, 1999).

Beijerse (2009) afirma que o conhecimento é algo além da ponta do iceberg, ou seja, algo além da informação comum. Na ponta do iceberg temos o conhecimento explícito, que é aquele que externalizado, de acesso facilitado. Já, de acordo com Nonaka e Takeuchi (1997), o resto do iceberg é composto por conhecimento tácito, interno, difícil de ser propagado e formalizado.

O conhecimento tácito compõe a maior parte do iceberg, por justamente estar presente em todas as pessoas e organizações, de diferentes formas e meios. O conhecimento pode ser compreendido como um conjunto de dados e informações que juntas acrescem valor e formam um determinado conceito (NONAKA;TAKEUCHI,1997).

Para manusear esse importante ativo, desenvolve-se através dos anos a gestão do conhecimento, essa ciência pode ser definida como uma ferramenta administrativa que aumenta o valor agregado das informações, contextualizando-as, gerindo o conhecimento explícito e tácito (SILVA, 2011).

2.2 RECUPERAÇÃO DA INFORMAÇÃO

Sendo uma subárea da ciência da computação, a Recuperação da Informação (RI) surgiu como estudo a partir da identificação da necessidade de processos de recuperação cada vez mais sofisticados, dado o crescente número de dados e informações disponíveis aos usuários (CARDOSO, 2005).

A RI está presente toda vez em que se faz necessário o uso de determinada informação (normalmente contida em documentos) para realizar uma atividade ou ação, consiste no ato de encontrar um material que satisfaça determinada necessidade informacional (MANNING *et al*, 2009).

Segundo Silva (2005), o modelo básico da RI é composto por uma base de dados, uma interface, e um usuário. Se um usuário deseja consultar um conteúdo presente na base de dados, é por meio da interface que o mesmo a realizará. A

partir desse momento, as tecnologias empregadas no Sistema de Recuperação de Informação realizam a busca na base de dados, retornando o requisitado ao usuário.

A Recuperação de Informação é usada como sinônimo de recuperação de documentos e de texto (JONES; WILLETT, 1997), e pode ser considerada como a porção tecnológica da Ciência da Informação (SARACEVIC, 1999).

Essa ciência trata dos aspectos intelectuais da informação, bem como suas especificações para a busca (recuperação), além de abranger todos os si*stem*as, técnicas e máquinas que compõe a operação de recuperar uma informação (MOOERS, 1951).

Michael Lesk (1995) propõe em seu artigo, *The Seven Ages of Information Retrieval*, uma evolução cronológica da recuperação da informação. Para isso, o autor divide a RI em diferentes momentos, a seguir uma descrição adaptada das etapas da RI descritas por Lesk:

- **Infância (*Childhood*) 1945 – 1955:** em um contexto de início de Guerra Fria, com a corrida armamentista e tecnológica entre a URSS e os Estados Unidos. Teve-se a criação dos primeiros si*stem*as de recuperação, que englobavam técnicas já conhecidas, como uso de índices e concordâncias.
- **O estudante (*The schoolboy*) década de 60**: si*stem*as de informação foram criados em larga escala. Embora muitos trabalhos de RI fossem feitos de modo tradicional (uso de índices em documentos físicos) é nessa década que ocorre o Boom da RI, a comunidade científica passa a dar mais atenção a conferências e produções intelectuais acerca do tema. Grandes órgãos, como a NASA, passam a dar mais atenção a melhorar o processo de recuperação da informação utilizando computadores.
- **Maioridade (*Adulthood*) década de 70**: a recuperação da informação começa a se fixar nos si*stem*as. Um dos causadores são os novos formatos tipográficos que permitem, efetivamente, com que se possam obter documentos textuais e passíveis de leitura em computadores. Outro motivo foi à evolução dos processos tecnológicos na ação de se recuperar a informação. Agora os computadores podem processar a informação de forma muito mais rápida.
- **Maturidade (*Maturity*) 1980 – 1990**: houve um enorme crescimento de base de dados disponíveis em si*stem*as online. Técnicas tradicionais foram adaptadas a web. Os avanços tecnológicos permitiram com que a

informação fosse indexada e recuperada de forma mais rápida, mais ampla, mais segura e confiável. Novas mídias de armazenamento como o CD-ROM, permitem com que se guardem muito mais informações, de diferentes tipos e tamanhos. O Boom informacional e tecnológico incita a comunidade científica a realizar cada vez mais pesquisas na área.

- **Realização (*Fulfillment*) 2000**: consolidada, junto à indexação e a web, como ferramenta indispensável a toda a informação produzida diariamente, a RI vislumbra novos desafios. Como a utilização de técnicas mais apuradas para recuperar informações em diferentes mídias e plataformas. Um vídeo, por exemplo, necessita de um sistema de recuperação de informação extremamente amplo e apurado, pois é uma mídia que carrega diversos canais informacionais (imagem, som, texto, etc).

Atualmente, vive-se em uma era onde o montante informacional cresce cada vez mais. E cada vez mais se tem espaço para armazenar dados e informação, portanto, quanto mais documentos indexados existirem, melhor terá de ser o sistema de RI, para que o usuário possa recuperar de forma precisa a informação desejada.

2.3 TIPOS DE RECUPERAÇÃO DA INFORMAÇÃO

Segundo Hiemstra (2001) e Silva (2005), Existem três processos básicos que a recuperação da informação deve dar suporte: a representação do conteúdo do documento; a representação das necessidades informacionais do usuário; e a combinação (*matching*) das duas representações. Esses processos podem ser visualizados na figura 2. Os retângulos representam dados e os retângulos de bordas arredondadas representam processos:

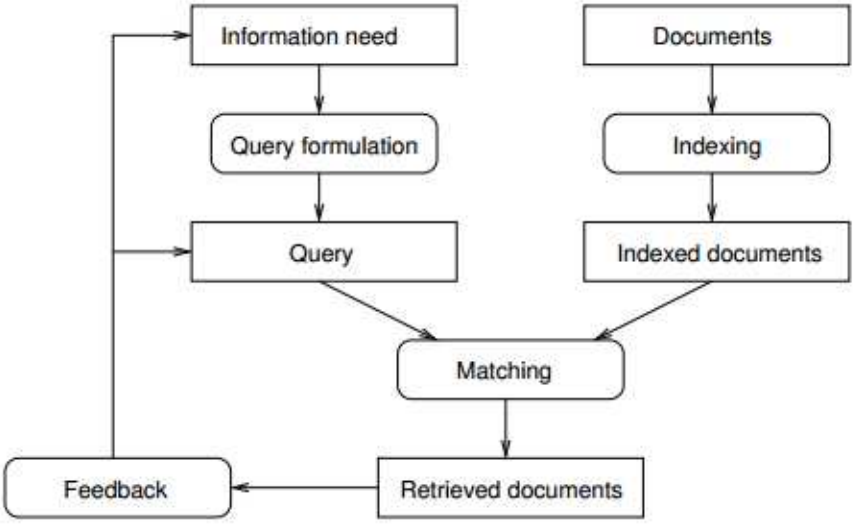

FIGURA 2 - FLUXOGRAMA DOS TRÊS PROCESSOS BÁSICOS DA RECUPERAÇÃO DA INFORMAÇÃO
FONTE: HIEMSTRA (2001)

No modelo da figura 2, pode-se observar duas rotinas principais. A primeira refere-se às etapas em que o usuário participa diretamente, normalmente a essa etapa se da o nome de nível de *view*, de modo que o usuário interaja com o sistema por meio de uma interface.

Assim que imposta a necessidade informacional, o sistema busca, por meio de uma linguagem de consulta, os documentos indexados. A partir de então, o terceiro processo, o de combinação, compara a consulta realizada pelo usuário, com os documentos indexados na base que possuam conteúdo relativo à consulta, resultando nos documentos que satisfaçam a necessidade informacional do usuário.

2.3.1 Modelos Quantitativos

Os tipos de recuperação da informação baseados em modelos quantitativos são representados por conjuntos de termos de indexação, baseados em termos lógicos e estatísticos (FERNEDA, 2003). Os modelos quantitativos, em sua maioria,

buscam suporte na lógica, estatística e outras fórmulas matemáticas que justificam seus tipos de recuperação da informação.

2.3.1.1 Modelo Booleano

O modelo booleano é representado por 0 e 1, cada qual representando verdadeiro ou falso. Esses termos são conectados por operadores lógicos, E, OU e NÃO. E resultam em uma linguagem que satisfaça as restrições lógicas (FERNEDA, 2003). Sendo assim, um modelo fundamentado na teoria dos conjuntos, que expressa seus resultados de recuperação da informação como uma combinação de índices (SILVA, 2002). A seguir, os operadores lógicos e suas funcionalidades segundo Ferneda (2003).

- **Operador lógico "E"**: recupera documentos indexados que satisfaçam a condição Termo 1 (T1) E Termo 2 (T2). Representada pela intersecção dos conjuntos

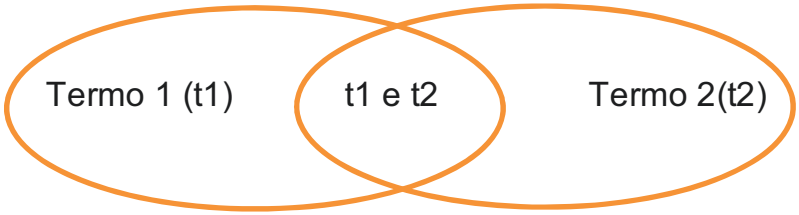

FIGURA 3 - RECUPERAÇÃO DA INFORMAÇÃO – MODELO BOOLEANO, OPERADOR LÓGICO E
FONTE: Adaptado de FERNEADA (2003)

- **Operador Lógico "OU"**: recupera documentos indexados que satisfaçam a condição T1 OU T2. Representada pela união dos conjuntos.

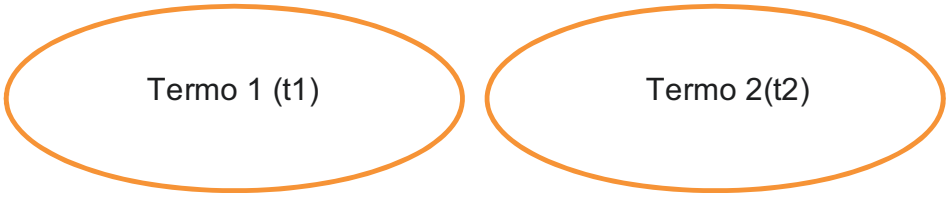

FIGURA 4 - RECUPERAÇÃO DA INFORMAÇÃO: MODELO BOOLEANO, OPERADOR LÓGICO OU
FONTE: Adaptado de FERNEADA (2003)

- **Operação lógica "Não"**: a operação NÃO t1 ignora documentos que são recuperados através do termo t1.

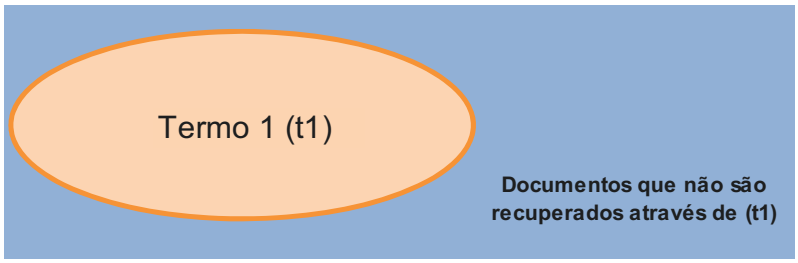

FIGURA 5 - FIGURA 5 – RECUPERAÇÃO DA INFORMAÇÃO: MODELO , OPERADOR LÓGICO NÃO
FONTE: Adaptado de FERNEADA (2003)

Outra operação possível para esse operador lógico é caso de uso t1 E NÃO t2, ou seja, o sistema deve recuperar documentos com o termo um, que não contenha o termo dois.

FIGURA 6 - RECUPERAÇÃO DA INFORMAÇÃO: MODELO BOOLEANO,OPERADOR LÓGICO NÃO, CASO T1 E NÃO T2
FONTE: Adaptado de FERNEADA (2003)

Dado os modelos básicos da lógica booleana, pode-se combiná-los a fim de refinar a pesquisa. Expressando, através da combinação de índices e operadores lógicos, a necessidade informacional do usuário (SILVA, 2002).

2.3.1.2 Modelo Fuzzy

A lógica fuzzy objetiva trabalhar com a diversidade, incertezas e verdades imparciais dos fenômenos da natureza de uma forma si*stem*ática (SHAW; SIMÕES, 1999). O modelo fuzzy não é prestigiado pela comunidade de recuperação da informação, os estudos realizados com esse modelo consideram pequenas

amostras, que não comprovam sua superioridade perante outros modelos (BAEZA–YATES; RIBEIRO-NETO, 1998).

A representação fuzzy de um documento é definida pela função F(d, t), o valor numérico produzido representa o peso do termo t para o documento d. Essa função baseia-se no cálculo da frequência de ocorrência dos termos no texto, fornecendo uma representação estatística do texto (FERNEDA, 2003).

2.3.1.3 Modelo de Espaço Vetorial

O modelo de espaço vetorial é representado por um vetor construído por termos (SINGHAL, 2001). Nesse modelo é possível obter documentos que correspondam à parte da consulta realizada, através da associação do peso dos termos, e também do peso das expressões de busca (SILVA, 2002).

Qualquer texto pode ser representado por um vetor na dimensão espaço vetorial. Se um termo estiver presente em um texto, o mesmo recebe um valor maior que zero no vetor-texto, ao longo da dimensão correspondente ao termo (SINGHAL, 2001).

Normalmente, o ângulo entre dois vetores é usado para medir a divergência entre os vetores, o cosseno desse ângulo é usado para representar a similaridade numérica entre os vetores representados pelos termos. Se todos os vetores são unidades de comprimento, então o cosseno do ângulo entre os dois vetores é igual aos seus produtos escalares (SINGHAL, 2001).

$$Sim(\vec{D}, \vec{Q}) = \sum_{t_i \in Q,D} w_{t_i Q} \cdot w_{t_i D}$$

FIGURA 7 - EQUAÇÃO DO MODELO ESPAÇO VETORIAL
FONTE: SINGHAL (2001)

A equação é representada na figura 7, onde D é o vetor documento, e Q é o vetor de consulta, a similaridade do documento D em relação à consulta Q pode ser representado de modo que wtiq é o valor do enésimo componente (i) no vetor consulta (Q), e wtid é o enésimo componente (i) no vetor documento (D) (desde que wtiq e wtid sejam maiores do que zero).

2.3.1.4 Modelo Probabilístico

A teoria da probabilidade é uma ferramenta na resolução de problemas matemáticos. O método probabilístico tenta provar que se uma estrutura com certas propriedades existe, pode-se definir uma probabilidade dentro do espaço amostral, provando padrões que ocorrem no conjunto (ALON; SPENCER, 2000).

O modelo probabilístico é definido por P(e) = n(E)/n(S), onde a probabilidade E de um espaço amostral é definida pela razão entre o número de elementos (E), e o número total de elementos (S) (SILVA, 2002).

O modelo pode ser definido entre eventos dependentes e independentes. Quando os eventos são independentes entre si, o modelo probabilístico atua na ordem de P(E) = E(1).E(2), onde a probabilidade de determinado espaço amostral é definido pela multiplicação do número de possibilidades do evento um (E1), pelo número de possibilidades do evento dois (E2).

Já quando os eventos são dependentes, pode-se definir a probabilidade como sendo P(A | B) = P(A e B)/P(B), onde a probabilidade de A, dado B, é definida pela multiplicação da probabilidade de A e B, fracionado pela probabilidade de B.

A recuperação probabilística ocorre, segundo Ferneda (2003), com a divisão de quatro subconjuntos dentro do conjunto do tema: os documentos recuperados (Rec), o conjunto dos documentos relevantes que foram recuperados (RR) e o conjunto dos documentos não relevantes e não recuperados (rr).

A equação da recuperação probabilística é definida tendo um conjunto relevante (REL), e um conjunto de documentos não relevantes (REL), sendo similaridade (sim) de um documento “d”: sim(d, e) + p(REL | d)/p(REL | d).

2.3.2 Modelos Dinâmicos

Os modelos dinâmicos colocam o usuário como fator de importância a recuperação da informação. De forma que os usuários interajam e influenciem na representação dos documentos, moldando a relevância dos documentos recuperados (FERNEDA, 2003).

2.3.2.1 Sistema Especialista

Segundo Ferneda (2003), o sistema especialista atua em um campo específico do conhecimento, auxiliando as ações e fatores que englobam o domínio, guiado pela priori de que quanto mais informações possam ser armazenadas, melhor o sistema será.

Esse sistema está orientado a um grande armazenamento de informação.

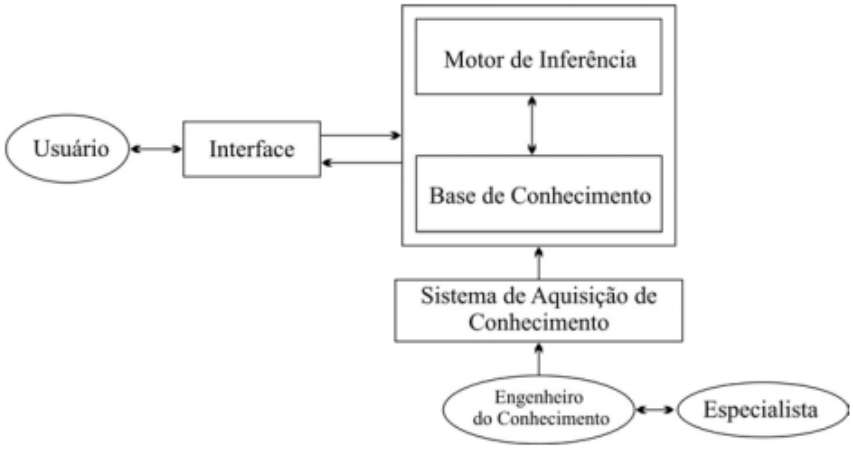

FIGURA 8 - SI*STEM*A ESPECIALISTA
FONTE: FERNEDA (2003)

A figura 8 ilustra os processos do sistema especialista, o sistema é composto pelos métodos de recuperação (motor de inferência), conectado a base de conhecimento. O sistema possui interação com o usuário por meio de uma interface, por esse intermédio o usuário pode recuperar documentos. Além disso, da outra extremidade, o sistema é alimentado por um sistema de aquisição de conhecimento, que por sua vez é alimentado pelos especialistas do tema específico do si*stem*a.

A aplicação do sistema especialista na recuperação da informação, segundo Ferneda (2003), é formada, normalmente, por cálculos que associam palavras, buscando qualificar os termos na sua recuperação. As regras de associação tentam transformar as palavras-chave que classificam um documento, em uma informação, melhorando o processo de recuperação da informação.

2.3.2.2 Redes Neurais

Uma rede neural é um processador que atua de forma paralela, orientado por uma unidade de processamento simples. Deve possuir a capacidade de armazenar informações experimentais e torná-las disponíveis para o uso (HAYKIN, 1999).

Segundo Haykin (1999), os modelos de redes neurais procuram se assemelhar ao processamento do cérebro humano, particularmente em dois aspectos. O primeiro é referente ao fato de que as informações adquiridas pela rede são advindas de seu ambiente (semelhante ao sistema especialista), por meio do processo de aprendizagem. A outra semelhança reside no fato de que o método de armazenagem de informações se assemelha aos pesos sinápticos, nomenclatura dada às forças de conexão entre os neurônios.

De modo simplista, a ação de recuperação da informação pode ser visto como um processo baseado em redes neurais de três etapas (FERNEDA, 2003), semelhantes às definidas por Hiemstra (2001) e Silva (2005) como processos básicos da recuperação da informação.

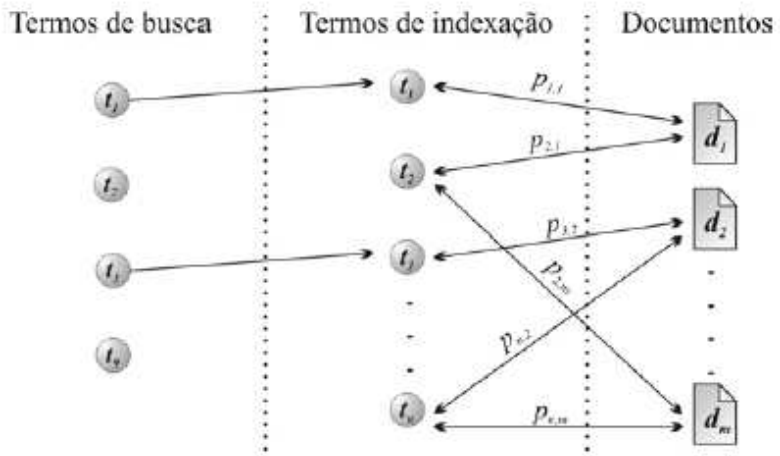

FIGURA 9 - RECUPERAÇÃO DA INFORMAÇÃO: REDES NEURAIS
FONTE: FERNEDA (2003)

A figura 9 ilustra o modelo de Rede Neural aplicado a Recuperação da Informação, na figura se vê que os termos inseridos, que pertencem ao conjunto indexado na base de documentos, são ativados simultaneamente para achar correlação com os termos de indexação. Logo os termos de indexação entram em contato com os documentos. Assim que ocorrido o matching, o sistema retorna os documentos que satisfaçam a necessidade informacional do usuário.

2.3.2.3 Algoritmos Genéticos

Algoritmos Genéticos utilizam da probabilidade para fornecer um mecanismo de busca baseada no princípio de sobrevivência dos mais adaptados, inspirados no princípio Darwiniano da evolução das espécies (PACHECO, 1999).

O modelo utiliza-se de um processo repetitivo que mantém um determinado montante de dados (população) que possam representar possíveis soluções a um determinado problema, a cada repetição do problema (geração), os dados e informações passam por uma avaliação para saber se são capazes de solucionar o problema (cálculo fitness) (FERNEDA, 2003).

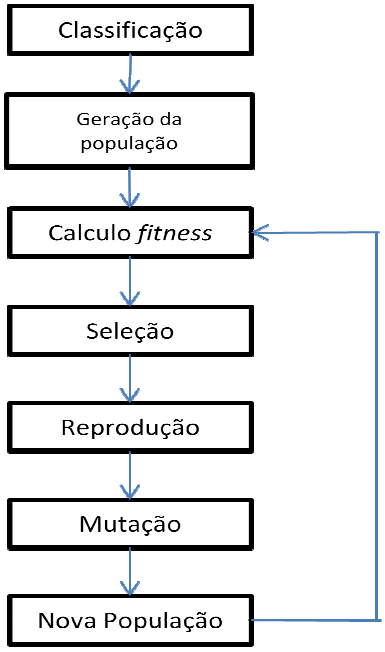

FIGURA 10 - ALGORITMOS GENÉRICOS
FONTE: Adaptado de FERNEDA (2003)

Na figura 10 se nota o processo do algoritmo genético, primeiramente os dados e informações são codificados (codificação dos indivíduos), após essa etapa as informações e dados são indexados (geração de População Inicial). Então se é executada a primeira rotina *fitness*, procurando testar a eficiência e eficácia dos elementos indexados na resolução de um determinado problema. O processo de seleção define quais são os elementos que estão aptos a tratar um problema. O processo de mutação modifica o elemento a fim de potencializar sua utilidade. Após

essas etapas, ocorre a atualização dos elementos retornando o sistema ao seu processo inicial.

Na aplicação dos algoritmos genéticos na recuperação da informação, os termos de pesquisa são agrupados em “genes”, cada termo é representado por 0 ou 1, (possui correlação com o documento ou não possui). Então é realizada a etapa de fitness para testar os genes. Logo, passa-se a etapa seguinte, que é a mutação dos genes a procura do conjunto de termos que possam prover uma recuperação mais assertiva da informação. Assim que fora provado a melhor combinação de termos, o documento é recuperado.

2.4 NUVEM DE *TAG*

A nuvem de *tag* pode ser definida como uma lista de termos, normalmente apresentada em ordem alfabética, e diferenciadas pelo tamanho da fonte de cada termo, quanto maior o termo maior a relevância do mesmo dentro de uma estrutura textual (TAGCROWD, 2012). Abaixo um exemplo de nuvem de tag, retirada do blog do site tagcrowd.com.

FIGURA 11 - NUVEM DE TAG
FONTE: TAGCROWD.COM (2012)

A nuvem de *tag* pode ser uma importante ferramenta de análise qualitativa, os termos destacados podem esclarecer cenários que antes não eram possíveis de

serem esclarecidos. A desvantagem da nuvem de tag, justamente por depender de uma análise subjetiva, é a de que dificilmente haverá 100% de certeza quanto ao que for constatado.

2.5 MINERAÇÃO DE DADOS e KDD

A Mineração de Dados transforma uma grande quantidade de dados em conhecimento, e pode ser vista como o resultado natural da evolução da tecnologia da informação (HAN et al., 2007). Evolução essa que acompanha a dita era da informação, onde o conhecimento adentra em uma íngreme ascendente no que tange sua produção (CARVALHO, TAVARES, 2001).

A Mineração de Dados é um conceito desenvolvido a partir de um conjunto de técnicas, conceitos e ferramentas, como a estatística, banco de dados, armazenamento de dados, computação de alto desempenho, algoritmos e probabilidades (HAN et al.,2007). É uma das etapas do método de Knowleged Discovery in Databases (KDD), que é abordado no decorrer do trabalho.

Dentro do KDD, a mineração de texto tem a função de análise dos dados, descobrindo, através de algoritmos, um determinado número de padrões dentro do escopo selecionado (FAYYAD; SHAPYRO; SMITH, 1996).

O KDD preocupa-se com o desenvolvimento de métodos e técnicas para dar sentido aos dados, sendo um processo interativo e interativo que envolve inúmeros processos com múltiplas decisões que podem ser tomadas (FAYYAD; SHAPYRO; SMITH, 1996).

Tendo que a mineração de dados tem como função extrair de uma determinada base de dados algo além de suas formas quantificáveis, Rezende et al (2011), descreve os principais métodos de mineração existentes, são eles:

- **classificação**: neste tipo de tarefa, ocorre o mapeamento dos dados de entrada, em um determinado número de categorias. O mapeamento resulta em classes, formada por exemplos que consi*stem* em um conjunto de atributos, e um atributo-meta discreto. Esse tipo de algoritmo tem como objetivo encontrar algum relacionamento entre os atributos e uma classe.
- **regressão**: a diferença entre o método de regressão para o método de classificação, é que o atributo-meta é contínuo, e não discreto. O objetivo dessa tarefa é encontrar uma relação entre um atributo e um atributo-meta

contínuo. Nesse modelo pode-se encontrar diversas formas de representação do conhecimento, Regras de associação e Árvores de Decisão são exemplos.

- **regras de associação**: o objetivo das regras de associação é encontrar tendências que possam ser usadas como descoberta de padrões dado a base de dados. São muito utilizadas como insumos estratégicos para descobrir comportamentos de classes de consumidores por exemplo. Utilizam também dos operadores lógicos SE, ENTÃO, OU e E.

Observada as afirmativas, é fato que o KDD utiliza diversas áreas do conhecimento para formar suas características. É também um processo não rotineiro que identifica, valida, potencializa o uso e promove significância aos dados. Tem como uma de suas premissas o processamento de dados de baixo nível dentro de outras formas que possam ser mais compactas, abstratas e úteis (FAYYAD; SHAPYRO; SMITH, 1996).

Antes do KDD, o principal meio para transformar dados em conhecimento era através de análises manuais e interpretativas. Porém, na medida em que a informação e a produção de dados tornaram-se mais complexas e densas, os métodos antigos tornaram-se impraticáveis. O KDD, na condição de processo, possui etapas distintas (FAYYAD; SHAPYRO; SMITH; 1996).

A Figura 12 ilustra as principais etapas do KDD.

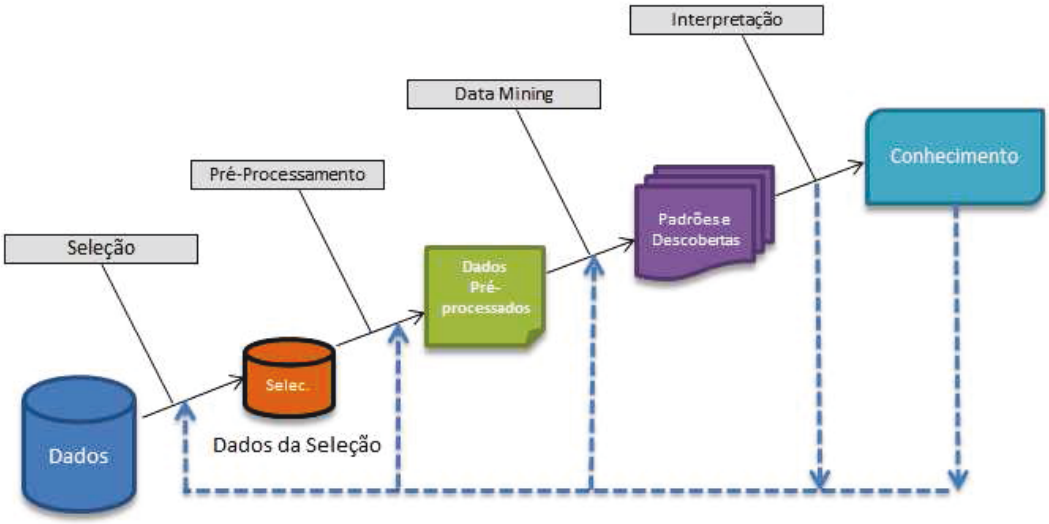

FIGURA 12 - KDD
FONTE: Adaptado de FAYYAD; SHAPIRO; SMYTH (1996)

As etapas do KDD como representado na figura 12 se iniciam a partir de um conjunto de dados. Neste momento deve-se compreender o ambiente de onde os dados existem ou foram extraídos, bem como o propósito da mineração (FAYYAD, SHAPIRO, SMYTH, 1996).

O segundo passo consiste na seleção dos dados com o propósito de criar um conjunto de dados alvo, onde as descobertas serão realizadas. O terceiro passo consiste na limpeza e pré-processamento dos dados. São realizadas operações que moldem o dado de forma orientada ao objetivo.

Na quarta etapa observa-se a redução e projeção, que consiste em encontrar características úteis para a representação do dado orientado ao objetivo. Com isso o número de variáveis não utilizáveis se reduz, encontrando possíveis inconsistências na base.

O quinto passo compete à aplicação da heurística de mineração de dados selecionada, nota-se que esse passo deve ser executado alinhado ao propósito da mineração da base, pois a escolha errada do método pode resultar em padrões irrelevantes.

O sexto passo consiste no reconhecimento dos padrões obtidos através da mineração de dados. Os padrões são o produto do processo, já a interpretação e validação dos mesmos são as ações que agregam valor aos resultados, provendo então, o conhecimento aprofundado da amostra escolhida.

A mineração de dados e o descobrimento de conhecimento vêm sendo usada em diversas áreas. Fayyad, Shapiro e Smith (1996) elucidam algumas aplicações nas principais áreas em que a mineração de dados é aplicada, são elas:

- **marketing:** a principal aplicação da mineração de dados para o marketing é para os sistemas de informação, que analisam dados de consumidores para identificar grupos e prever os seus comportamentos;
- **investimentos**: organizações utilizam técnicas de mineração de dados, na maioria dos casos coletando dados de mercado fora de seus si*stem*as de apoio à decisão;
- **detecção de fraude bancária**: bancos utilizam outliers para identificar possíveis transações financeiras que podem indicar crimes bancários ou operações ilegais;
- **indústria**: diferentes heurísticas são utilizadas para identificar e prever erros de fabricação nos produtos;

- **filtro de dados**: a mineração de dados vem sendo largamente utilizada para agregar qualidade a recuperação de informação. Inserindo-se uma determinada quantidade de termo, por meio de algoritmos de busca, podem-se recuperar informações mais precisas e que, consequentemente, atendem as necessidades informacionais do usuário.

2.6 MINERAÇÃO DE TEXTOS

Com a forma mais natural de armazenamento de informação (texto), credita-se a mineração de texto como campo científico de poder comercial maior do que a mineração de dados. Ainda, ressalta-se que a mineração de texto é mais complexa que a mineração de dados, pois trata de dados não estruturados (TAN, 1999).

Esse tipo de mineração trata o problema do volume crescente de textos digitalmente armazenados, procurando identificar novas informações e conhecimentos até então desconhecidos, através da extração automática de fontes de texto diferentes (GUPTA; LEHAL, 2009).

A mineração de textos se refere ao processo de extração de conhecimento ou padrões não estruturados em documentos de texto. É semelhante à mineração de dados, exceto pelo fato de que as ferramentas de mineração de dados são desenhadas para lidar com dados estruturados dentro de uma base, enquanto que na mineração de texto se deve lidar com dados não estruturados ou semiestruturados (GUPTA; LEHAL, 2009).

Existem alguns elementos chaves que compõe um processo de mineração de texto, a seguir a leitura desses elementos, adaptados de Feldman e Sanger (2006).

O primeiro compete à base de dados e documentos, que pode ser definida por basicamente qualquer coleção de textos, artigos e outras mídias que sejam compostas por palavras.

As coleções dos documentos podem ser tanto estáticas, caso em que não há alteração na base durante o processo de mineração, ou dinâmicas, caso em que pode haver a implementação de novos documentos ou palavras-chave durante o processo de mineração de texto. Nesse elemento, Feldman e Sanger explanam algumas subcategorias:

- **documento:** uma base de documentos textuais, mesmo que pequena, pode possuir muitas características diferentes entre os documentos que a

compõe. Um dos objetivos do pré-processamento é tentar simplificar ou modelar os documentos não-estruturados afim de refinar a base para que haja relevância nos resultados que poderão ser obtidos. Por outro lado, são as características de uma base textual que fará possível a mineração de texto, portanto, é necessário encontrar um equilíbrio entre a modelagem da base textual, para que não haja depreciação das características dos documentos da base;

- **características dos documentos comumente utilizadas:** para representar o documento dentro de um processo de mineração de texto, são comumente utilizadas palavras, letras, termos e conceitos. A escolha do objeto que orientará a mineração de texto dependerá da base de texto que o usuário possuirá. Ainda, observa-se que quanto mais próximo à mineração baseada em conceito, mais difícil será o processo, mas, provavelmente, embora em quantidade reduzidas, mais satisfatórios serão os resultados;
- **domínios do conhecimento:** os domínios do documento podem ser usados para limitar os elementos relevantes dentro de uma base textual. Na etapa de pré-processamento, esse elemento é um importante adjunto de classificação e metodologia para extração de conceitos, palavras, caracteres ou letras.

Os próximos elementos, segundo os mesmos autores, referem-se a padrões e tendências. A função principal da mineração de texto é analise da ocorrência de padrões dentre documentos de uma base.

Na figura 13 se observa o modelo de mineração de textos de Grupta e Lehal, semelhante às etapas do KDD, defendido também por Almeida (2003), exceto pelo fato de que o autor aprofunda as fases, e semelhante também ao modelo de Feldeman e Senger.

FIGURA 13 - PROCESSO DE MINERAÇÃO DE TEXTO
FONTE: Adaptado de GRUPTA e LEHAL (2009)

Na primeira etapa se seleciona um conjunto de documentos, uma ferramenta de mineração de texto deve recuperar um documento particular e o pré-processa, analisando o formato e os conjuntos de caracteres.

O pré-processamento é útil para identificar grupos similares em um conjunto de texto. Ainda, devem-se construir representações desse conjunto, para tornar o processamento da mineração mais assertivo (ALMEIDA, 2003). Para tal etapa, Existem diferentes técnicas que permitem com que os documentos estejam preparados para a mineração de texto, como definidos por Feldman e Sanger (2006), Lopes (2004) e Soares, Prati e Monard (2008):

- ***Bag of Words:*** para transformar os textos em dados estruturados se realiza a contagem das palavras independente de seu contexto ou significado;
- ***Token*ização** (*Tokenization*): a *token*ização corresponde à quebra do texto em sentenças e palavras. O principal desafio é realizar as separações respeitando as ordens gramaticais da língua nativa dos textos da base de dados;
- ***Stemming:*** o processo de *stem*ming procura reduzir os *token*s agrupando as flexões das palavras por meio de seu sufixo. As palavras: acompanha, acompanhado, acompanhante seriam reduzidas ao *stem* acompan, por exemplo. Esse processo é extremamente eficiente para a língua inglesa, por conta da simplicidade das regras gramaticais da língua, já para línguas como espanhol ou português, o processo pode não ser muito preciso. Um exemplo pode ser observado com as palavras barata e barato, ambas possuem dois valores semânticos diferentes, mas poderiam ser reduzidas ao *stem "*barat";
- ***Stop-words*:** segundo Lopes (2004), *stop-words* são palavras que não agregam valor semântico dentro do contexto da base de dados textual;
- **Seleção de partes do texto** (*Part-of-speech tagging)*: a técnica prevê a seleção e divisão de palavras em categorias, baseadas no papel que as mesmas atuam em cada sentença. A seleção de partes do texto (SPT) permite com que se observe o papel semântico da palavra dentro do texto. As categorias são comumente divididas em artigos, nomes, verbos, pronomes, adjetivos etc;

- **Tesauro:** o Tesauro é um vocabulário controlado organizado em estrutura lógica de modo que as várias relações entre os termos são visualizadas indicadores padronizados (ANSI/NISO Z39.19-2005 [R2005]);
- ***Case Folding*:** o *case folding* tem a função de modelar as palavras, como, por exemplo, deixando-as todas minúsculas. Essa rotina agiliza todo o processo da mineração de texto (LOPES, 2004);
- ***Parsing***: é responsável por separar o texto em sua forma livre em palavras, realizar a limpeza da base removendo palavras redundantes e sem conteúdo (FRANCIS, 2005);
- ***N-Grama:*** o n-grama (onde n é um número qualquer) calcula a ocorrência de palavras em sequência, que podem ter maior significância dentro de uma base do que simplesmente as frases isoladas.

2.6.1 Processo de descoberta de padrões em dados não estruturados

No pré-processamento, deve-se segmentar o texto. Em outras palavras, deve-se preparar a estrutura textual retirando espaçamentos e sinais de pontuação. Após o ajuste na base textual, é prudente realizar a exclusão de termos, como preposições e conjunções. Pois os mesmos não possuem valor informacional (ALMEIDA, 2003), além de reduzir verbos a sua raiz, excluir palavras que não são relevantes entre outros procedimentos já citados.

2.6.1.1 Seleção de Palavras

O primeiro passo para encontrar termos que possam se transformar em valores-atributos e assim agregar relevância ao texto, para a língua portuguesa, é utilizar medidas de frequência. Dada à complexidade da língua, se torna difícil em um primeiro momento a análise semântica de cada termo, ainda sim, as medições utilizadas podem fornecer interpretações interessantes para as bases em português.

No presente trabalho se utiliza o conceito do N-grama, onde N é o número de junções de palavras a ser realizadas, o n-grama varre o texto a procura de palavras em sequência, que podem ter mais relevância do que palavras isoladas, um exemplo é o termo “Dano Moral”, para n =1, ou seja, 1-grama, o resultado seria a

separação da palavra DANO da palavra MORAL, quando que, para o contexto do trabalho, utilizando-se 2-grama podemos utilizar o termo “Dano Moral”, que possui mais relevância.

2.6.1.2 Cortes de Palavras Baseado na Frequência

A mineração de texto pode ser analisada simplesmente pela frequência de termos, porém, como um termo pode-se ser muito mais frequente que outro, e mesmo assim o termo de menor frequência possuir maior valor informacional é necessário utilizar algum tipo de ponderação de termos.

O modelo mais utilizado de ponderação de texto é o TFIDF (*Term Frequency X Inverse Document Frequency*) proposto por Jones (1972). Este modelo, segundo Almeida (2003), combina a frequência de ocorrência de um termo com a frequência inversa dos documentos em que o termo ocorre.

Exemplificando, em uma base de documentos de 200 arquivos, supõe-se que o termo (t) “árvore” apareça duas vezes em um documento de 200 palavras (p), pode-se afirmar que a frequência do termo é definida por 2. Ao se analisar a base em sua totalidade, observa-se que o termo “’arvore” aparece em 100 (a) dos 200 documentos da base (b), pode-se então aferir que o idf do documento é (Log 100/200). Assim o peso do termo, segundo o cálculo td-idf é definido por (2*log 200/100) ou td-idf = ((t/p)*(log b/a)).

O resultado expressado em idf é a capacidade que o termo possui para discriminar uma estrutura textual. Após a ponderação do termo, pode-se aplicar outras medidas para qualificar a mineração.

Como o caso da frequência do documento, que define em quantos documentos um termo apareceu; a força do termo, que indica a probabilidade de ocorrência de um termo em um documento; a contribuição do termo, que indica o quanto um termo contribuiu para a similaridade entre todos os documentos; a qualidade da variância do termo, que define em quantos documentos o termo apareceu pelo menos uma vez; a seleção do termo indexador, que define a capacidade discriminatória de um termo e a ordenação baseada na entropia, que calcula a qualidade de uma palavra no documento.

Outro modo para o calculo de palavras relevantes é o *Term Frequency Linear,* utiliza-se um fator de ponderação para que os termos que apareçam na maioria dos documentos tenham um fator de representação semelhante a outros termos que não apareçam em tantos documentos diferentes (SOARES; PRATI; MONARD, 2008).

A equação é definida por *linear(t) = 1 – (dt)/N*, onde o numero de vezes em que o termo aparece (dt) é divido pelo numero de documentos da base, subtrai-se -1 do resultado.

Se um termo aparecer pelo menos uma vez em todos os documentos, ocorrerá que os fatores de ponderação se anularão, para isso utiliza-se o fator de suavização, que aumenta em 10% o N (número de documentos) para que o t (número de vezes que um termo apareça) possa obter valor de ponderação.

Geralmente, nos processos de mineração de texto, observam-se padrões como resultado de suas descobertas. O problema reside no fato de que se faz necessário a ponderação dos termos ditos como descoberta, de modo a aperfeiçoar o resultado.

Identificada a necessidade de um modelo que expressasse a mineração de texto, Tan (1999) desenvolveu um modelo básico. No modelo da figura 14 proposto por Tan (1999), observa-se um processo de duas fases.

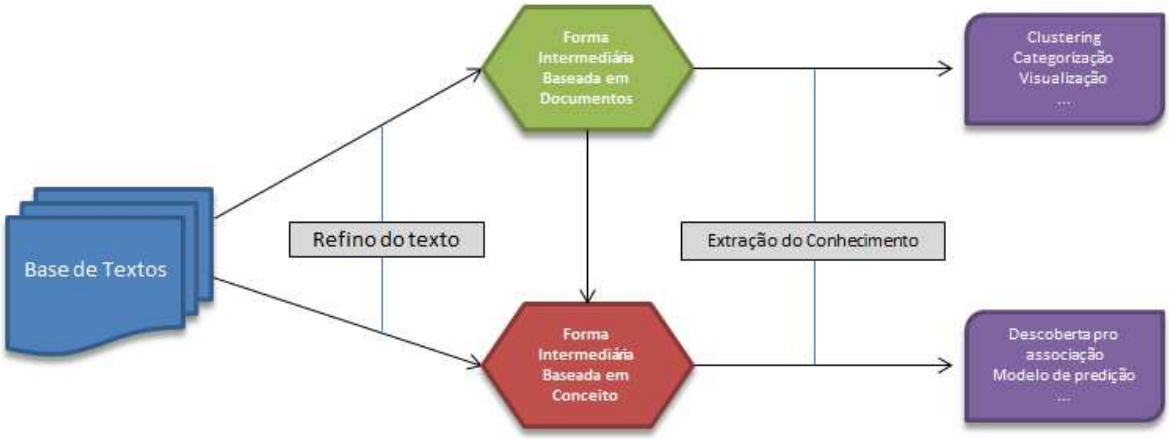

FIGURA 14 - MODELO BÁSICO DE MINERAÇÃO
FONTE: Adaptado de TAN (1999)

A primeira fase compete ao refino (preparação) do texto, que transforma textos de escrita livre em textos com forma semiestruturada, provendo o que o autor chama de forma intermediária em documentos. A partir da forma intermediária (FI), pode-se orientar a mineração de texto para dois propósitos distintos:

- **forma Intermediária baseada em documentos**: nesse modelo cada entidade representa um documento, a extração de conhecimento é feita, por exemplo, pó métodos como clustering e categorização.
- **forma intermediária baseada em conceito**: nesse modelo cada entidade representa um conceito, a extração de conhecimento é feita, por exemplo, por modelos de predição e descobertas associativas.

Ainda, FI baseada em documento pode se transformar em FI baseada em conceito, tratando a FI baseada em documento como etapa antecessora a FI baseada em conceito.

2.7 PODER JUDICIÁRIO

Segundo o Portal Brasil (2012), o poder judiciário é um órgão que possui autonomia administrativa. Forma, juntamente com os poderes legislativo e executivo, os pilares da governança do Brasil, e tem a função de garantir os direitos individuais, coletivos e sociais de cidadãos, entidades e Estado.

O Poder Judiciário é estruturado em sete diferentes órgãos, são eles:

- **Supremo Tribunal Federal (STF)**: Composto por 11 ministros nomeados pelo Presidente da República, com a aprovação do Senado Federal, o STF possui a função de garantir o cumprimento da constituição, além de regulamentar as normas constitucionais.
- **Superior Tribunal de Justiça (STJ)**: O STJ possui, sob a luz de interpretar de modo uniforme da legislação federal, os encargos de julgar causas criminais relevantes, que envolvam nomes e órgãos públicos.
- **Justiça Federal (JF)**: A Justiça Federal julga os casos em que figurem entre os atores a União Federal e empresas públicas federais.
- **Justiça do Trabalho**: A justiça do trabalho possui juízes representantes nos Tribunais Regionais do Trabalho e ministros do Tribunal Superior do Trabalho, além de juízes da primeira instância, atende casos em haja conflito entre empregador e empregado.
- **Justiça Eleitoral**: Os juízes eleitorais atuam na primeira instância, no Tribunal Regional Eleitoral (TER) e ministros que atuam no Tribunal

Superior Eleitoral (TSE). Tem como objetivo garantir e monitorar o direito ao voto previsto na constituição e julgar casos de irregularidade em cargos públicos ocupados por pessoas escolhidas por meio de votação direta.

- **Justiça Militar**: Composta por juízes da primeira e segunda instância e por ministros do Superior Tribunal Militar (STM). Processa e Julga crime militares.
- **Justiças Estaduais**: De competência de cada estado, a Justiça Estadual é composta por juízes de direito e desembargadores. Julga qualquer caso que não compete aos outros órgãos aqui citados.

Na figura 15 se observa um organograma simplificado do poder judiciário, destacando os órgãos que lidam com os processos que o presente trabalho se apoia:

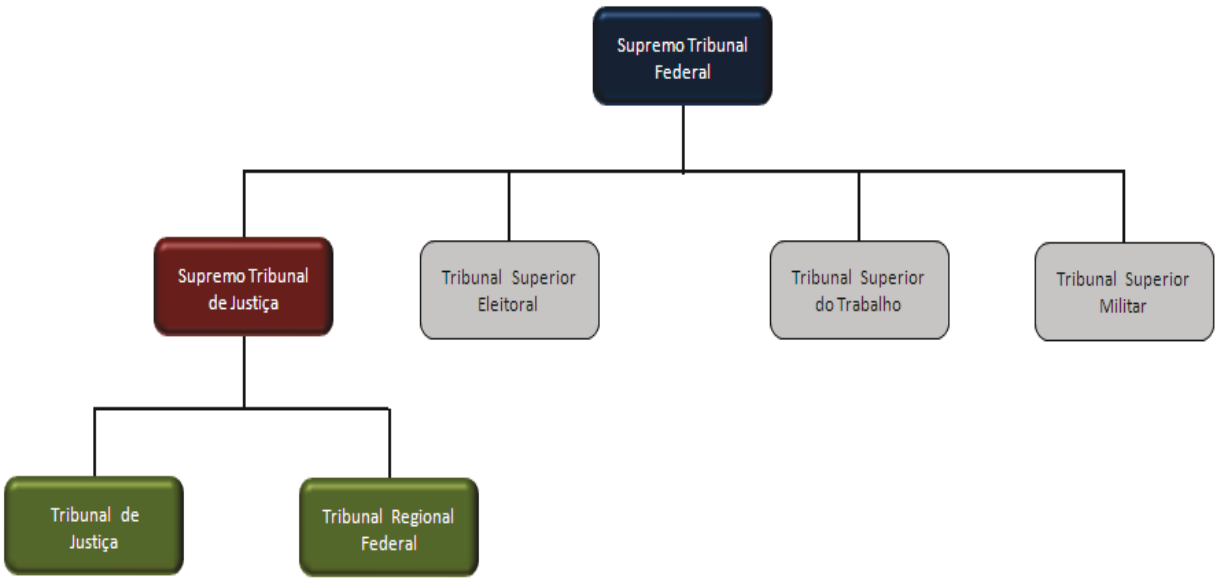

FIGURA 15 - ORGANOGRAMA BÁSICO DO PODER JUDICIÁRIO
FONTE: O AUTOR (2013)

Os casos encaminhados à justiça são julgados na primeira instância, composta por juízes federais ou estaduais. Caso haja recurso por algumas das partes, o processo se qualifica a segunda instância, caso em que o apelo é enviado ao Tribunal Regional Federal, aos Tribunais de Justiça e aos Tribunais de Segunda Instância. Ainda havendo a possibilidade do apelo por algumas das partes, o processo é encaminhado à última instância, ou seja, o Supremo Tribunal Federal.

3 METODOLOGIA DA PESQUISA

Os processos realizados no projeto caracterizam-no como um “trabalho de pesquisa experimental, quando selecionada as ferramentas, observa-se os efeitos que as mesmas causarão no objeto (GIL, 2008).” No caso, as ferramentas e técnicas serão aplicadas em determinada base de textos, e a partir de então se realizarão as análises conformes os resultados que essas variáveis impactaram no objeto de estudo.

Para atingir o objetivo do trabalho, primeiro se faz necessária a construção de uma base de dados, assim seguindo os passos do método KDD. Para tanto, é preciso validar dentre os milhares de processos encontrados, os que apresentam características suficientes para compor a base.

Trabalhando os casos específicos de comércio eletrônico do Paraná, encontra-se no portal do Tribunal de Justiça do Estado do Paraná uma aba de consultas para jurisprudência. É nessa seção em que se podem consultar os processos que já foram julgados.

A partir de uma pesquisa simples, onde o usuário insere dados em apenas um campo, ou a partir de uma pesquisa refinada (com campos e utilização de ranking de termos, por exemplo), pode-se obter acesso ao download de processos de diferentes matérias.

No presente trabalho, coletam-se todos os processos eletrônicos do tribunal regional federal que envolva o tema “comércio eletrônico”, formando assim, uma base de dados composta por arquivos do mesmo tema. Ressalta-se ainda que só serão utilizados processos disponíveis em ambiente virtual, de modo a tornar possível o reconhecimento de caracteres e, consequentemente, a mineração de texto.

Após a construção, se processa a base em um sistema de nuvem de tags, que irá prover a contagem das palavras, método semelhante a *bag of words*.

Após construir e processar a base sob a luz do método descrito, utilizar-se-á o software PreText, o programa é livre e permite com que o usuário, dentre outras funções, realize a mineração de texto.

A partir do início de um novo projeto criado no software, é possível selecionar os parâmetros da mineração, especificando o diretório onde os documentos serão lidos, bem como a heurística da ação, como *TF-IDF*, *Team Frequêncy*, *Team Occurrences* e *Binary Team Ocurrences*, n-grama etc. Após a escolha, o software realiza a mineração de texto provendo os resultados conforme a configuração escolhido.

3.1 BASES DE DADOS DE PROCESSOS JURÍDICOS: COMÉRCIO ELETRÔNICO

A base de processos jurídicos baseadas no comércio eletrônico foi criada a partir do *download* dos arquivos relacionados ao tema no portal do tribunal da justiça do Paraná.

A seleção dos processos criou uma base interessante mesmo sem qualquer análise estatística ou de mineração, uma vez que aglomera todos os processos judiciais de um tema específico. Isso foi possível graças à ferramenta de busca avançada do portal. Vale ainda afirmar que foram apenas considerados processos identificados como jurisprudência, ou seja, processos que, mesmo na primeira instância, já foram julgados.

Três termos orientaram a busca para a formação da base, são eles “comércio eletrônico”, “compras pela internet” e *“e-commerce”*. Para o termo “comércio eletrônico”, foram recuperados 64 processos, já para o termo “e-commerce”, foram recuperados 13 processos, por fim, para o termo “compras pela internet”, foram recuperados 4 processos. Em um total de 81 documentos.

Utilizando a aba de conectores de busca disponíveis na pesquisa avançada, foi possível recuperar os documentos conforme o termo de recuperação. Ainda na pesquisa avançada, selecionaram-se somente acórdãos, e a pesquisa foi realizada por tema da matéria.

Todos os processos analisados foram caracterizados como acórdão, que corresponde ao mesmo que uma sentença, com a diferença de que na sentença a decisão é tomada por um julgador único de primeira instância, e o acórdão e composto por um conjunto de julgadores em instância superiores, também conhecido como apelação. Nesse modo, todos os processos relacionados a esse termo já passaram por um juizado menor, e foram julgados em segunda instância.

Também é prudente ressaltar que os recursos são, em sua totalidade, acionados pelas empresas de comércio eletrônico, ou seja, a empresa não concordando com a sentença da primeira instância recorre a instâncias superiores, com o propósito de anular ou amenizar o veredito dado na primeira instância.

3.2 FERRAMENTA PARA A MINERAÇÃO DE TEXTO

Para o projeto, fora utilizado a ferramenta PreText, por ser um software livre, de código aberto, adaptado a língua portuguesa, e que atende as necessidades do projeto. O PreText é um programa orientado ao objeto, desenvolvido na linguagem de programação Perl, que utiliza o conceito da *bag of words* para análise, predição e mineração de textos. As métricas utilizadas pelo programa são: *tokenização*, *stemming*, taxonomias, Remoção de *Stopwords*, cortes de palavras baseada em frequência, valores de atributos, suavização e normalização.

Outra análise fora realizada pela contagem das palavras, para tanto se utilizou a ferramenta de nuvem de tag chamada *tagcrowd,* disponível de forma gratuita no site tagcrowd.com, através da inserção dos textos a ferramenta faz a contagem das palavras, utilizando o conceito da *bag of words.*

3.3 PROCEDIMENTOS METODOLÓGICOS

Após a formulação da base, dividiram-se as jurisprudências entre seus termos de busca (como já citado) e também entre causas ganhas (que favoreceram quem entrou com o recurso), e causas perdidas (que favoreceu o recorrido, o que já havia ganhado a causa em primeira instância).

Nessa etapa, foram identificados 23 processos em que a turma recursal concedeu provimento ao recursante, ou seja, para a empresa de e-commerce que entrou com o recurso contra o cliente, e 58 processos em que a turma recursal negou provimento ao recursante, ou seja, mantiveram a decisão do juiz de primeira instância favorável ao cliente.

Primeiramente se dividiu a base em três grupos, formados pelos termos de busca no site do Tribunal de Justiça do Paraná, são eles “Comércio Eletrônico”, “E-commerce” e “Compras pela internet”.

Feita a divisão, procurou-se aplicar a mineração de textos de diferentes maneiras para realizar análises. A figura 16 ilustra o processo de mineração de texto da presente pesquisa.

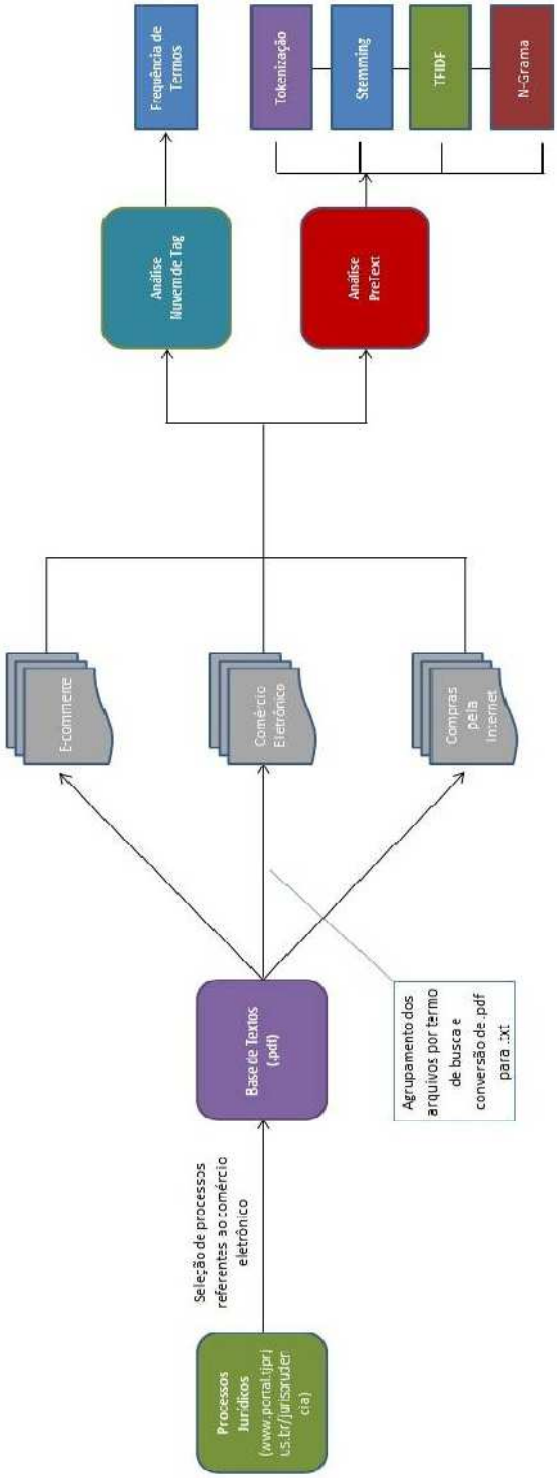

FIGURA 16 - PROCESSO DE MINERAÇÃO DE TEXTO DA PRESENTE BASE
FONTE: O AUTOR (2012)

Todas as análises realizadas no trabalho se caracterizam por serem associativas como descritas por Tan (1999), o fato se da, pois a mineração de texto escritos em português não possuem compatibilidade com determinados algoritmos de descoberta que se encaixam bem para a língua inglesa, por exemplo. A seção quatro detalha a descrição dos processos de mineração de texto realizados nesta pesquisa.

3.3.1 Software PreText

O PreText[1] é um software de processamento de texto desenvolvido em linguagem Perl pelo Laboratório de Inteligência Computacional (LABIC), pertencente ao Departamento de Ciência da Computação do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo.

No presente trabalho, as bases foram dividas entre seus termos de busca, e subdivididas entre recursos ganhos e recursos perdidos, logo após esse processo se realizou a conversão dos arquivos, originalmente baixados em .pdf, para a extensão .txt – essa conversão é necessária para deixar os arquivos compatíveis com os procedimentos do software, que podem ser observados na figura 17.

[1] PRETEXT. Disponível em: <http://sites.labic.icmc.usp.br/pretext2/>. Acesso em: 24 mar. 2013.

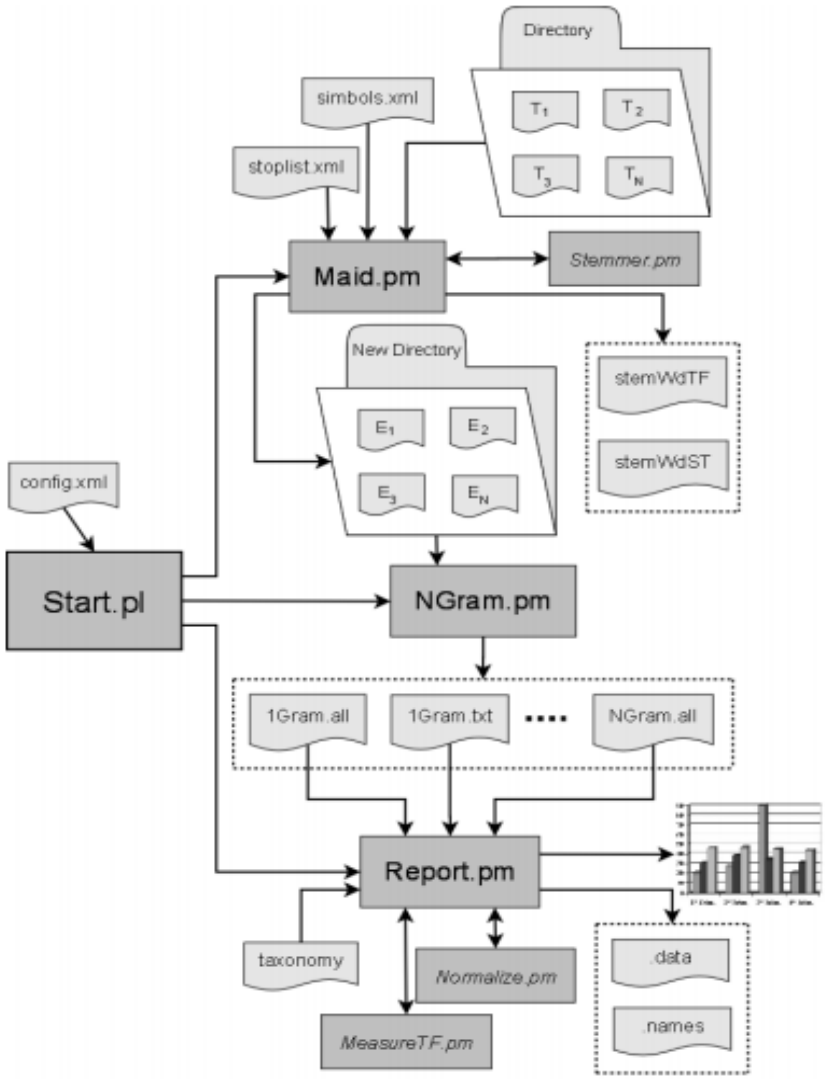

FIGURA 17 - PRETEXT II
FONTE: SOARES; PRATIL; MONARD (2008)

Para execução do programa é preciso configurar os arquivos de entrada (Figura 18), a execução é realizada pelo arquivo *start.pl*. Todos os comandos são executados pelo prompt de comando do windows.

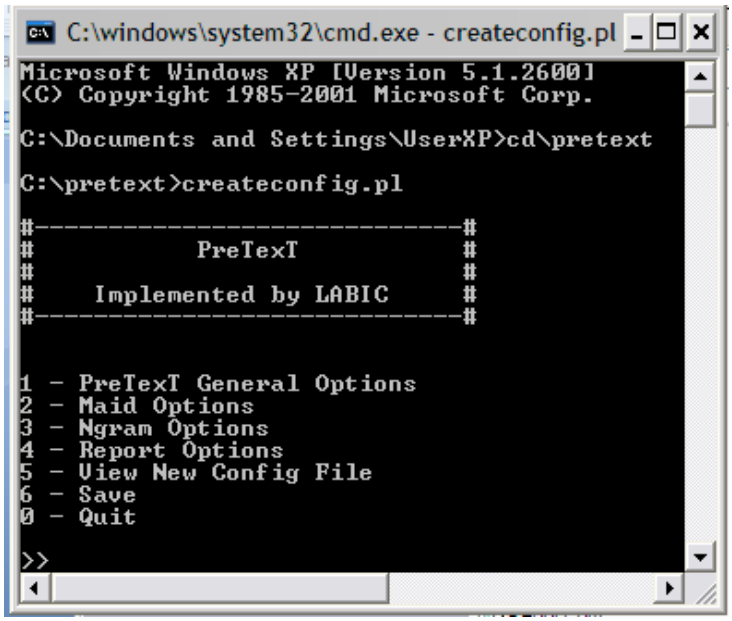

FIGURA 18 – TELA DE CONFIGRAÇÃO DO PRETEXT III
FONTE: SOARES; PRATIL; MONARD (2008)

Por meio do menu do arquivo *createconfig.pl* pode-se configurar todos as rotinas do software PreText.

Os arquivos são escritos em linguagem Perl, porém, com o propósito de integrar o software a outros tipos de estruturas, o arquivo config é construído em linguagem XML.

Na opção *1 – PreText General Options*, edita-se o diretório onde os arquivos estão armazenados, bem como o diretório onde os arquivos com as *sotpwords* (contruídas em XML) se encontram.

Os primeiros referem-se ao *stoplist.xml* e *simbols.xml*. O primeiro compete a formulação das stopwords, palavras que são retiradas do texto por não agregarem valor ao texto, como preposições, por exemplo. No trabalho foram utilizadas tanto stopwords em português quanto inglês, devido a propriedades dos arquivos, que podiam conter a palavra estrangeira *"page"* (página), por exemplo. Já o arquivo *simbols.xml* se refere a remoção de símbolos, pontuações gramaticais etc.

Com o diretório das bases textuais definidos e com os arquivos de *stopwords* e símbolos configurados é possível executar o módulo *maid.pm,* que pode ser configurado pela opção 2 no arquivo *createconfig.pl,* esse módulo é responsável pela limpeza dos documentos (remoção de *stopwords* e símbolos) e o *stemming*

para a *token*ização. Nessa opção habilita-se ou não o *stemming,* bem como a tokenização.

O processo de *stemming* é realizado através da função *snowball* criada por Porter em 1980 (ORENTO, HYUK, 2009). Nessa etapa reduz-se os plurais; as palavras no sentido feminino são reorganizadas (Ex: japonesa -> japonês); os advérbios são reduzidos, o sufixo do advérbio é “mente”, mas nem todas as palvras que terminam com esse sufixo são advérvbios, portanto o algoritmo conta com uma lista de excessão; reduz-se as formas diminutivas e superlativas bem como os sufixos de nomes, pronomes e verbos; removem-se as últimas vogais das palavras.

O processo do arquivo *maid.pm* gera um novo diretório com os arquivos de entrada já processados, tendo os *token*s criados com base no *stemming,* além disso, são gerados os arquivos *stemWdST.all* e *stemWdFT.all,* o primeiro contém todos os *stem*s identificados na base por ordem alfabética, quanto que o segundo arquivo contém todos os *stem*s identificados na base por frequência.

```
conhec : 64(20/23)
        conheceram:1
        conhecia:1
        conhecimentos:1
        conheceu:2
        conhecimento:7
        conhecer:18
        conhecido:34
```

FIGURA 19 – EXEMPLO DE *STEM*MING
FONTE: O AUTOR (2013)

Notam-se na base alguns casos de *under-stemming*, em que dois ou mais *tokens* são reduzidos equivocadamente para dois *stems* diferentes. E também casos de *mis-stemming* quando são retirados os prefixos de determinados *tokens* formando *stems* sem significado.

A opção 3 – *N-gram Options* compete a configuração dos n-grama, nele são definidos a frequência mínima, e também a quantidade de N (junções) que serão testadas.

As contrações (*stems*) são observadas no novo diretório gerado pelo módulo *maid.pm,* e são com base nos *stems* que o módulo *n.grama.pm* é executado. A métrica utilizada para o corte de n-grama é baseada no cálculo tf-idf.

Quanto maior o N do grama, maior será a profundidade semântica que poderá ser encontrada, o problema reside no fato de que para se utilizar um n-grama

alto, é preciso ter uma base de um tema muito específico, sem perder a variância de termos de um texto para o outro, ou seja, o cenário ideal para a formulação de um n-grama é um texto onde a semântica seja extremamente semelhante entre um documento e outro, mas utilizando diferentes termos para expressa-la.

No presente trabalho, testou-se 1-grama, 4-grama, 5-grama e 9-grama, sendo que se testaram os n-gramas de 1 a 10. Os resultados não são relevante após o 4-grama.

A base de dados de jurisprudências relacionadas ao comércio eletrônico pode ser vista como específica quanto ao seu tema, porém, a semelhança dos termos utilizados desde a explanação até a sentença acaba por dificultar o processo de descoberta. Ainda sim, observa-se que, por conta da singularidade de cada caso, podem-se realizar análises na base.

Utilizando os N-grama como entrada de dados se é executado o módulo *Report.pm,* esse módulo provê uma tabela atributo-valor, cada atributo (*token*) recebe um valor em cada documento conforme a métrica tfidf, ainda, o módulo cria gráficos de variância de n-grama, mostrando o valor máximo atingido por um determinado atributo conforme o N.

É também nesse módulo que é realizada a normalização dos documentos, além da classificação via taxonomia na tabela atributo-valor, caso essa seja habilitada.

Por fim, a opção 4 – *Report Option* que compete ao módulo *report.pm* , permite com que se habilite a opção de relatórios gráficos e de documentos por palavras, e provê gráficos de frequência das métricas utilizadas, bem como a contagem de termos por documentos, de forma geral, o *report.pm* gera relatórios de *feedback* para a pesquisa, ideal para definir os cortes por frequência e documentos da base de texto.

Os documentos providos pelo software permitem uma análise associativa por meio da observação dos *tokens* gerados, tanto separados quanto unidos pelos n-grama. A seguir a análise dos agrupamentos de documentos pela observação dos n-grama.

4 DESCRIÇÃO DO EXPERIMENTO E ANÁLISE DE RESULTADOS

Nesta seção estão detalhadas as etapas da mineração de texto realizada na base de dados de processos jurídicos e apresentadas as análises dos resultados obtidos.

4.1 ANÁLISE POR FREQUÊNCIA DE PALAVRAS

Com o propósito de compreender os termos mais frequentes na base dentro dos dois grupos (processo ganho, processo perdido), utilizou-se um medidor de frequência de palavras que possui como resultado uma nuvem de tags.

Trata-se do site tagcrowd.com, que permite com que a partir da entrada de arquivos textuais se faça a contagem de palavras por meio da nuvem de *tag*. Segundo o site, o *TagCrowd* é definido por um aplicativo web para visualização da frequência de palavras, criado em 2006.

Com a ação, podem-se realizar análises prévias com a intenção de orientar a mineração de texto, bem como facilitar a interpretação dos resultados pós-mineração.

No Quadro 1 pode-se analisar os resultados após o processo de contagem de palavras.

TB: Comércio Eletrônico			
Concedeu Provimento	Frequência	Negou Provimento	Frequência
Serviços	112	Produto	226
Produto	103	Valor	166
Valor	90	Serviços	148
Marca	88	Indenização	139
Contrato	82	Dano Moral	136
Juros	74	Pagamento	89
Pagamento	53	Comércio	75
Direito	52	Dever	61
Crédito	50	Materiais	61
Pedido	43	Pedido	60
Consumo	41	Fornecedor	59
Dano Moral	39	Empresa	58
Mercado	38	Consumo	57
Indenização	36	Prestação	56
Uso	31	Condenação	55
Agravante	29	Razão	55
Atividade	26	Mercado	53
Empresa	26		
Ônus	26		

QUADRO 1 - FREQUENCIA DE TERMOS: COMÉRCIO ELETRÔNICO
FONTE: O AUTOR (2013)

No que compete ao termo de busca (TB) “Comércio Eletrônico” dos 20 processos que concederam provimento (em um total de 64 para esse TB), observa-se que a palavra “Serviços” fora a que apresentou maior frequência, com 112 citações, seguida da palavra “Produto”, com 103 citações, de fato os domínios de internet que trabalham com venda de produtos oferecem tanto um serviço (o de compra e entrega) quanto um produto (o site em si, e também o produto comprado).

Logo abaixo, com 90 citações, nota-se a palavra “Valor”, seguida de “Marca” com 88 citações e a palavra “Contrato”, com 82 citações. Logo, observa-se que os recursos que concederam provimento ao recursante citam, com base na análise de frequência de termos, os aspectos do produto, do site e também, não em menor importância, o contrato.

Já dos 44 documentos que negaram provimento ao TB “Comércio Eletrônico”, observa-se que a palavra com mais frequência foi “Produto”, com 226 citações, seguida da palavra “Valor”, com 166 citações e “Serviços” com 148 citações. É

interessante ressalvar que a palavra “Contrato”, a quinta mais citada nos que concederam provimento, não aparece nem nas 17 primeiras citadas no que tange os documentos que negaram provimento ao recursante.

Dos que negaram provimento, as palavras mais usadas, referem-se ao produto, ao valor, ao serviço, e ao dano moral.

Uma possível conclusão para a análise do quadro 1 é a de que, quando se concede provimento ao recursante (no caso, as empresas de e-commerce recorrendo a decisão desfavorável a mesma em primeira instância), as palavras referentes a contrato, pagamento, serviços e produtos são destacadas pela turma recursal.

Já quando se nega o provimento (mantendo a decisão do juiz de primeira instância), observa-se que as palavras referentes a valor, produto e dano moral se destacam pela turma recursal.

A seguir a análise pelo termo de busca “*E-commerce*”:

TB: e-commerce			
Concedeu Provimento	Frequência	Negou Provimento	Frequência
Serviço	48	Serviços	21
Consumo	32	Produto	12
Produto	25	Retido	11
Ônus	20	Indenização	10
Pagamento	16	Consumo	9
Valor	16	Valor	9
Empresa	15	Pagamento	8
Inversão	15	Prestação	8
Tecnologia	15	Pedido	7
E-Commerce	14	Tecnologia	7
Julgamento	14	Contrato	6
Omissão	14	Falha	6
Prestação	14	Informação	6
Informação	13	E-Commerce	5
Dano Moral	13	Vulnerabilidade	4

QUADRO 2 - FREQUENCIA DE TERMOS: E-COMMERCE
FONTE: O AUTOR (2013)

Para o TB “e-commerce”, no que se refere aos 3 documentos que negaram provimento, nota-se que a palavra “Serviço” fora citada 48 vezes, sendo a primeira da lista, seguida dos termos “Consumo”, “Produto” e “Ônus”. A quantidade reduzida

de documentos que possuem resultado favorável ao recursante dificulta quaisquer análises com base na frequência. O impacto de um único documento pode ser decisivo na contagem de palavras.

Ainda sim, as palavras mais citadas referem-se aos produtos, serviços, valor e pagamento.

Já para os 10 documentos do mesmo TB que negaram provimento, observa-se que palavra mais citada também foi “Serviço”, seguida dos termos “Produto”, “Retido” e “Indenização”. Conclui-se que para o TB e-commerce as turmas recursais analisaram os casos baseados no serviço prestado pela organização de comércio eletrônico, e os acontecimentos em torno dos produtos que a mesma comercializa.

O último grupo, do termo de busca “Compras pela Internet”, não contabiliza nenhum processo que teve como resultado provimento do recurso, como observado na tabela abaixo.

TB: compras pela internet			
Concedeu Provimento	Frequência	Negou Provimento	Frequência
-	-	Dano Moral	17
-	-	Valor	16
-	-	Produto	13
-	-	Compra	10
-	-	Entrega	10
-	-	Pagamento	10
-	-	Indenização	9
-	-	Ausência	7

QUADRO 3 - FREQUENCIA DE TERMOS: COMPRAS PELA INTERNET
FONTE: O AUTOR (2013)

Usando o TB “compras pela internet”, observam-se apenas resultados desfavoráveis ao recursante. Dos 4 processos analisados, ressalta-se que, com 17 ocorrências, o termo mais citado é “Dano Moral”, seguido de “Valor”e “Produto”. Assim, pode-se concluir que o dano moral foi assunto relevante dentre as turmas recursais do presente TB.

No Quadro 4 apresenta analisa-se a frequência das palavras divididas apenas por recursos ganhos e perdidos.

Recurso Ganho		Recurso Perdido	
Serviços	160	Produto	251
Produto	128	Valor	191
Valor	106	Serviços	169
Marca	88	Indenização	158
Contrato	82	Dano Moral	153
Juros	74	Pagamento	107
Consumo	73	Comércio	75
Pagamento	69	Pedido	67
Dano Moral	52	Consumo	66
Direito	52	Prestação	64
Crédito	50	Dever	61
Ônus	46	Materiais	61
Pedido	43	Fornecedor	59
Empresa	41	Empresa	58
Mercado	38	Condenação	55
Indenização	36	Razão	55
Uso	31	Mercado	53

QUADRO 4 - FREQUENCIA DE TERMOS: COMPRAS PELA INTERNET
FONTE: O AUTOR (2013)

Tanto nos recursos ganhos, quanto nos recursos perdidos, os três termos mais frequentes são “Produto”, “Valor” e “Serviços”. No âmbito do comércio eletrônico, pode-se afirmar que de forma simplificada são esses três termos que representam as interações entre cliente e empresa em um processo de compra.

A empresa oferece o serviço de compra *online*, e o cliente, por usa vez, adquire o produto. Uma possível interpretação é a de que quando um dos serviços ou produtos não corresponde à expectativa, e a empresa não soluciona o problema, o cliente procura seus direitos legais.

O termo “Contrato” está entre os cinco termos mais citados nos recursos ganhos, por outro lado, nos recursos perdidos, o termo não se encontra na lista dos mais citados. É possível que as turmas recursais que julgam procedente o recurso acionado pelas empresas julguem os aspectos contratuais para tal ato.

Já o termo “Dano Moral” está entre os cinco termos mais citados nos recursos perdidos, e é o nono nos recursos ganhos. Possivelmente as turmas recursais consideram o grau do dano moral, o ponto onde se ultrapassam os simples dissabores da vida, sobrepõe outros fatores de modo a ser de direito do cliente o recebimento de uma indenização (quarto termo mais citado) advinda da empresa.

4.2 ANÁLISE TF-IDF E N-GRAMA

A seguir a análise pelas técnicas TF-IDF e N-GRAMA, aplicadas pelo software PreText, baseada na Figura 16

4.2.1 Análise Pretext: Grupo Comércio eletrônico

Abaixo a análise dos n-grama gerados pelo processamento dos 64 documentos correspondentes ao grupo comércio eletrônico, sendo 44 documentos onde a turma recursal negou o recurso, e 20 onde a turma recursal concedeu o recurso.

No PreText, foram removidas 1024 *stopwords* e foram gerados 1308 *stems*. A contagem de *stems* foi feita a partir da soma dos resultados obtidos com o grupo comércio eletrônico – recurso negado e grupo comércio eletrônico – recurso provido.

Para os N-grama acima de 4 não foram encontrados resultados relevantes.

4.2.1.1 Comércio Eletrônico - Análise 1-grama

Para o 1-grama configurado para uma frequência mínima de 10 *stem*, foram gerados 250 *stems* para os 20 processos que proveram o recurso e 540 *stems* para os processos em que a turma recursal negou o recurso. O 1-grama nada mais é do que o processamento dos *stems* conforme a métrica, nesse caso o tfidf, no gráfico 1 se observa a frequência de termo do 1-grama para o subgrupo recurso negado:

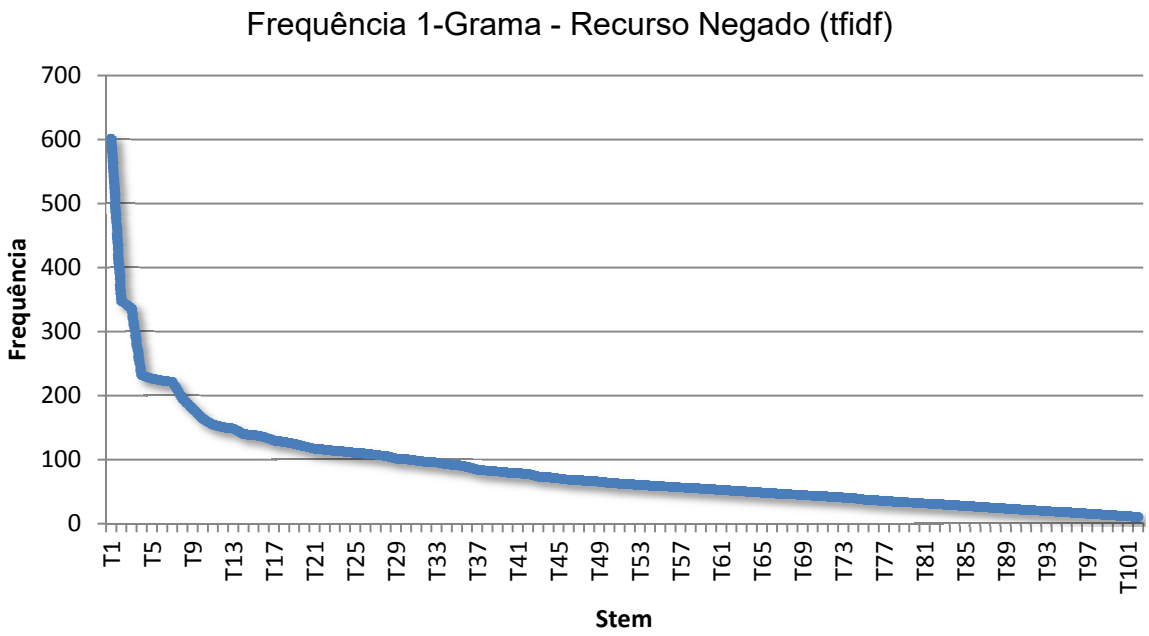

GRÁFICO 1 - FREQUENCIA 1 GRAMA: RECURSO NEGADO (TFIDF)
FONTE: O AUTOR (2013)

Nos resultados para 1-grama se observa o *stem* “ser” (T1) com 601 ocorrências em 44 dos 44 documentos, assim todos os documentos possuem esse *stem*. Em uma análise de todos os *token*s pertencentes a esse *stem*, conclui-se que o mesmo em uma análise isolada, sem a junção de outros *stem* em um n-grama de maior grau, não possui relevância, pois se trata de um verbo de ligação.

```
ser : 601(44/44)
      serem:2
      fossem:2
      seria:4
      sejam:6
      era:7
      sera:8
      sido:10
      fosse:13
      foram:17
      sendo:58
      foi:128
      es:142
      ser:204
```

FIGURA 20 – 1 GRAMA: SER
FONTE: O AUTOR (2013)

O terceiro *stem* com maior frequência (T3), “dan”, com 347 ocorrências em 37 dos 44 documentos se refere a palavra dano ou o seu plural, danos. Esse *stem* faz menção tanto a danos em bens materiais quanto ao dano moral.

A palavra “dano” conta com 131 ocorrências, a maioria conectada com a palavra “morais”, que possui 136 ocorrências. Assim, nota-se que a palavra dano ou

danos está presente em 84,1% dos documentos. Os danos morais ou materiais são temas recorrentes nesse agrupamento, sendo então um termo que caracteriza os documentos desse grupo.

O *stem* “produt” se refere ao termo “produto” e ao seu plural “produtos”, e o terceiro termo mais recorrente com 223 ocorrências em 36 dos 44 documentos, o termo está presente em 81,8% dos documentos.

A relação entre produtos e danos justifica as jurisprudências acerca do tema, bem como a busca por uma resolução do problema, explicada pelo décimo *stem* “indeniza”, com 139 ocorrências em 33 dos 44 documentos que se refere à indenização a que tem direito o recorrido até a primeira instancia.

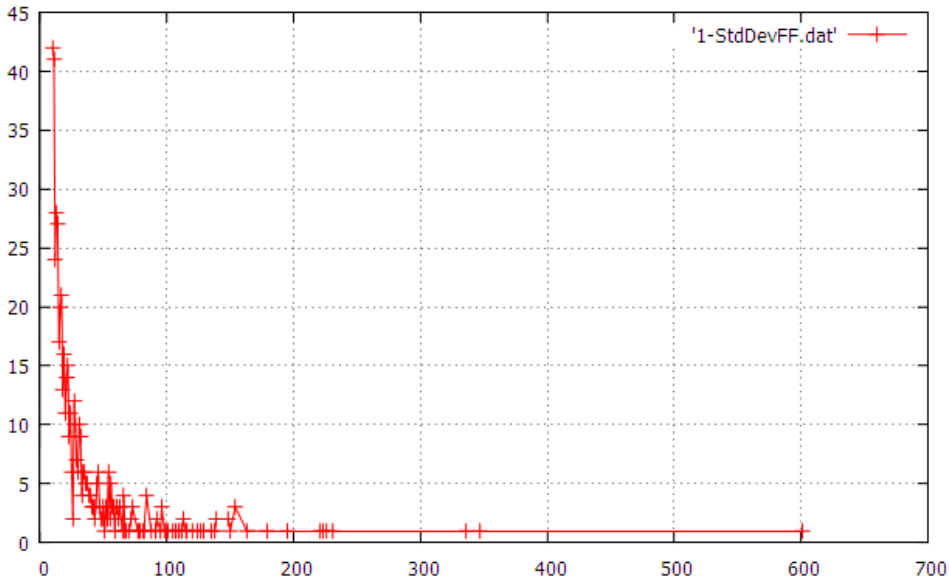

GRÁFICO 2 – 1 GRAMA: GRÁFICO TERMO X FREQUÊNCIA – RECURSO NEGADO
FONTE: O AUTOR (2013)

O gráfico 2, gerado pelo PreText, permite com que se analise quantos *stem*s possuem a mesma frequência, esse tipo de gráfico é útil para entender até que ponto é interessante analisar os n-gramas, o ponto máximo para o presente grupo são os 41 *stem*s com a mesma frequência de 9 ocorrências. Na soma de 10 *stem*s é notada a frequência de 31 ocorrências, a partir de então todos os *token*s se assemelham em sua frequência menor que 10 ocorrências.

No gráfico 3, observa-se a frequência de termo do 1-grama para o rupo comércio eletrônico recurso provido:

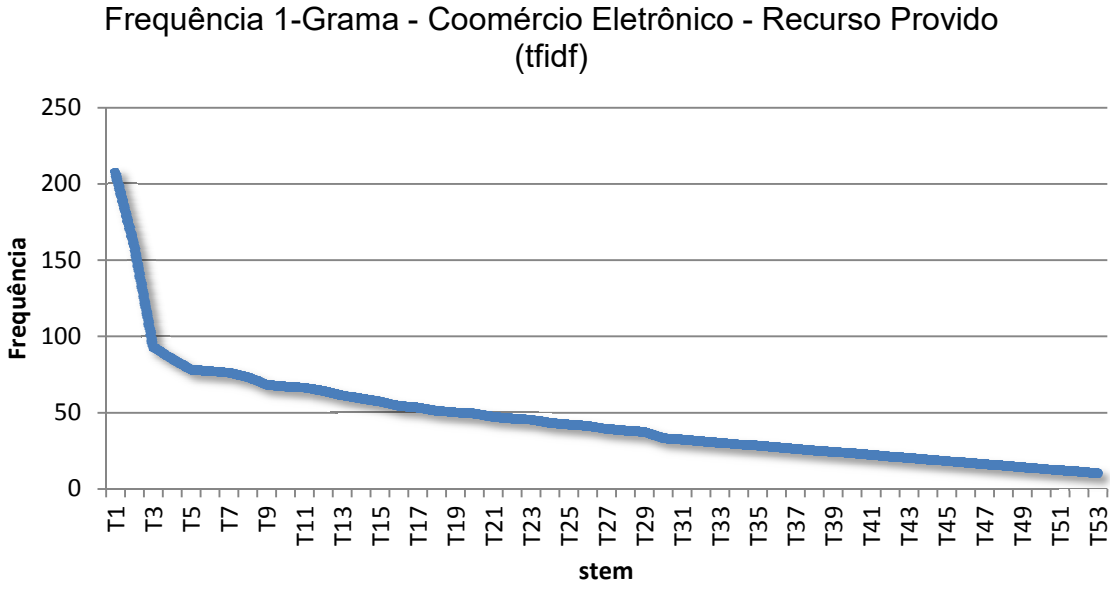

GRÁFICO 3 - FREQUENCIA 1-GRAMA: COMÉRCIO ELETRÔNICO: RECURSO PROVIDO (TFIDF)
FONTE: O AUTOR (2013)

Assim como nos 1-grama dos recursos negados, a primeira palavra com 200 ocorrências em 20 dos 20 documentos é o verbo de ligação se. A partir de então se analisam algumas diferenças entre o 1-grama do recurso provido e o 1-grama do recurso negado.

O *stem* “jur” aparece em 85 ocasiões em 12 dos 20 documentos, sendo o quarto *stem* mais frequente, sendo que 71 ocorrências são referentes a palavra Juro ou Juros. O sexto *stem* mais frequente diz respeito aos contratos, com 78 ocorrências em 8 dos 20 documentos, é interessante ressaltar que esse *stem* não é observado nem entre os 10 primeiros no grupo recurso negado. Com 73 ocorrências em 16 dos 20 documentos nota-se o *stem* “val”, referente as palavras valor e valores.

No gráfico 4, visualiza-se a semelhança de frequência entre os *stem*s:

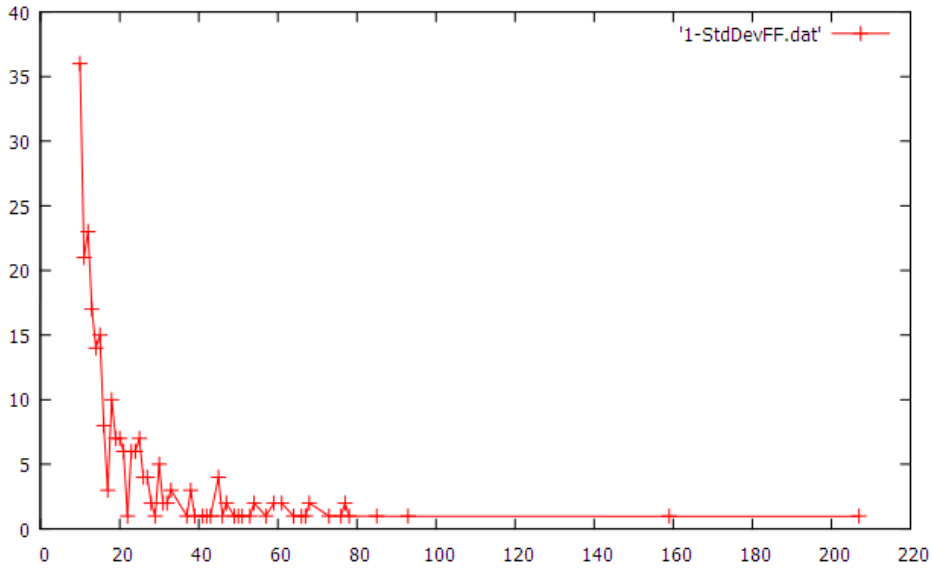

GRÁFICO 4 – 1 GRAMA: GRÁFICO TERMO X FREQUÊNCIA – RECURSO NEGADO
FONTE: O AUTOR (2013)

Para esse agrupamento a maior semelhança entre os *stem*s são a soma de 10 com 36 de frequência A partir de 17 *stem*s a semelhança entre a frequência se torna menor que 10.

No quadro 5, observa-se as principais diferenças entre os agrupamentos na base comércio eletrônico:

Análise 1-Grama - Comércio Eletrônico		
Concedeu Recurso	X	**Negou Recurso**
Termo "Contrato"entre os 10 mais Frequentes	X	Termo "Deveres" entre os 10 mais frequentes
Termo "Juros" mais frequente que "Danos"	X	Termo "Danos" mais frequente que "Juros"
Termo "Responsabilidade" entre os mais frequentes	X	Termo "Consumidor" entre os mais frequentes
Termo "lei"entre os mais frequentes	X	Termo "Indenização"entre os mais frequentes

QUADRO 5 - ANALISE 1 GRAMA: COMÉRCIO ELETRÔNICO
FONTE: O AUTOR (2012)

Ressalta-se que as análises são associativas, portanto dependente do processo cognitivo do analista dos resultados.

Na coluna “Concedeu Recurso” referente a um dos grupos do comércio eletrônico, observam-se os termos Lei, Contrato, Responsabilidade e Juros. Já na coluna “Negou Recurso”, encontram-se os termos Consumidor, Indenização, Danos e Deveres.

Interpreta-se que a turma recursal, no caso dos processos que favoreceram a empresa prestadora do serviço de e-commerce, tende a considerar fatores como os aspectos legais, o contrato entre cliente-empresa e além da responsabilidade (nesse caso de algumas das partes) e os Juros.

Já para os processos que favoreceram o cliente, verificam-se aspectos do consumidor, dos danos morais e materiais e da indenização sentenciada.

4.2.1.2 Grupo Comércio Eletrônico: Análise 4-grama

Para o 4-grama, junção de 4 *stem*s, são encontrados 2920 *stem*s para os processos em que a turma recursal concedeu provimento de recurso e 4922 *stem*s para os processos em que o recurso foi negado. No gráfico 5, nota-se a frequência dos 4-grama para os recursos negados:

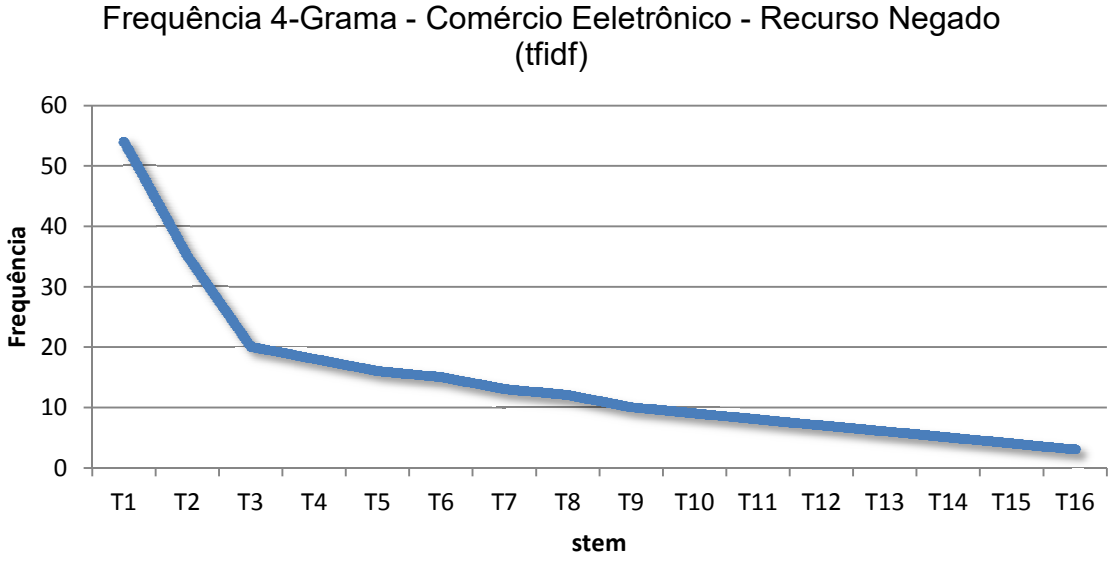

GRÁFICO 5 - FREQUÊNCIA 4 GRAMA: COMÉRCIO ELETRÔNICO RECURSO NEGADO (TFIDF)
FONTE: O AUTOR (2013)

O T1 possui 3 *stem*s com 54 ocorrências em 11 dos 44 documentos são indicativos de que o presente documento pode ser acesso via meio eletrônico,a partir de T2 até T4 ressaltam-se *stem*s de frases corriqueiras em processos judiciais, como “julgamento presidido pelo juiz”, ou “participam os senhores e juízes, todos estes termos não possuem relevância a pesquisa.

O conjunto T5 possui como um de seus *stem* com o nome da juíza da Primeira Turma Recursal do Paraná Dra. Léo Henrique Furtado Araújo, com 16 ocorrências em 15 dos 44 documentos. Outro nome presente e da Dra. Ana Paula Kaled A. Rotunno, também integrante da Primeira Turma Recursal.

O 4-grama T10 e T11 possuem em seus conjuntos a junção Dano Moral. O 4-grama T12 o conjunto “redução do quantum indenizatório” expresso pelo *stem* “redu_quantum_indenizat_rio” (observado um caso de *mis-stemming* no prefixo “rio”), com 7 ocorrências em 7 dos 44 documentos. Já o 4-grama T9 é compreendido pelo novo valor fixado do valor indenizatório.

No gráfico 6, analisa-se a semelhança de frequência entre os *stem*s:

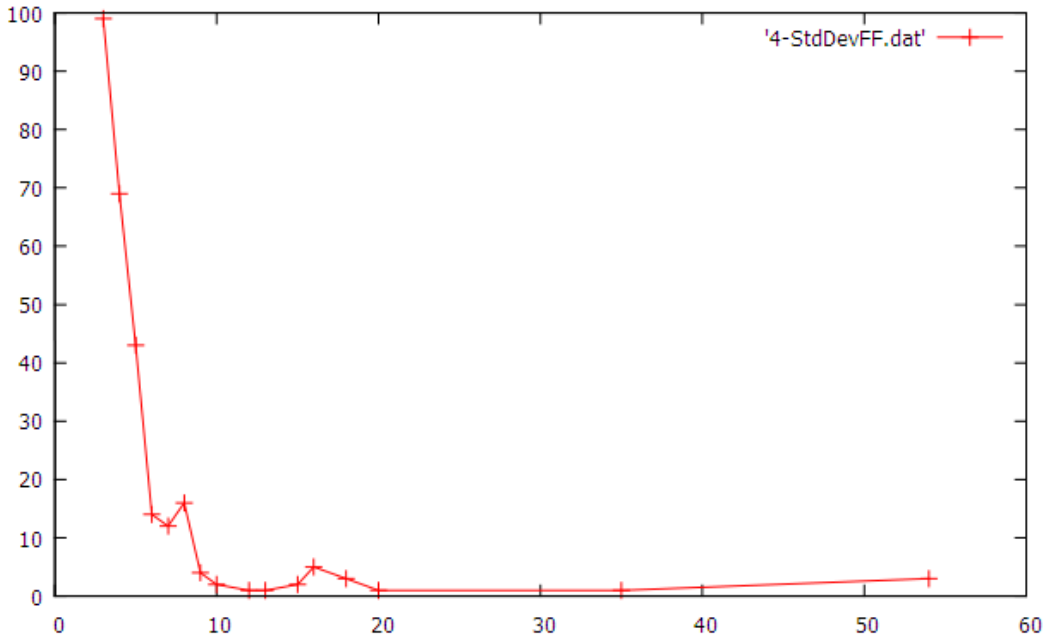

GRÁFICO 6 - 4-GRAMA: COMÉRCIO ELETRÔNICO GRÁFICO TERMO X FREQUÊNCIA - RECURSO NEGADO
FONTE: O AUTOR (2013)

No gráfico 6, observa-se que aproximadamente 100 *stem*s possuem frequência de 3 ocorrências, até a linha de 10 ocorrências o gráfico apresenta uma drástica queda, elevando-se apenas com o conjunto de 5 *stem*s que possuem frequência de 5 ocorrências. A partir da queda para 1 *stem*s a frequência não ultrapassa 3 ocorrências.

No gráfico 7, visualiza-se a frequência dos 4-grama para o grupo comércio eletrônico recursos providos:

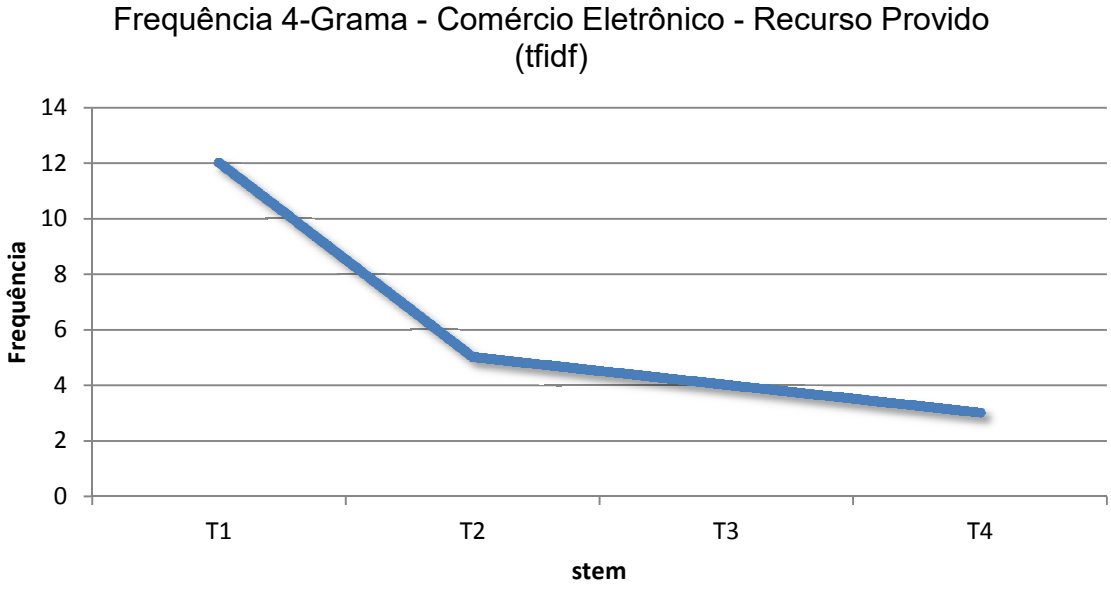

GRÁFICO 7 - FREQUENCIA 4 GRAMA: COMÉRCIO ELETRÔNICO RECUROS PROVIDO (TFIDF)
FONTE: O AUTOR (2012)

O conjunto T1 composto 3 *stem* com 12 ocorrências em 7 dos 20 documentos cada um é composto por indicativos de que o documento pode ser acessado em meio eletrônico, irrelevante para a pesquisa. O conjunto T3 possui como um de seus *stem* os nomes dos Juízes Dr. Horário Ribas Teixeira, Dra. Leo Henrique Furtado Araújo, com 3 ocorrências em 3 dos 20 documentos cada um.

No gráfico 8, pode-se verificar a semelhança entre os termos quanto sua frequência:

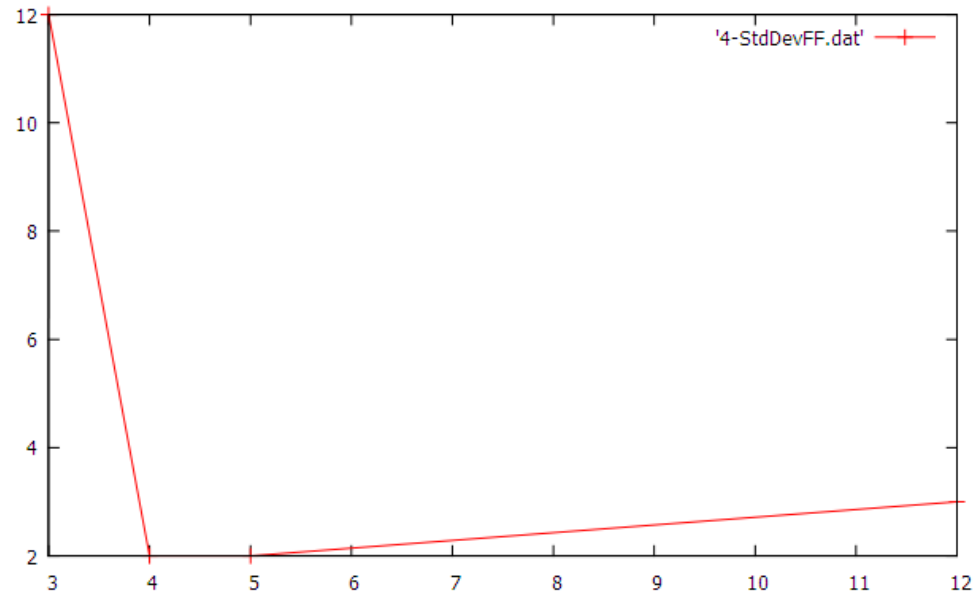

GRÁFICO 8 – 4-GRAMA: COMÉRCIO ELETRÔNICO GRÁFICO TERMO X FREQUÊNCIA – RECURSO PROVIDO
FONTE: O AUTOR (2012)

O gráfico mostra que a base de 20 documentos não possui muita representatividade, uma vez após a observação de 3 conjuntos de *stem* com frequência igual a 12, nenhum outro conjunto ultrapassa a frequência de 4 ocorrências.

4.2.2 Análise Pretext: Grupo Compras pela internet

No grupo compras pela internet são observados 4 documentos, que representam apenas recursos negados pela turma recursal. O número de documentos é insuficiente para gerar *stem*s para n-grama acima de 1.

Para o processamento em 1-grama foram encontrados 33 *stem*s.

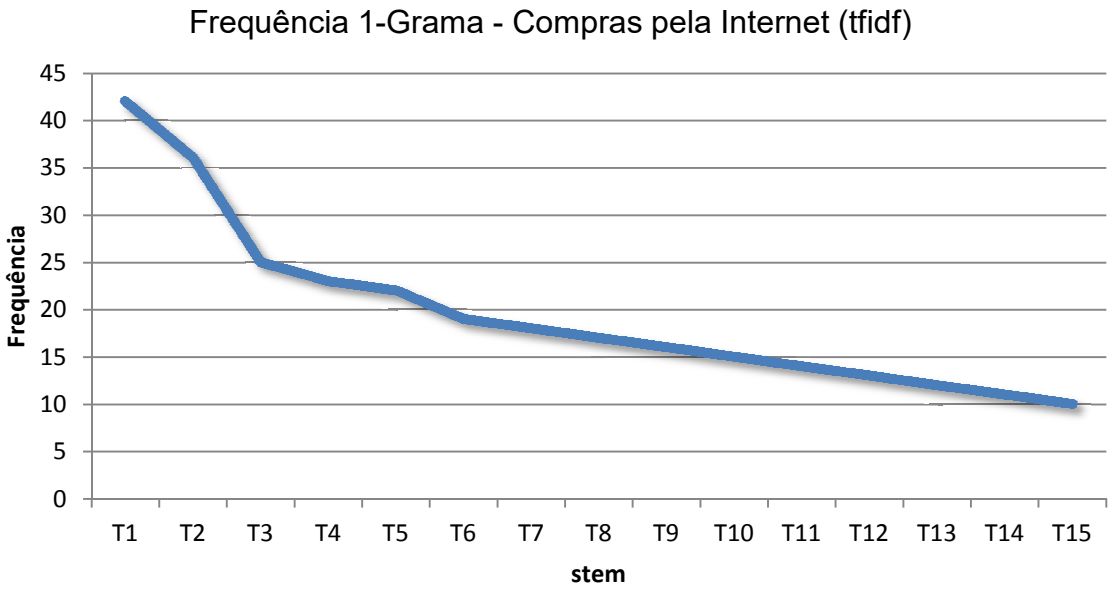

GRÁFICO 9 - FREQUENCIA 1 GRAMA: COMPRAS PELA INTERNET (TFIDF)
FONTE: O AUTOR (2013)

O *stem* com mais frequência é o verbo de ligação “ser”, com 44 ocorrências em 4 dos 4 documentos. O *stem* T5 se refere aos danos tanto em bens quanto em moral, com 22 ocorrências em 4 dos 4 documentos.

O *stem* “moral” possui 17 ocorrências em 3 dos 4 documentos, seguido das 16 ocorrências em 4 dos 4 documentos do *stem* “produtos”. O *stem* “entrega” possui 12 ocorrências em 4 dos 4 documentos.

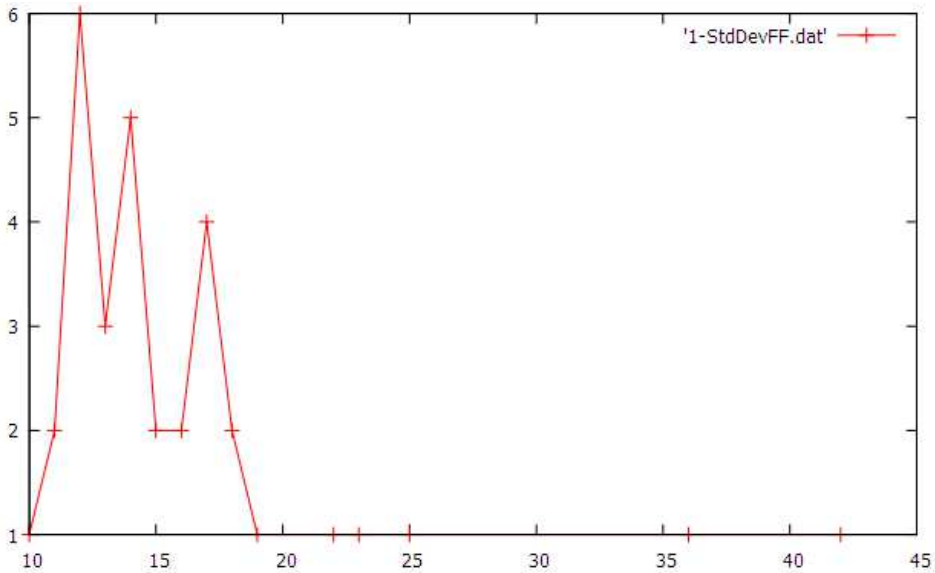

GRÁFICO 10 - 1-GRAMA: GRÁFICO TERMO x FREQUËNCIA - COMÉRCIO ELETRÔNICO
FONTE: O AUTOR (2013)

O gráfico 10 demonstra que em frequências mais baixas são observados maiores agrupamentos de *stems,* quando a frequência ultrapassa as 17 ocorrências os *stem* passam a variar entre 0 e 1.

4.2.3 Análise Pretext: Grupo E-commerce

Após o processamento do PreText na base e-commerce, foram encontrados 233 *stem*s, sendo 28 *stem*s nos 3 documentos que possuem resultado favorável ao recorrente e 205 *stem*s nos 10 documentos que possuem resultado favorável ao recorrido.

4.2.3.1 Grupo E-commerce Análise 1-grama

No gráfico 11 se observam os termos com maior frequência dentro dos documentos que mantiveram a decisão em primeira instância.

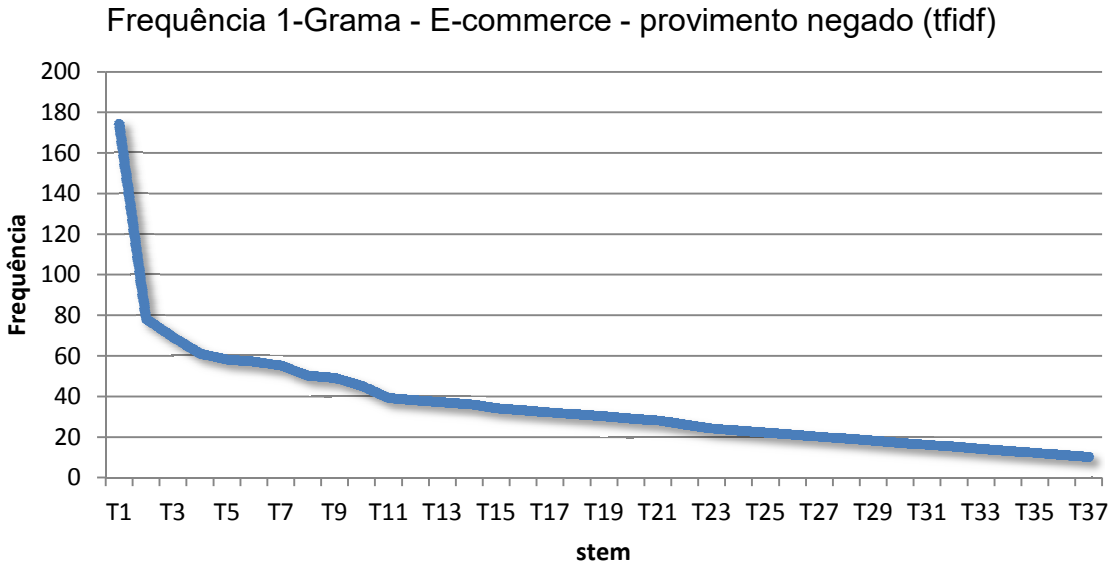

GRÁFICO 11 - FREQUENCIA 1 GRAMA: E-COMMERCE PROVIMENTO NEGADO (TFIDF)
FONTE: O AUTOR (2032)

Tendo T1 a T5 como *stem*s irrelevantes a pesquisa, nota-se em T6 o termo consumidor, com 57 ocorrências em 8 de 10 documentos. O *stem* lei possui 38 ocorrências em 9 dos 10 documentos, com uma aparição a menos se observa o termo dano. Próximo a esses termos se encontra o termo “produto”, com 36 ocorrências em 6 dos documentos, ou seja, 60% dos documentos deste

agrupamento citam o termo produto, e em 90% dos casos a palavra dano. A palavra “pagamento” possui 16 ocorrências em 7 dos 10 documentos.

Embora se observe as palavras produto, dano, lei e pagamento, o *stem* “entreg”, referente à palavra entrega, aparece apenas 2 vezes em 2 dos 10 documentos da base.

No gráfico 12, visualiza-se a frequência dos termos:

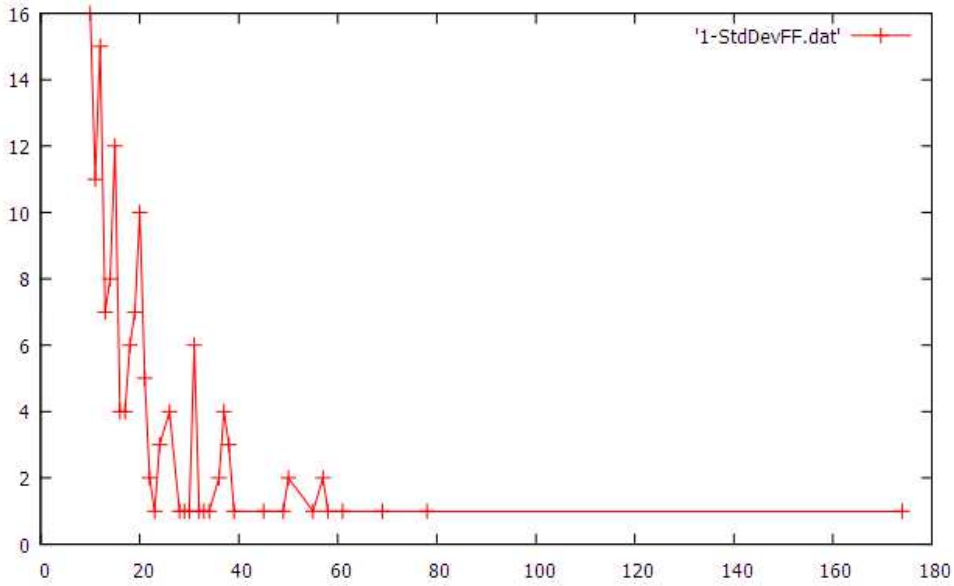

GRÁFICO 12 - 1-GRAMA: E-COMMERCE GRÁFICO TERMO x FREQUËNCIA - RECURSO NEGADO
FONTE: O AUTOR (2013)

No gráfico 12, pode-se visualizar que o pico de *stem*s que possuem a mesma frequência ocorre em um conjunto de 16 possuintes de 5 aparições, a partir da linha de frequência 57 os *stem*s se estabilizam.

Já para os 3 documentos que representam os recursos que foram favoráveis ao recorrente no termo de busca comércio eletrônico, realiza-se a seguinte análise com base nos *stem*s representados no gráfico 13:

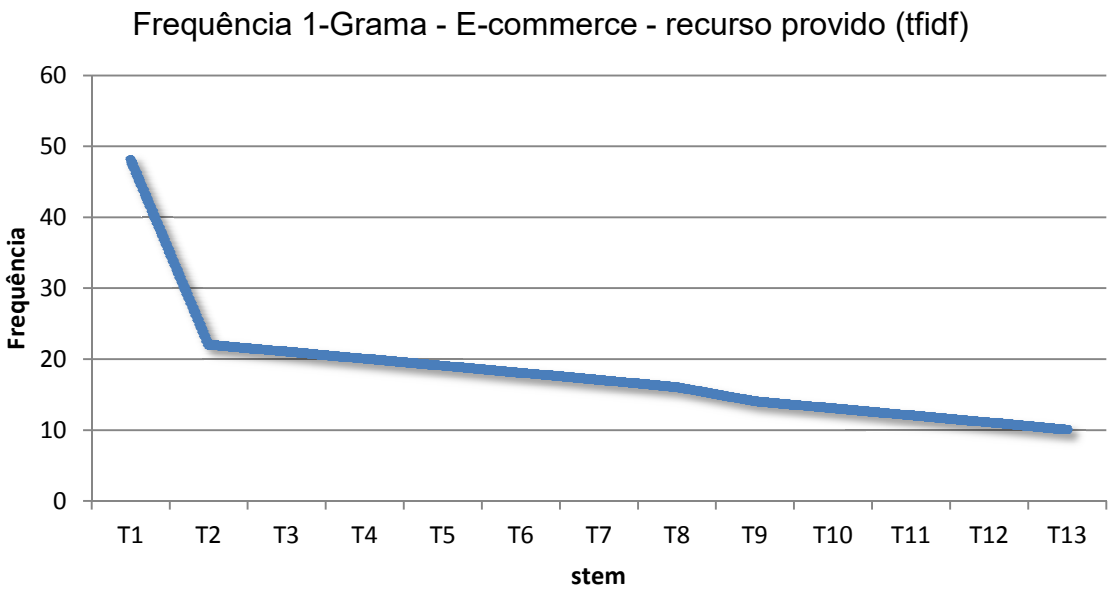

GRÁFICO 13 - FREQUENCIA 1 GRAMA: E-COMMERCE RECURSO PROVIDO (TFIDF)
FONTE: O AUTOR (2013)

O *stem* T2 referente a consumidores possui 22 ocorrências em 3 de 3 documentos, seguido de juros e produtos. Em um base pequena como essa a análise se torna impraticável devido ao número de documentos, em 3 documentos são encontrados alguns padrões que podem simplesmente não manter a proporção a medida em que a base aumenta.

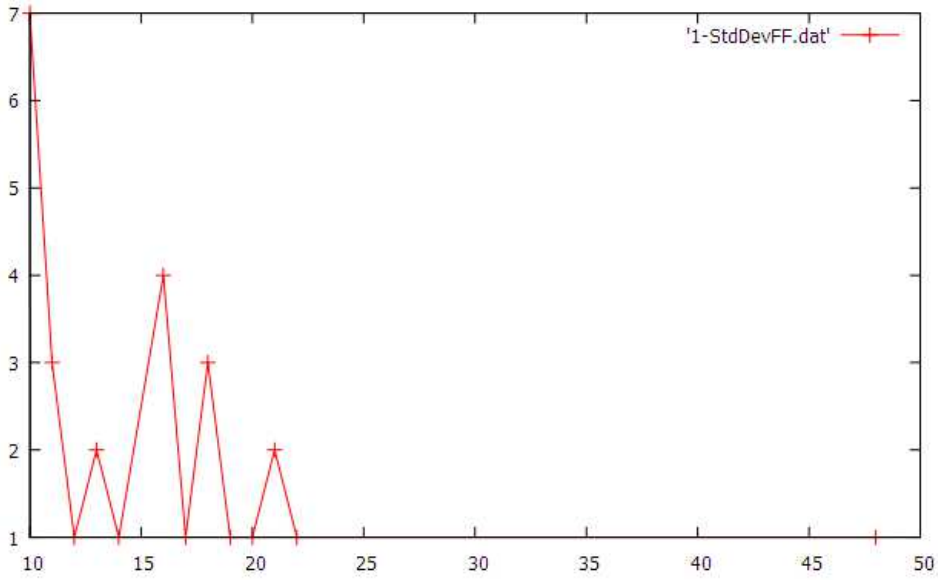

GRÁFICO 14 - 1-GRAMA: E-COMMERCE GRÁFICO TERMO x FREQUËNCIA - RECURSO PROVIDO
FONTE: O AUTOR (2012)

Isso se confirma a partir da análise dos agrupamentos por frequência, após identificado dois conjuntos de *stem* com 22 ocorrências cada um se observa uma estabilidade em 0 conjuntos de termos.

Não são visualizados casos de relevância para n-grama superior a 1.

4.2.4 Análise Pretext: Processamento geral

Na última etapa da análise, dividem-se todos os processos em apenas dois grupos, o grupo em que a turma recursal concedeu o recurso, favorecendo a empresa, e o grupo em que a turma recursal negou o provimento, favorecendo o cliente.

Nesse modo são analisados um total de 85 documentos, sendo 62 processos que possuem como resultado o recurso negado, e 23 artigos que possuem como resultado o recurso provido.

Do processamento no software PreText foram gerados 1707 *stem*s, sendo 389 referentes aos recursos providos e 1318 referentes aos recursos negados.

4.2.4.1 Processamento Geral: Análise 1 Grama

Dos recursos que foram negados pela turma recursal se oberva a geração de 669 *stem* para 1-grama, no gráfico 15, pode-se obervar a frequência dos *stem*s:

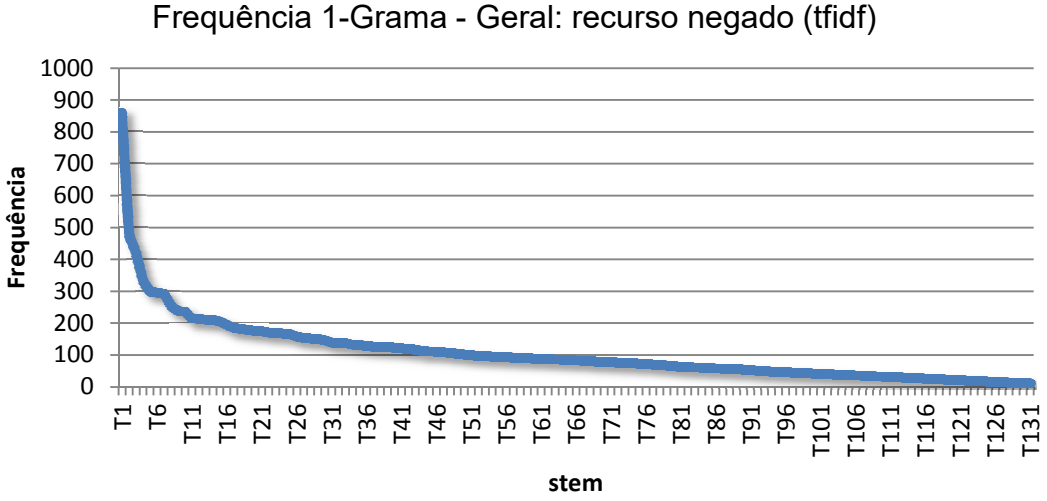

GRÁFICO 15 - FREQUENCIA 1 GRAMA GERAL: RECURSO NEGADO (TFID)
FONTE: O AUTOR (2013)

O *stem* T3 se refere ao termo dano, tanto moral como em bens, com 417 ocorrências em 54 dos 62 documentos, o terceiro termo mais recorrente, e primeiro com relevância, está presente em 84,3% dos documentos do grupo.

O *stem* T5 se refere ao consumidor com 295 ocorrências em 49 dos 62 documentos, presente em 76,6% dos documentos do grupo. Com o mesmo porcentual de frequência encontra-se o termo produto, com 291 ocorrências em 50 dos 62 documentos.

O *stem* T22 com 173 ocorrências em 42 dos 62 documentos se refere a “mor” que contém o *token* “morais”. Em um caso de *under-stemming* se observa também o *stem* “moral” que contém o *token* moral com 150 ocorrências em 35 dos 62 documentos, quando na verdade ambos os *token*s deveriam estar no mesmo *stem*.

O *stem* T33 se refere a responsabilidades, possuindo 136 ocorrências em 40 dos 62 documentos, com 5 ocorrências a menos, observa-se o *stem* “compr” que contém os *token*s referentes a compras. Logo após na condição de trigésimos sexto termos mais frequente se encontram os *token*s referentes à palavra “pagamento” com 126 ocorrências em 56 dos 62 documentos.

Abaixo o gráfico 16, que demonstra a semelhança entre conjuntos de *stem*s:

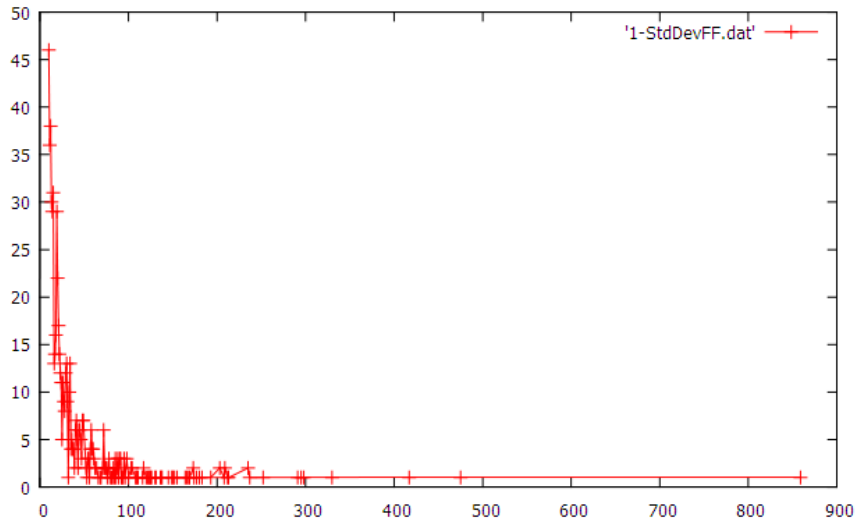

GRÁFICO 16 - 1-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO NEGADO
FONTE: O AUTOR (2013)

No gráfico se pode notar que o ponto máximo de semelhança de frequência entre os 1-grama é um conjunto de 10 *stem*s que possuem 45 ocorrências, a partir do conjunto de 25 1-grama a semelhança entre frequência entre 0 e 5 ocorrências. A

variância é contínua até o agrupamento de 200 *stem*s, quando a frequência segue de forma linear em 1 aparição de frequência por *stem*.

Para o 1-grama visualiza-se a seguinte curva de frequência:

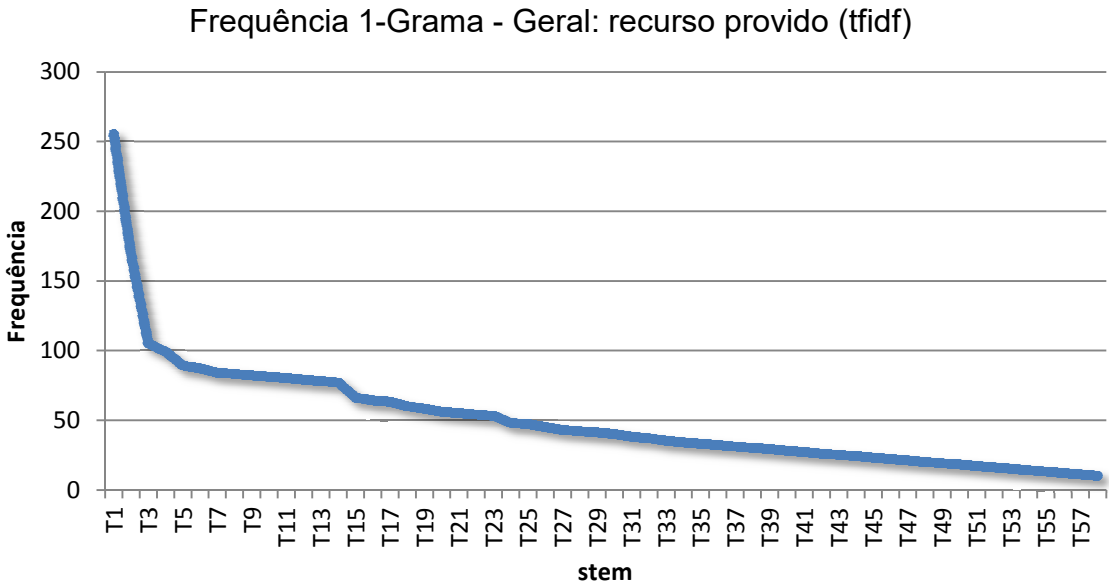

GRÁFICO 17 - FREQUENCIA 1 GRAMA GERAL: RECUROS PROVIDO (TFIDF)
FONTE: O AUTOR (2013)

O terceiro *stem* mais frequente, e primeiro a agregar relevância à pesquisa se refere ao *token* “juros”, com 105 ocorrências em 14 dos 23 documentos. Na quarta posição com 89 ocorrências em 18 de 23 documentos se encontra o *token* referente a palavra consumidor.

Na sexta posição com 87 ocorrências em 17 dos 23 documentos da base se encontra o *stem* que se refere aos *token*s dano e danos, logo após, com 84 ocorrências em 17 de 23 documentos se encontra o *token* contrato. Na décima posição se observa o *token* “dever” com 19 ocorrências em 23 documentos.

Com 78 ocorrências em 17 dos 23 documentos se encontra o *stem* para produtos, seguido de pagamento com 54 ocorrências em 19 dos 23 documentos.

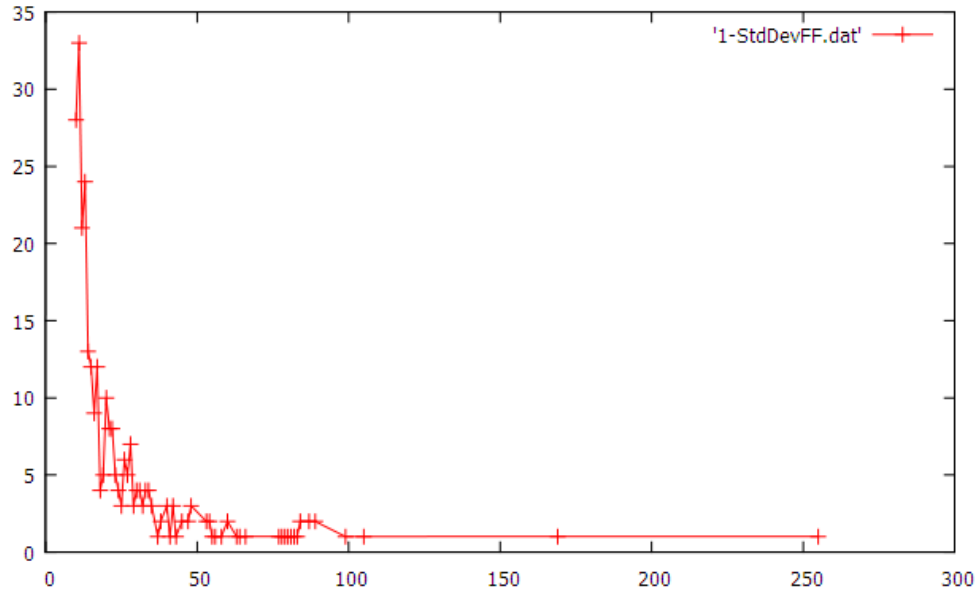

GRÁFICO 18 - 1-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO PROVIDO
FONTE: O AUTOR (2013)

Para o 1-grama do grupo dos que concederam provimento ao recurso, observa-se que 10 *stem*s possuem 32 ocorrências, sendo essa a maior frequência em comum entre os *stem*s. O gráfico possui acentuada queda estabilizando entre as frequências 0 e 5 para um conjunto de 50 *stem*s em diante.

No quadro 6 se podem observar as principais diferenças entre os processos qe provem recurso e os que negam:

Análise 1-Grama - Todos os Documentos		
Concedeu Recurso	X	**Negou Recurso**
Terceiro termo mais frequente: "Juros"	X	Terceiro termo mais frequente: "Danos"
Termo "Contrato" mais frequente que "Produto"	X	Termo "Produto" mais frequente que "Contrato"
Termo "Pagamento" entre os mais frequentes	X	Termo "Consumidor" entre os mais frequentes
Termo "Consumo" entre os mais frequentes	X	Termo "Indenização"entre os mais frequentes

QUADRO 6 - ANÁLISE 1 GRAMA: TODOS OS DOCUMENTOS
FONTE: O AUTOR (2013)

Na coluna dos processos que possuem como resultado o provimento do recurso, observam-se os termos Juros, Contrato, Pagamento e Consumo entre os

mais frequentes. Já na coluna dos que negaram o recurso, observam-se os termos Danos, Produto, Consumidor e Indenização.

Assim como na análise pela Frequência de Termos (TF) realizada na nuvem de tag, percebe-se, quando o recurso favorece a empresa, ou seja, o recurso é provido, que os principais termos se referem aos aspectos legais e contratuais.

Já para os casos em que o recurso é negado, favorecendo o cliente, observam-se termos relacionados aos problemas que o cliente enfrenta em caso de entrave no processo de compra pela internet. Assim, os termos mais frequentes são Dnos, Produto, Consumidor e Indenização.

4.2.4.2 Processamento Geral: Análise 4 Grama

No processamento da base de dados no software PreText configurado para 4-grama foram gerados um total de 421 *stem*s, sendo 53 dos recursos que concederam provimento e 368 dos recursos que negaram provimento.

No gráfico 19, visualiza-se a frequência dos 4-grama para os recursos negados:

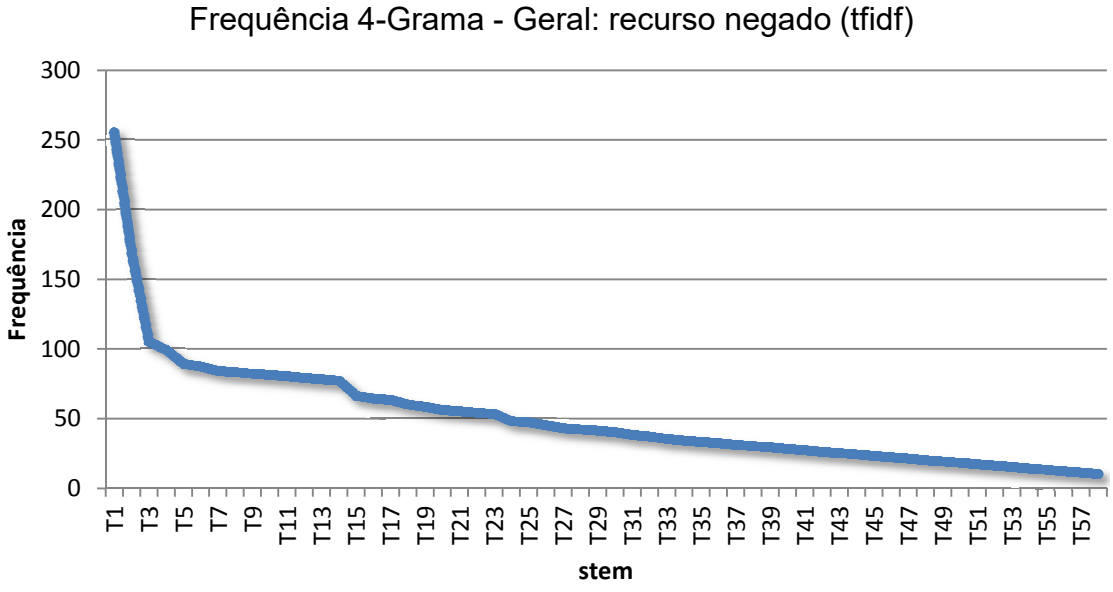

GRÁFICO 19- FREQUENCIA 4 GRAMA GERAL: RECURSO NEGADO (TFIDF)
FONTE: O AUTOR (2013)

Os quatro primeiros *stem*s não possuem relevância para a pesquisa por se tratarem de informes sobre a possibilidade do documento ser encontrado em ambiente virtual, além de termos presentes em todos os recursos inominados.

O *stem* T5 se refere a Dr. Léo Henrique Furtado Araújo com 26 ocorrências em 24 dos 64 documentos da base, outros juízes citados por ordem de frequência são a Dra, Ana Paula Kaled Accioly e o Dr. Moacir Antonio Dala Costa.

A junção dos termos “Dano” e “Moral” aparece em dois dos 4-grama, ambos *stem* com 12 aparições em 62 documentos.

A redução do quantum indenizatório possui 11 ocorrências em 11 dos 62 documentos. Assim, em 17,7% dos documentos da base é citada a redução do valor indenizatório previsto em primeira instância. O *stem* “val_indeniza_dan_moral” possui 8 ocorrências em 8 dos 62 documentos, e é identificado como a sentença “valor indenizatório por dano moral”.

No gráfico 20, analisa-se a frequência por ocorrência dos *stem*s do 4-grama para recursos negados:

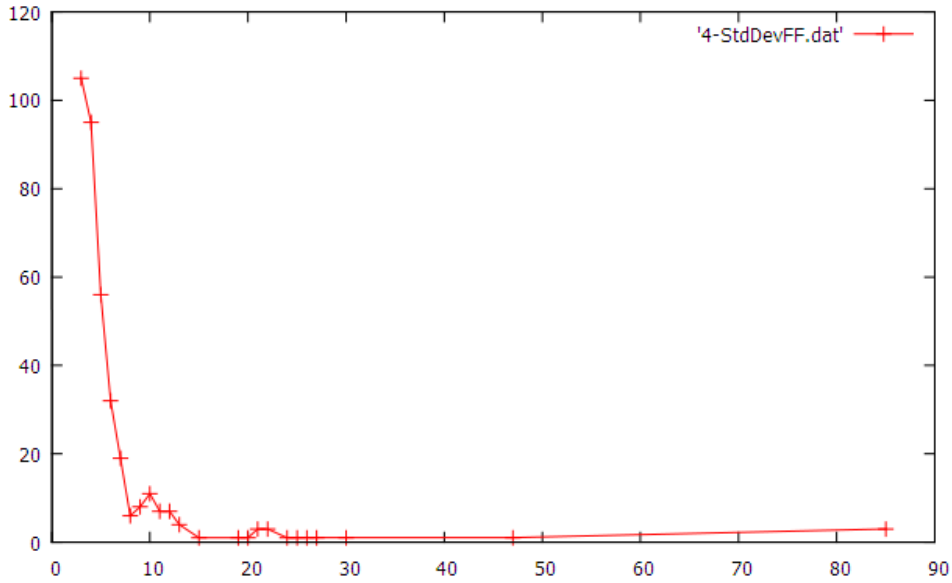

GRÁFICO 20 - 4-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO NEGADO
FONTE: O AUTOR (2013)

Visualiza-se que as frequências que possuem maior número de termos se encontram entre 0 e 10 ocorrências, com mais de 100 termos entre esses pontos. Tendo o pico observado nos 104 *stem* que possuem 3 ocorrências como sua frequência.

No gráfico 21, analisa-se a frequência por ocorrência dos *stem*s do 4-grama para os recursos concedidos:

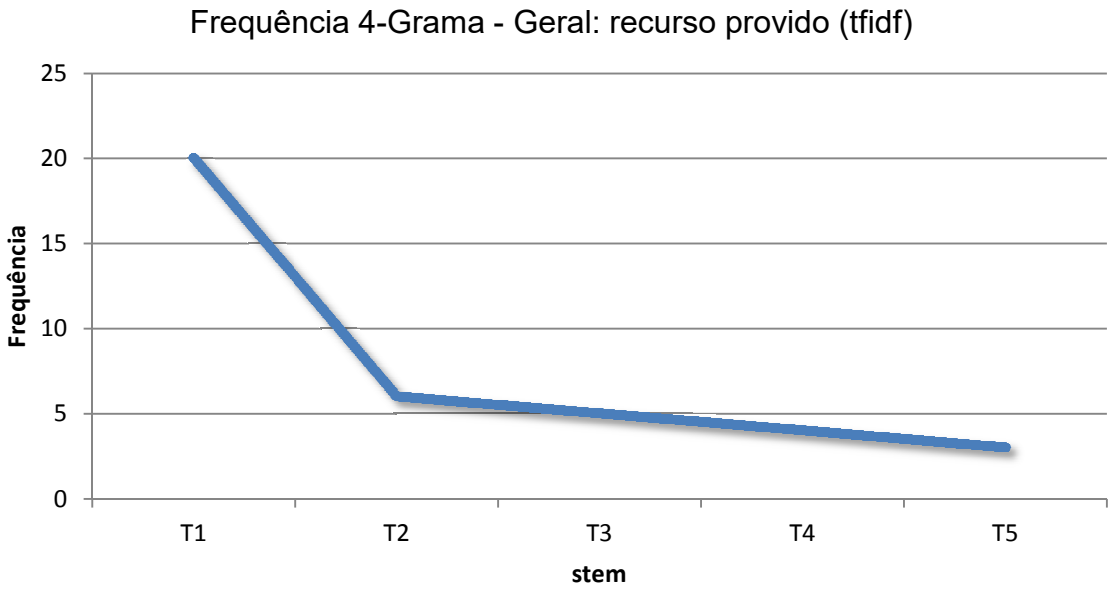

GRÁFICO 21 - FREQUENCIA 4 GRAMA GERAL: RECURSO PROVIDO (TFIDF)
FONTE: O AUTOR (2013)

O 4-grama T1 com frequência de 20 ocorrências, não se aplica a pesquisa, por se tratar de um informativo sobre a disponibilidade do documento em meios virtuais. O 4-grama T2 se refere a sentenças introdutórias do documento, presentes em todos os processos analisados.

Uma das descobertas interessantes do 4-grama são os nomes dos juízes relatores do processo, em 5 ocorrências em 4 de 23 documentos aparecem o nome do Juíz Dr. Horácio Ribeiro Teixeira, seguido da Juíza Dra. Ana Paula Kaled Accioly com 4 ocorrências em 4 dos 23 documentos, Dr. Helder Luiz Henrique Tacaguchi com 4 ocorrências em 3 dos 23 documentos e Dr. Leo Henrique Araujo com 4 ocorrências em 4 dos 23 documentos.

No gráfico 22, nota-se a frequência por ocorrência dos *stem*s do 4-grama para os recursos providos:

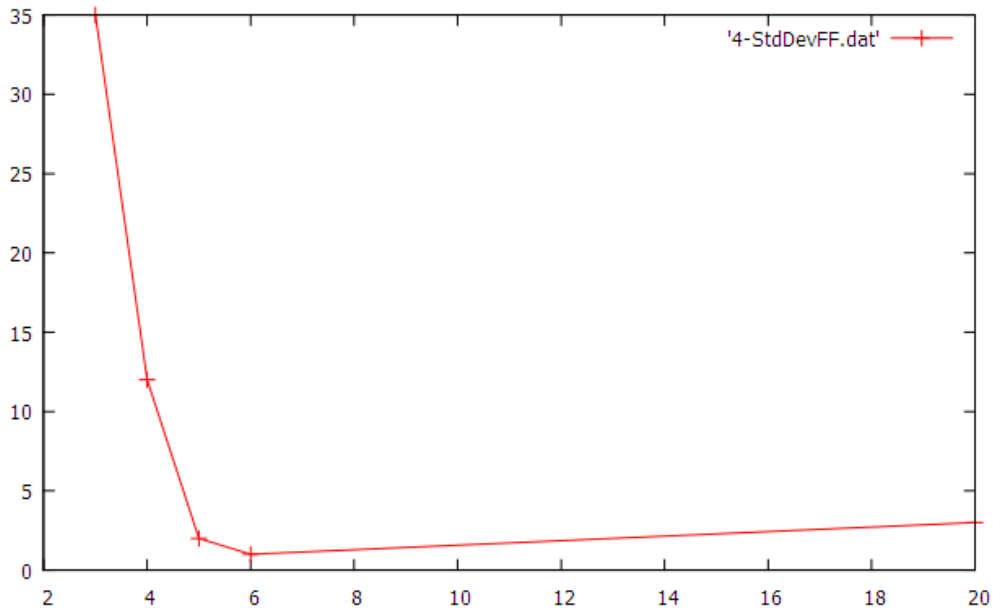

GRÁFICO 22 - 4-GRAMA: GRÁFICO TERMO x FREQUËNCIA - RECURSO PROVIDO
FONTE: O AUTOR (2013)

O gráfico demonstra que o pico da frequência de *stem* é encontrado com 35 *stem* do 4-gram que possuem 3 ocorrências, o aumento é gradual a partir da frequência 6.

O quadro 7 propõe um ranking de juízes, independente de sua participação como relator ou como integrante de determinada turma recursal, conforme um comparativo entre recursos ganhos e recursos perdidos:

Análise 4-Grama - Ranking de Juízes				
Concedeu Recurso	**TF-IDF**	X	**TF-IDF**	**Negou Recurso**
Dr. Horácio Ribeiro Teixeira	5 (4/23)	X	26 (24/64)	Dr. Léo Henrique Araujo
Dra. Ana Paula Kaled Accioly	4 (4/23)	X	19 (15/62)	Dra. Ana Paula Kaled Accioly
Dr. Helder Luiz Henrique Tacaguchi	4 (4/23)	X	12 (11/62)	Dr. Moacir Antonio Dala Costa
Dr. Léo Henrique Araujo	4 (4/23)	X	7 (7/62)	Dr. Telmo Zaions Zainko

QUADRO 7 - ANALISE 4-GRAMA: RANKING DE JUÍZES
FONTE: O AUTOR (2012)

Pode-se observar que todos os juízes são citados em pelo menos 4 documentos no caso dos recursos ganhos, sendo que, por conta de 1 ocorrência a mais o Dr. Horácio Ribeiro Teixeira é considerado o Juiz que mais participou das jurisprudência acerca do tema.

A Dr. Ana Paula Kaled Accioly participou tanto dos recursos ganhos quanto dos recursos perdidos, posicionando-se no segundo lugar em ambas as colunas. Ilustrado também nas duas colunas está o Dr. Léo Henrique Araujo, sendo que na coluna de recursos perdidos ele é o que possuir maior frequência dentre os documentos.

Figurando apenas entre os recursos ganhos se encontram o Dr. Helder Luiz Henrique Tacaguchi e Dr. Horácio Ribeiro Teixeira. Já do lado dos recursos perdidos figuram com exclusividade a essa coluna o Dr. Telmo Zaions Zainko e o Dr. Moacir Antonio Dala Costa.

4.3 SÍNTESE DAS ANÁLISES

Após as análises se concluiu que dos recursos ganhos, observa-se termos que remetem ao ato da compra online, como contrato, lei, produto, entrega. J á quanto aos recursos perdidos, observam-se termos que remetem aos aspectos que lesam o consumidor, como dano, moral, indenização. Em todos os agrupamentos e métricas 1-grama ou TF se observou esse mesmo fato.

A utilidade para esse fato se dá, por exemplo, a uma defesa. Ressalta-se que não há como afirmar que o uso de termos específicos garantirá maior chance ao ganho de um processo, pois os juízes tem a premissa de serem representantes da lei, realizando suas sentenças não através de termos, mas sim através da semântica. Porém, de fato, observam-se termos preponderantes a outros conforme o resultado do recurso.

O mesmo acontece quando se analisa o Ranking de Juízes. Não se pode constatar que se um caso for presidido ou julgado por um juiz se terá maior chance de se obter um determinado resultado, pois, teoricamente, todos representam a Lei de forma idêntica. Porém, de fato, alguns juízes possuem maior padrão nas decisões que outros conforme o caso analisado.

5 CONSIDERAÇÕES FINAIS

A partir do referencial teórico, que baseia as ações do projeto, realiza-se a mineração de texto a fim de demonstrar a técnica aplicada.

A recuperação e gestão da informação orienta o projeto de modo a garantir que as etapas sejam fundamentadas em vieses acadêmicos e teóricos, permitindo acurácia nos resultados a serem analisados.

O meio escolhido, de processos jurídicos do estado do Paraná baseados em comércio eletrônico, é um dos ambientes que não possuem estudos aprofundados se não os dados puramente estatísticos.

As escolhas das jurisprudências do tema recuperaram em sua totalidade recursos inominados, o seja, recursos dado uma determinada sentença em primeira instância. Assim, os documentos da analise compõe em sua totalidade recursos propostos pela empresa de e-commerce, na tentativa de reverter uma sentença em primeira instância favorável ao cliente.

Analisando a base foi percebido que os recursos são acionados pela prestadora do serviço de e-commerce, em sua maioria na tentativa de anular ou amenizar os efeitos da decisão do juiz de primeira instância.

Assim, com o propósito de orientar a mineração de texto, dividiu-se a base de documentos entre recursos negados, que favorecem o cliente, e recursos providos, que favorecem a empresa prestadora do serviço de comércio eletrônico. Dividiu-se ainda por categorias de termo de busca, para que se fosse possível a realização de comparativos entre os termos de busca encontrados.

A primeira análise realizada, utilizando um contador de termos em formato de nuvem de *tags* pode prover análises simples analisando os termos mais recorrentes.

A partir de então, por meio da conversão dos arquivos baixados do portal da justiça, antes em pdf e depois em txt, fora possibilitada a análise dos documentos por meio da ferramenta de mineração de texto PreText

O objetivo principal foi atingido a partir da descoberta de padrões na base do presente trabalho através da análise associativa sob o processamento dos documentos por meio do software PreText e pela contagem de termos utilizando a nuvem de *tag.*

Por meio da utilização de métricas de valoração de n-grama geradas por frequência de corte tfidf se observou os termos que possuem mais peso dentro da base de documentos, e a partir deles se pode refletir conforme associações entre significado do termo e temática da base.

O processo fora descrito de forma detalhada com base nas etapas de Descoberta de Conhecimento em Bases de Dados (KDD); desde a coleta, passando pelo processamento até o resultado, desse modo se pode chegar a um resultado satisfatório da pesquisa, além de agregar valor científico à mesma.

Termos como “Dano Moral” ou “Indenização” aparecem em posições diferentes conforme o agrupamento realizado, possibilitando, por meio de análises associativas, a extração do conhecimento.

Utilizando-se a metodologia proposta, o KDD, pode-se realizar de forma eficiente a mineração de texto, mesmo que se pondere o risco de não obter quaisquer descobertas relevantes, sempre haverá a etapa de análise, seja essa associativa ou baseada em um sistema, dentre outros. Tendo o descrito, uma vez realizada a mineração de texto, de fato, se está realizando a gestão da informação.

Especificamente para a descoberta em bases do âmbito judiciário paranaense, pode-se observar a necessidade da definição de grupos para a comparação entre os processamentos de mineração de texto. O fato se deu, pois documentos jurídicos são pouco voláteis quanto a sua terminologia, ou seja, os termos utilizados em documento são em sua maioria, encontrados em qualquer outro documento seja qual for à área jurídica.

De certa forma o fato é positivo tanto para o poder judiciário quanto para a mineração de texto. Para o poder judiciário a padronização terminológica é interessante e necessária para facilitar a classificação e recuperação dos documentos, mesmo que essa padronização não seja planejada para isso.

Para estudos futuros, demonstra-se que é possível realizar a mineração de texto em bases jurídicas de outros temas. A mineração de texto é uma importante ferramenta de apoio à análise e tomada de decisão. Seguindo as etapas do KDD, o projeto poderá servir como um exemplificador ou guia para futuros projetos que também envolvam a gestão de dados e textos.

A exploração de outras matérias jurídicas pode não somente expandir as descobertas, como gerar um banco de dados que unam as características de cada temática. Denota-se também que a utilização de outras técnicas de Mineração de

Textos aliada à mostrada no presente trabalho, como a sumarização, pode complementar a pesquisa, de modo a agregar insumos a análise.

REFERÊNCIAS

ALMEIDA, L. G. P. de. **Análise de algoritmos de agrupamento para base de dados textuais.** 161 p. Tese (Mestrado) - Laboratório Nacional de Computação Científica, Petrópolis, 2007.

ALON, N.; SPENCER, J. H. **The probabilistic method.** 2. ed. *[S.l.]:* Wiley, 2000. 301 p.

BAEZA YATES, R; RIBEIRO NETO, B. **Modern information retrieval.** *[S.l.]***:** Addison-Wesley, 1999.

BEIJERSE, R.P. Question in knowledge management: defining and conceptualizing a phenomenon. **Journal of Knowledge Management**, *[S.l]:* 1999.

BRASPAG. Disponível em: < http://www.braspag.com.br>. Acesso em: 10/01/2013.

BRASIL. República Federativa do Brasil. **Portal Brasil.** Disponível em: <http://www.brasil.gov.br>. Acesso em: 12 set. 2012.

CARDOSO, O. N. P. **Recuperação da informação.** UFLA – Universidade Federal de Lavras, Lavras, 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf>. Acesso em: 25 maio 2012.

CARVALHO, G. M. R.; TAVARES, M. S. **Informação & conhecimento**: uma abordagem organizacional. Rio de Janeiro: Qualitymark, 2001.

CRUZ, E.M. K; SEGATTO, A.P. Processos de comunicação em cooperação tecnológica universidade-empresa**:** um estudo de caso em universidades federais do Paraná. **Revista de Administração Contemporânea,** *[S.l],* v.13, n.3, 2009.

FAYYAD, U. *et al.* From Data Mining to knowledge discovery in databases. **AI Magazine**, Rhode Island, p.1-54, 1996.

FERNEDA, E. **Recuperação de informação: a**nálise sobre a contribuição da ciência da computação para a ciência da informação. 147 p. Tese (Doutorado) - Universidade de São Paulo, São Paulo, 2003.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2008.

GRUPTA, V.; LEHAL, G. S. A Survey of Text Mining Techniques and Applications. **Journal Of Emerging Technologies In Web Intelligence,** Chandigarh, p. 60-76. ago. 2009.

HAN, J. *et al.* **Data mining:** concepts and techniques. *[S.l]:* Hardcover, 2007.

HIEMSTRA, D. **Using language models for information retrieval.** Twente: Taaluitgeverij Neslia Paniculata, 2001.

HAYKIN, S. S. **Redes neurais:** princípio e prática. 2. ed. *[S.l]:* Bookman, 2001. 900 p.

JONES, S.; WILLETT, P. **Readings in information retrieval.** San Francisco: Morgan Kaufmann Publishers Inc, 1997.

LABORATORY OF COMPUTATIONAL INTELLIGENCE. Disponível em: <http://labic.icmc.usp.br/>. Acesso em: 25 maio 2012.

LE COADIC, Y. F. **A ciência da informação.** Brasília: Briquet de Lemos/Livros, 1996.

LESK, M. **The seven ages of information retrieval.** International Federation Of Library Associations And Institutions, 1995. Disponível em: <http://archive.ifla.org/VI/5/op/udtop5/udtop5.htm>. Acesso em: 25 maio 2012.

MANNING, C. D. *et al.* **Introduction to information retrieval.** Cambridge: Cambridge University Press, 2009.

MCGEE, J.; PRUSAK, J. **Gerenciamento estratégico da informação**: aumente a competitividade e a eficiência de sua empresa utilizando a informação como uma ferramenta estratégica. Rio de Janeiro: Campus, 1994.

MOODY, D.; WALSH, P. Measuring the value of information: an asset valuation approach. **European Conference on Information Sy*stem*s**, Copenhagen, p.9-9, 1999.
Disponível em: <http://wwwinfo.deis.unical.it/~zumpano/2004-2005/PSI/lezione2/ValueOfInformation.pdf>. Acesso em: 12 set. 2012.

MOOERS, C. (1951). Zatocoding applied to mechanical organization of knowledge. **American Documentation**, *[S.l],* v.2, n.1, p.20-32.

NISO STANDARDS. **Ansi/niso z39.19 - guidelines for the construction, format, and management of monolingual controlled vocabularies**. Bethesda, 2005. 172 p. Disponível em: <http://www.niso.org >. Acesso em: 10 out. 2012.

NONAKA, I.; TAKEUCHI, H. **The knowledge-creating company:** how Japanese companies create the dynamics of innovation. New York: Oxford University Press, 1995. 284 p.

ORENGO, V. M.; HUYCK, C. R. A. ***Stem*ming algorithm for the portuguese languague.** In: 8th NTERNATIONAL SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL (SPIRE). 2001, Laguna de San Raphael, Chile, p. 183-193.

PACHECO, M. A. C. **Algoritmos genéticos:** princípios e aplicações. – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 1999. Disponível em: <http://www.ica.ele.puc-rio.br/downloads/38/ce-apostila-comp-evol.pdf>. Acesso em: 25 maio 2012.

REZENDE, S. O. ; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Si*stem*as de Informação da FSMA**, *[S.l],* v. 7, 2011.

SARACEVIC, T. Information Science. **Journal of the American society for information science,** *[S.l],* v.50, n.12, 1999.

SETZER, V. W. Dado, Informação, Conhecimento e Competência. **DataGramaZero – Revista de Ciência da Informação**, *[S.l],* n. zero, dez. 1999.

SHAW, I.S; SIMÕES, M.G. **Controle e modelagem fuzzy**. São Paulo: Edgard Blucher, 1999.

SILVA, A. C. M. da. **Um estudo comparativo de ferramentas de descoberta de conhecimento em texto:** a análise da Amazônia. 110 p. Tese (Mestrado) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

SILVA, H. **Comunidades de prática**: programa de pós-graduação. Curitiba, 2011. Slides.

SINGHAL, A. **Modern information retrieval**: a brief overview**.** Google, Inc., 2001. Disponível em: <http://ilps.science.uva.nl/Teaching/0405/AR/part2/ir_overview.pdf>. Acesso em: 25 maio 2012.

SOARES, M. V.; PRATI, R.C.; MONARD, M. C. PreText: A Reestruturação da Ferramenta de Pré-Processamento de Textos. **Universidade de São Paulo**, 2008.

TAGCROWD. Disponível em: <http://www.tagcrowd.com>. Acesso em: 12 maio 2012.

TAN, A. H. Text Mining: **The State of the Art and Challenges**, IN: WORKSHOP OF KNOWLEGDE DISCOVERY FROM ADVANCED DATABASES, PAKDD, *[S.l]:* 1999.

ANEXOS

ANEXO I – REGISTRO DE PROCESSAMENTO PRETEXT: CONCEDEU PROVIMENTO

```
#----------------------------#
#           PreTexT          #
#                            #
#    Implemented by LABIC    #
#----------------------------#

#======== PARAMETERS =========#
 language     = pt
 directory    = Concedeu provimento geral
 log file     = pretext.log

#----------------------------#
#           Maid.pm          #
#----------------------------#

#======== PARAMETERS =========#
 html clear   = ENABLED
 number clear = ENABLED
 simbol clear = ENABLED
 stoplist     = ENABLED
  > directory = stoplist
  > stopfile  = stoplist/port.xml
  > stopfile  = stoplist/ingl.xml
 stemming     = ENABLED
  > directory = steminfo

    ### STOP LIST ###
    Total StopWords 1024
    Total StopFiles 2

Maid            :.........:.........:.........:.........:.........: OK

#----------------------------#
#          NGram.pm          #
#----------------------------#

#======== PARAMETERS =========#
 directory    = ngraminfo
  1-Gram      = ENABLED
  4-Gram      = ENABLED
  5-Gram      = ENABLED
  7-Gram      = ENABLED
  9-Gram      = ENABLED

Criando 1Gram   :.........:.........:.........:.........:.........: OK
Criando 4Gram   :.........:.........:.........:.........:.........: OK
Criando 5Gram   :.........:.........:.........:.........:.........: OK
Criando 7Gram   :.........:.........:.........:.........:.........: OK
Criando 9Gram   :.........:.........:.........:.........:.........: OK

#----------------------------#
#          Report.pm         #
#----------------------------#

#======== PARAMETERS =========#
 NGram Dir    = ngraminfo
```

```
 Discover Dir = discover
 Graphics Dir = graphics
 Taxonomy     = taxonomia.txt
  1-Gram      = ENABLED
   > Max      =  900
   > Min      = 10
   > Measure  = tfidf
   > Smooth   = ENABLED
  4-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  5-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  7-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  9-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

#------ 1-Gram -----#
1Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max = 900                 0
 Cut with user defined restriction min = 10                  1683
 Number of 1-gram loaded from ngraminfo/1Gram.all            289

1Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/1-StdDev.dat created!
 graphics/1-StdDevFF.dat created!
 graphics/1-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK

#------ 4-Gram -----#
4Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             3146
 Number of 4-gram loaded from ngraminfo/4Gram.all            53

4Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/4-StdDev.dat created!
 graphics/4-StdDevFF.dat created!
 graphics/4-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 5-Gram -----#
5Gram.all       :.........:.........:.........:.........:.........: OK
```

```
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             2105
 Number of 5-gram loaded from ngraminfo/5Gram.all            30

5Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/5-StdDev.dat created!
 graphics/5-StdDevFF.dat created!
 graphics/5-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 7-Gram -----#
7Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             949
 Number of 7-gram loaded from ngraminfo/7Gram.all            13

7Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/7-StdDev.dat created!
 graphics/7-StdDevFF.dat created!
 graphics/7-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 9-Gram -----#
9Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             481
 Number of 9-gram loaded from ngraminfo/9Gram.all            4

9Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/9-StdDev.dat created!
 graphics/9-StdDevFF.dat created!
 graphics/9-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

======= Summary =======
   N-Gram   : 1
Total Stems : 289
Total Texts : 23
-----------------------
   N-Gram   : 4
Total Stems : 53
Total Texts : 23
-----------------------
   N-Gram   : 5
Total Stems : 30
Total Texts : 23
-----------------------
   N-Gram   : 7
Total Stems : 13
```

```
Total Texts : 23
----------------------
   N-Gram   : 9
Total Stems : 4
Total Texts : 23
----------------------

Matrix Density                                                675.24

#--- Discovery Table ---#
Discover Data   :.........:.........:.........:.........:.........: OK
Number of Texts                                               23
Discover Names  :.........:.........:.........:.........:.........: OK
Number of Stems                                               389
Total Time: 13
#-------------------#
```

ANEXO II – REGISTRO DE PROCESSAMENTO PRETEXT: NEGOU PROVIMENTO

```
#-----------------------------#
#           PreTexT           #
#                             #
#    Implemented by LABIC     #
#-----------------------------#

#======== PARAMETERS =========#
 language     = pt
 directory    = NEGOU PROVIMENTO GERAL
 log file     = pretext.log

#-----------------------------#
#           Maid.pm           #
#-----------------------------#

#======== PARAMETERS =========#
 html clear   = ENABLED
 number clear = ENABLED
 simbol clear = ENABLED
 stoplist     = ENABLED
  > directory = stoplist
  > stopfile  = stoplist/port.xml
  > stopfile  = stoplist/ingl.xml
 stemming     = ENABLED
  > directory = steminfo

    ### STOP LIST ###
    Total StopWords 1024
    Total StopFiles 2

Maid            :.........:.........:.........:.........:.........: OK

#-----------------------------#
#          NGram.pm           #
#-----------------------------#

#======== PARAMETERS =========#
 directory    = ngraminfo
  1-Gram      = ENABLED
  4-Gram      = ENABLED
  5-Gram      = ENABLED
  7-Gram      = ENABLED
  9-Gram      = ENABLED

Criando 1Gram   :.........:.........:.........:.........:.........: OK
Criando 4Gram   :.........:.........:.........:.........:.........: OK
Criando 5Gram   :.........:.........:.........:.........:.........: OK
Criando 7Gram   :.........:.........:.........:.........:.........: OK
Criando 9Gram   :.........:.........:.........:.........:.........: OK

#-----------------------------#
#          Report.pm          #
#-----------------------------#

#======== PARAMETERS =========#
 NGram Dir    = ngraminfo
```

```
 Discover Dir = discover
 Graphics Dir = graphics
 Taxonomy     = taxonomia.txt
  1-Gram      = ENABLED
   > Max      = 900
   > Min      = 10
   > Measure  = tfidf
   > Smooth   = ENABLED
  4-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  5-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  7-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  9-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

    ### Taxonomy was NOT Loaded ###

#------ 1-Gram -----#
1Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max = 900                 0
 Cut with user defined restriction min = 10                  2163
 Number of 1-gram loaded from ngraminfo/1Gram.all            669

1Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/1-StdDev.dat created!
 graphics/1-StdDevFF.dat created!
 graphics/1-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK

#------ 4-Gram -----#
4Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             6054
 Number of 4-gram loaded from ngraminfo/4Gram.all            368

4Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/4-StdDev.dat created!
 graphics/4-StdDevFF.dat created!
 graphics/4-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK
```

```
#------ 5-Gram -----#
5Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             4003
 Number of 5-gram loaded from ngraminfo/5Gram.all            201

5Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/5-StdDev.dat created!
 graphics/5-StdDevFF.dat created!
 graphics/5-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 7-Gram -----#
7Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             1694
 Number of 7-gram loaded from ngraminfo/7Gram.all            61

7Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/7-StdDev.dat created!
 graphics/7-StdDevFF.dat created!
 graphics/7-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 9-Gram -----#
9Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             830
 Number of 9-gram loaded from ngraminfo/9Gram.all            19

9Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/9-StdDev.dat created!
 graphics/9-StdDevFF.dat created!
 graphics/9-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

======= Summary =======
   N-Gram   : 1
Total Stems : 669
Total Texts : 62
-----------------------
   N-Gram   : 4
Total Stems : 368
Total Texts : 62
-----------------------
   N-Gram   : 5
Total Stems : 201
Total Texts : 62
-----------------------
```

```
   N-Gram   : 7
Total Stems : 61
Total Texts : 62
-----------------------
   N-Gram   : 9
Total Stems : 19
Total Texts : 62
-----------------------
Matrix Density                                          1163.62
#--- Discovery Table ---#
Discover Data   :.........:.........:.........:.........:.........: OK
Number of Texts                                         62
Discover Names  :.........:.........:.........:.........:.........: OK
Number of Stems                                         1318

#------------------#
Total Time: 38
#------------------#
```

ANEXO III – REGISTRO DE PROCESSAMENTO PRETEXT: COMÉRCIO ELETRÔNICO - NEGOU PROVIMENTO

```
#----------------------------#
#           PreTexT          #
#                            #
#    Implemented by LABIC    #
#----------------------------#

#======== PARAMETERS =========#
 language     = pt
 directory    = CE negou provimento
 log file     = pretext.log

#----------------------------#
#           Maid.pm          #
#----------------------------#

#======== PARAMETERS =========#
 html clear   = ENABLED
 number clear = ENABLED
 simbol clear = ENABLED
 stoplist     = ENABLED
  > directory = stoplist
  > stopfile  = stoplist/port.xml
  > stopfile  = stoplist/ingl.xml
 stemming     = ENABLED
  > directory = steminfo

    ### STOP LIST ###
    Total StopWords 1024
    Total StopFiles 2

Maid            :.........:.........:.........:.........:.........: OK

#----------------------------#
#          NGram.pm          #
#----------------------------#

#======== PARAMETERS =========#
 directory    = ngraminfo
  1-Gram      = ENABLED
  4-Gram      = ENABLED
  5-Gram      = ENABLED
  7-Gram      = ENABLED
  9-Gram      = ENABLED

Criando 1Gram   :.........:.........:.........:.........:.........: OK
Criando 4Gram   :.........:.........:.........:.........:.........: OK
Criando 5Gram   :.........:.........:.........:.........:.........: OK
Criando 7Gram   :.........:.........:.........:.........:.........: OK
Criando 9Gram   :.........:.........:.........:.........:.........: OK

#----------------------------#
#          Report.pm         #
#----------------------------#

#======== PARAMETERS =========#
```

```
 NGram Dir    = ngraminfo
 Discover Dir = discover
 Graphics Dir = graphics
 Taxonomy     = taxonomia.txt
  1-Gram      = ENABLED
   > Max      = 900
   > Min      = 10
   > Measure  = tfidf
   > Smooth   = ENABLED
  4-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  5-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  7-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED
  9-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED


    ### Taxonomy was NOT Loaded ###

#------ 1-Gram -----#
1Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max = 900                 0
 Cut with user defined restriction min = 10                  1899
 Number of 1-gram loaded from ngraminfo/1Gram.all            540

1Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/1-StdDev.dat created!
 graphics/1-StdDevFF.dat created!
 graphics/1-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK

#------ 4-Gram -----#
4Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             4922
 Number of 4-gram loaded from ngraminfo/4Gram.all            276

4Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/4-StdDev.dat created!
 graphics/4-StdDevFF.dat created!
 graphics/4-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
```

```
Linear          :.........:.........:.........:.........:.........: OK

#------ 5-Gram -----#
5Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             3343
 Number of 5-gram loaded from ngraminfo/5Gram.all            148

5Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/5-StdDev.dat created!
 graphics/5-StdDevFF.dat created!
 graphics/5-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 7-Gram -----#
7Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             1491
 Number of 7-gram loaded from ngraminfo/7Gram.all            51

7Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/7-StdDev.dat created!
 graphics/7-StdDevFF.dat created!
 graphics/7-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 9-Gram -----#
9Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             750
 Number of 9-gram loaded from ngraminfo/9Gram.all            18

9Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/9-StdDev.dat created!
 graphics/9-StdDevFF.dat created!
 graphics/9-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

======= Summary =======
   N-Gram   : 1
Total Stems : 540
Total Texts : 44
-----------------------
   N-Gram   : 4
Total Stems : 276
Total Texts : 44
-----------------------
   N-Gram   : 5
Total Stems : 148
```

```
Total Texts : 44
-----------------------
   N-Gram   : 7
Total Stems : 51
Total Texts : 44
-----------------------
   N-Gram   : 9
Total Stems : 18
Total Texts : 44
-----------------------

Matrix Density                                          994.71

#--- Discovery Table ---#
Discover Data   :.........:.........:.........:.........:.........: OK
Number of Texts                                         44
Discover Names  :.........:.........:.........:.........:.........: OK
Number of Stems                                         1033

#-------------------#
Total Time: 26
```

ANEXO IV – REGISTRO DE PROCESSAMENTO PRETEXT: COMÉRCIO ELETRÔNICO – CONCEDEU PROVIMENTO

```
#----------------------------#
#           PreTexT          #
#                            #
#    Implemented by LABIC    #
#----------------------------#

#======== PARAMETERS =========#
 language     = pt
 directory    = CE Concedeu provimento
 log file     = pretext.log

#----------------------------#
#           Maid.pm          #
#----------------------------#

#======== PARAMETERS =========#
 html clear   = ENABLED
 number clear = ENABLED
 simbol clear = ENABLED
 stoplist     = ENABLED
  > directory = stoplist
  > stopfile  = stoplist/port.xml
  > stopfile  = stoplist/ingl.xml
 stemming     = ENABLED
  > directory = steminfo

    ### STOP LIST ###
    Total StopWords 1024
    Total StopFiles 2

Maid            :.........:.........:.........:.........:.........: OK

#----------------------------#
#          NGram.pm          #
#----------------------------#

#======== PARAMETERS =========#
 directory    = ngraminfo
  1-Gram      = ENABLED
  4-Gram      = ENABLED
  5-Gram      = ENABLED
  7-Gram      = ENABLED
  9-Gram      = ENABLED

Criando 1Gram   :.........:.........:.........:.........:.........: OK
Criando 4Gram   :.........:.........:.........:.........:.........: OK
Criando 5Gram   :.........:.........:.........:.........:.........: OK
Criando 7Gram   :.........:.........:.........:.........:.........: OK
Criando 9Gram   :.........:.........:.........:.........:.........: OK

#----------------------------#
#          Report.pm         #
#----------------------------#

#======== PARAMETERS =========#
 NGram Dir    = ngraminfo
 Discover Dir = discover
```

```
 Graphics Dir = graphics
 Taxonomy     = taxonomia.txt
  1-Gram      = ENABLED
   > Max      = 900
   > Min      = 10
   > Measure  = tfidf
   > Smooth   = ENABLED
  4-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  5-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  7-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  9-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

#------ 1-Gram -----#
1Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max = 900                 0
 Cut with user defined restriction min = 10                  1587
 Number of 1-gram loaded from ngraminfo/1Gram.all            250

1Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/1-StdDev.dat created!
 graphics/1-StdDevFF.dat created!
 graphics/1-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK

#------ 4-Gram -----#
4Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             2920
 Number of 4-gram loaded from ngraminfo/4Gram.all            19

4Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/4-StdDev.dat created!
 graphics/4-StdDevFF.dat created!
 graphics/4-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK
```

```
#------ 5-Gram -----#
5Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             1993
 Number of 5-gram loaded from ngraminfo/5Gram.all            6

5Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/5-StdDev.dat created!
 graphics/5-StdDevFF.dat created!
 graphics/5-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 7-Gram -----#
7Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             931
 Number of 7-gram loaded from ngraminfo/7Gram.all            0

7Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/7-StdDev.dat created!
 graphics/7-StdDevFF.dat created!
 graphics/7-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 9-Gram -----#
9Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             480
 Number of 9-gram loaded from ngraminfo/9Gram.all            0

9Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/9-StdDev.dat created!
 graphics/9-StdDevFF.dat created!
 graphics/9-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

======= Summary =======
   N-Gram   : 1
Total Stems : 250
Total Texts : 20
-----------------------
   N-Gram   : 4
Total Stems : 19
Total Texts : 20
-----------------------
   N-Gram   : 5
Total Stems : 6
Total Texts : 20
-----------------------
```

```
   N-Gram   : 7
Total Stems : 0
Total Texts : 0
-----------------------
   N-Gram   : 9
Total Stems : 0
Total Texts : 0
-----------------------

Matrix Density                                              658.31

#--- Discovery Table ---#
Discover Data   :.........:.........:.........:.........:.........: OK
Number of Texts                                             20
Discover Names  :.........:.........:.........:.........:.........: OK
Number of Stems                                             275

#-------------------#
Total Time: 10
#-------------------#
```

ANEXO V – REGISTRO DE PROCESSAMENTO PRETEXT: COMPRAS PELA INTERNET

```
#----------------------------#
#           PreTexT          #
#                            #
#    Implemented by LABIC    #
#----------------------------#

#======== PARAMETERS =========#
 language     = pt
 directory    = compras pela internet
 log file     = pretext.log

#----------------------------#
#           Maid.pm          #
#----------------------------#

#======== PARAMETERS =========#
 html clear   = ENABLED
 number clear = ENABLED
 simbol clear = ENABLED
 stoplist     = ENABLED
  > directory = stoplist
  > stopfile  = stoplist/port.xml
  > stopfile  = stoplist/ingl.xml
 stemming     = ENABLED
  > directory = steminfo

    ### STOP LIST ###
    Total StopWords 1024
    Total StopFiles 2

Maid            :.........:.........:.........:.........:.........: OK

#----------------------------#
#          NGram.pm          #
#----------------------------#

#======== PARAMETERS =========#
 directory    = ngraminfo
  1-Gram      = ENABLED
  4-Gram      = ENABLED
  5-Gram      = ENABLED
  7-Gram      = ENABLED
  9-Gram      = ENABLED

Criando 1Gram   :.........:.........:.........:.........:.........: OK
Criando 4Gram   :.........:.........:.........:.........:.........: OK
Criando 5Gram   :.........:.........:.........:.........:.........: OK
Criando 7Gram   :.........:.........:.........:.........:.........: OK
Criando 9Gram   :.........:.........:.........:.........:.........: OK

#----------------------------#
#          Report.pm         #
#----------------------------#

#======== PARAMETERS =========#
 NGram Dir    = ngraminfo
 Discover Dir = discover
```

```
 Graphics Dir = graphics
 Taxonomy     = taxonomia.txt
  1-Gram      = ENABLED
   > Max      = 900
   > Min      = 10
   > Measure  = tfidf
   > Smooth   = ENABLED
  4-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  5-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  7-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  9-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

#------ 1-Gram -----#
1Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max = 900                 0
 Cut with user defined restriction min = 10                  526
 Number of 1-gram loaded from ngraminfo/1Gram.all            33

1Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/1-StdDev.dat created!
 graphics/1-StdDevFF.dat created!
 graphics/1-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK

#------ 4-Gram -----#
4Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             377
 Number of 4-gram loaded from ngraminfo/4Gram.all            1

4Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/4-StdDev.dat created!
 graphics/4-StdDevFF.dat created!
 graphics/4-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK
```

```
#------ 5-Gram -----#
5Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             242
 Number of 5-gram loaded from ngraminfo/5Gram.all            0

5Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/5-StdDev.dat created!
 graphics/5-StdDevFF.dat created!
 graphics/5-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 7-Gram -----#
7Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             105
 Number of 7-gram loaded from ngraminfo/7Gram.all            0

7Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/7-StdDev.dat created!
 graphics/7-StdDevFF.dat created!
 graphics/7-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 9-Gram -----#
9Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             56
 Number of 9-gram loaded from ngraminfo/9Gram.all            0

9Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/9-StdDev.dat created!
 graphics/9-StdDevFF.dat created!
 graphics/9-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

======= Summary =======
   N-Gram   : 1
Total Stems : 33
Total Texts : 4
-----------------------
   N-Gram   : 4
Total Stems : 1
Total Texts : 4
-----------------------
   N-Gram   : 5
Total Stems : 0
Total Texts : 0
-----------------------
```

```
   N-Gram   : 7
Total Stems : 0
Total Texts : 0
-----------------------
   N-Gram   : 9
Total Stems : 0
Total Texts : 0
-----------------------

Matrix Density                                              97.37

#--- Discovery Table ---#
Discover Data   :.........:.........:.........:.........:.........: OK
Number of Texts                                             4
Discover Names  :.........:.........:.........:.........:.........: OK
Number of Stems                                             34

#-------------------#
Total Time: 3
```

ANEXO VI – REGISTRO DE PROCESSAMENTO PRETEXT: E-COMMERCE – NEGOU PROVIMENTO

```
#-----------------------------#
#           PreTexT           #
#                             #
#    Implemented by LABIC     #
#-----------------------------#

#======== PARAMETERS =========#
 language     = pt
 directory    = EC-nao concedeu
 log file     = pretext.log

#-----------------------------#
#           Maid.pm           #
#-----------------------------#

#======== PARAMETERS =========#
 html clear   = ENABLED
 number clear = ENABLED
 simbol clear = ENABLED
 stoplist     = ENABLED
  > directory = stoplist
  > stopfile  = stoplist/port.xml
  > stopfile  = stoplist/ingl.xml
 stemming     = ENABLED
  > directory = steminfo

    ### STOP LIST ###
    Total StopWords 1024
    Total StopFiles 2

Maid            :.........:.........:.........:.........:.........: OK

#-----------------------------#
#          NGram.pm           #
#-----------------------------#

#======== PARAMETERS =========#
 directory    = ngraminfo
  1-Gram      = ENABLED
  4-Gram      = ENABLED
  5-Gram      = ENABLED
  7-Gram      = ENABLED
  9-Gram      = ENABLED

Criando 1Gram   :.........:.........:.........:.........:.........: OK
Criando 4Gram   :.........:.........:.........:.........:.........: OK
Criando 5Gram   :.........:.........:.........:.........:.........: OK
Criando 7Gram   :.........:.........:.........:.........:.........: OK
Criando 9Gram   :.........:.........:.........:.........:.........: OK

#-----------------------------#
#          Report.pm          #
#-----------------------------#

#======== PARAMETERS =========#
 NGram Dir    = ngraminfo
 Discover Dir = discover
```

```
 Graphics Dir = graphics
 Taxonomy     = taxonomia.txt
  1-Gram      = ENABLED
   > Max      = 900
   > Min      = 10

  4-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  5-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  7-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  9-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

#------ 1-Gram -----#
1Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max = 900                 0
 Cut with user defined restriction min = 10                  1157
 Number of 1-gram loaded from ngraminfo/1Gram.all            149

1Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/1-StdDev.dat created!
 graphics/1-StdDevFF.dat created!
 graphics/1-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK

#------ 4-Gram -----#
4Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             1080
 Number of 4-gram loaded from ngraminfo/4Gram.all            31

4Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/4-StdDev.dat created!
 graphics/4-StdDevFF.dat created!
 graphics/4-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 5-Gram -----#
```

```
5Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             600
 Number of 5-gram loaded from ngraminfo/5Gram.all            18

5Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/5-StdDev.dat created!
 graphics/5-StdDevFF.dat created!
 graphics/5-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 7-Gram -----#
7Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             145
 Number of 7-gram loaded from ngraminfo/7Gram.all            6

7Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/7-StdDev.dat created!
 graphics/7-StdDevFF.dat created!
 graphics/7-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 9-Gram -----#
9Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             36
 Number of 9-gram loaded from ngraminfo/9Gram.all            1

9Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/9-StdDev.dat created!
 graphics/9-StdDevFF.dat created!
 graphics/9-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

======= Summary =======
   N-Gram   : 1
Total Stems : 149
Total Texts : 10
-----------------------
   N-Gram   : 4
Total Stems : 31
Total Texts : 10
-----------------------
   N-Gram   : 5
Total Stems : 18
Total Texts : 10
-----------------------
   N-Gram   : 7
```

```
Total Stems : 6
Total Texts : 10
-----------------------
   N-Gram   : 9
Total Stems : 1
Total Texts : 10
-----------------------

Matrix Density                                              455.35

#--- Discovery Table ---#
Discover Data   :.........:.........:.........:.........:.........: OK
Number of Texts                                             10
Discover Names  :.........:.........:.........:.........:.........: OK
Number of Stems                                             205

#-------------------#
Total Time: 6
#-------------------#
```

ANEXO VII – REGISTRO DE PROCESSAMENTO PRETEXT: E-COMMERCE – CONCEDEU PROVIMENTO

```
#----------------------------#
#           PreTexT          #
#                            #
#    Implemented by LABIC    #
#----------------------------#

#======== PARAMETERS =========#
 language     = pt
 directory    = EC-concedeu
 log file     = pretext.log

#----------------------------#
#           Maid.pm          #
#----------------------------#

#======== PARAMETERS =========#
 html clear   = ENABLED
 number clear = ENABLED
 simbol clear = ENABLED
 stoplist     = ENABLED
  > directory = stoplist
  > stopfile  = stoplist/port.xml
  > stopfile  = stoplist/ingl.xml
 stemming     = ENABLED
  > directory = steminfo

    ### STOP LIST ###
    Total StopWords 1024
    Total StopFiles 2

Maid            :.........:.........:.........:.........:.........: OK

#----------------------------#
#          NGram.pm          #
#----------------------------#

#======== PARAMETERS =========#
 directory    = ngraminfo
  1-Gram      = ENABLED
  4-Gram      = ENABLED
  5-Gram      = ENABLED
  7-Gram      = ENABLED
  9-Gram      = ENABLED

Criando 1Gram   :.........:.........:.........:.........:.........: OK
Criando 4Gram   :.........:.........:.........:.........:.........: OK
Criando 5Gram   :.........:.........:.........:.........:.........: OK
Criando 7Gram   :.........:.........:.........:.........:.........: OK
Criando 9Gram   :.........:.........:.........:.........:.........: OK

#----------------------------#
#          Report.pm         #
#----------------------------#

#======== PARAMETERS =========#
```

```
 NGram Dir    = ngraminfo
 Discover Dir = discover
 Graphics Dir = graphics
 Taxonomy     = taxonomia.txt
  1-Gram      = ENABLED
   > Max      = 900
   > Min      = 10
   > Measure  = tfidf
   > Smooth   = ENABLED
  4-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  5-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  7-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

  9-Gram      = ENABLED
   > Max Files= 900
   > Min Files= 3
   > Measure  = tfidf
   > Smooth   = ENABLED

#------ 1-Gram -----#
1Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max = 900                 0
 Cut with user defined restriction min = 10                  611
 Number of 1-gram loaded from ngraminfo/1Gram.all            28

1Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/1-StdDev.dat created!
 graphics/1-StdDevFF.dat created!
 graphics/1-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK

#------ 4-Gram -----#
4Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             469
 Number of 4-gram loaded from ngraminfo/4Gram.all            0

4Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/4-StdDev.dat created!
 graphics/4-StdDevFF.dat created!
 graphics/4-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
```

```
Linear          :.........:.........:.........:.........:.........: OK

#------ 5-Gram -----#
5Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             306
 Number of 5-gram loaded from ngraminfo/5Gram.all            0

5Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/5-StdDev.dat created!
 graphics/5-StdDevFF.dat created!
 graphics/5-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 7-Gram -----#
7Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             153
 Number of 7-gram loaded from ngraminfo/7Gram.all            0

7Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/7-StdDev.dat created!
 graphics/7-StdDevFF.dat created!
 graphics/7-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

#------ 9-Gram -----#
9Gram.all       :.........:.........:.........:.........:.........: OK
 Cut with user defined restriction max files = 900           0
 Cut with user defined restriction min files = 3             93
 Number of 9-gram loaded from ngraminfo/9Gram.all            0

9Gram.txt       :.........:.........:.........:.........:.........: OK
 graphics/9-StdDev.dat created!
 graphics/9-StdDevFF.dat created!
 graphics/9-GnuPlot.script created!

Loading TF-IDF
Writing Measure :.........:.........:.........:.........:.........: OK
Loading Linear Normalize by column
Linear          :.........:.........:.........:.........:.........: OK

======= Summary =======
   N-Gram   : 1
Total Stems : 28
Total Texts : 3
-----------------------
   N-Gram   : 4
Total Stems : 0
Total Texts : 0
-----------------------
   N-Gram   : 5
Total Stems : 0
```

```
Total Texts : 0
-----------------------
   N-Gram   : 7
Total Stems : 0
Total Texts : 0
-----------------------
   N-Gram   : 9
Total Stems : 0
Total Texts : 0
-----------------------

Matrix Density                                          74.19

#--- Discovery Table ---#
Discover Data   :.........:.........:.........:.........:.........: OK
Number of Texts                                         3
Discover Names  :.........:.........:.........:.........:.........: OK
Number of Stems                                         28

#-------------------#
Total Time: 3
```

APÊNDICES

APÊNDICE I – STOPWORDS EM PORTUGUES

a
abaixo
acaso
acerca
acima
acola
ademais
adentro
adiante
afinal
afora
agora
agorinha
ai
ainda
alem
algo
alguem
algum
alguma
algumas
alguns
ali
alias
amiude
ante
antes
ao
aonde
aos
apenas
apesar
apos
apud
aquela
aquelas
aquele
aqueles
aqui
aquilo
as
assim
ate
atras
atraves
basicamente
bastante
bastantes
bem
bom
ca
cada
cade
caso
certa
certamente
certas
certeiramente
certo
certos
chez
com
comigo
como
comumente
conforme
confronte
conosco
conquanto
consequentemente
consigo
consoante
contanto
contigo
contra
contudo
convosco
cuja
cujas
cujo
cujos
da
dai
dali
dantes
daquela
daquelas
daquele
daqueles
daqui
daquilo
das
de
debaixo
defronte
dela
delas
dele
deles
demais
dentre
dentro
depois
desde
dessa
dessas
desse
desses
desta
destas
deste
destes
detras
deveras
diante
disso
disto
diversos
do
donde
doravante
dos
dum
duma
dumas
duns
durante
e
eis
ela
elas
ele
eles
em
embaixo
embora
enfim
enquanto
entanto
entao
entre
entretanto
exceto
essa
essas
esse
esses
esta
estas
este
estes
eu
exatamente
exceto
felizmente
frequentemente
fora
graCas
hoje
ibidem
idem
in
inclusive
inda
infelizmente
inicialmente
isso
isto
ja
jamais
la
largamente
lha
lhas
lhe
lhos
logo
mais
mal
malgrado
mas
me
mediante
melhor
menos
meramente
mesma
mesmas
mesmo
mesmos
meu
meus
mim
minha
minhas
mui
muita
muitas
muitissimo
muito
muitos
mutuamente
na
nada
nadinha
nalgum
nalguma
nalgumas
nalguns
naquela
naquelas
naquele
naqueles
naquilo
nao
nas
nela
nelas
nele
neles
nem
nenhum
nenhuma
nessa
nessas
nesse
nesses
nesta

nestas
neste
nestes
ninguem
nisso
nisto
no
nos
nossa
nossas
nosso
nossos
noutra
noutras
noutro
noutros
novamente
num
numa
numas
nunca
nunquinha
nuns
o
onde
ontem
ora
os
ou
outra
outras
outrem
outro
outrora
outros
outrossim
para
pela
pelas
pelo
pelos
per
perante
pero
pois
por
porem
porquanto
porque
portanto
porventura
possivelmente
posteriormente
posto
pouca
poucas
pouco
poucos
pra
praquela
praquelas
praquele
praqueles
praquilo
pras
praticamente
prela
prelas
prele
preles
preste
prestes
previamente
primeiramente
principalmente
priori
pro
pros
pronto
propria
proprias
proprio
proximo
qual
qualquer
quais
quaisquer
quando
quanta
quantas
quanto
quantos
quao
quase
que
quem
quer
quiCa
raramente
realmente
recentemente
salvante
salvo
se
segundo
seguramente
seja
sem
sempre
senao
sequer
seu
seus
sim
simplesmente
so
sob
sobre
sobremaneira
sobremodo
sobretudo
somente
sua
suas
tal
tais
talvez
tambem
tampouco
tanta
tantas
tanto
tantos
tao
tao-so
tao-somente
te
teu
teus
ti
tirante
toda
todas
todavia
todo
todos
tras
tu
tua
tuas
tudo
um
uma
umas
uns
varias
varios
versus
vezes
via
vice-versa
visto
voce
voces
vos
vossa
vossos
vulgo
cadern
um
dois
duas
tres
quatro
cinco
seis
sete
oito
nove
dez
onze
doze
treze
quatorze
quinze
dezesseis
dezessete
dezoito
dezenove
vinte
trinta
quarenta
cinquenta
sessenta
setenta
oitenta
noventa
cem
duzentos
trezentos
quatrocentos
quinhentos
seiscentos
setecentos
oitocentos
novecentos
mil
milhao
bilhao
trilhao
pag
recurso
inominado
oriundo
juizado
especial
comarca
registrada
regiao
metropolitana
curitiba
recorrente
ltda

APÊNDICE I – STOPWORDS EM INGLÊS

a
able
about
above
according
accordingly
across
actually
after
afterwards
again
against
all
allow
allows
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
appreciate
appropriate
are
around
as
aside
ask
asking
associated
at
available
away
awfully
b
be
became
because
become
becomes
becoming
been
before
beforehand
behind
being
believe
below
beside
besides
best
better
between
beyond
both
brief
but
by
c
came
can
cannot
cant
cause
causes
certain
certainly
changes
clearly
co
com
come
comes
concerning
consequently
consider
considering
contain
containing
contains
corresponding
could
course
currently
d
definitely
described
despite
did
different
do
does
doing
done
down
downwards
during
e
each
edu
eg
eight
either
else
elsewhere
enough
entirely
especially
et
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
exactly
example
except
f
far
few
fifth
first
five
followed
following
follows
for
former
formerly
forth
four
from
further
furthermore
g
get
gets
getting
given
gives
go
goes
going
gone
got
gotten
greetings
h
had
happens
hardly
has
have
having
he
hello
help
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
hi
him
himself
his
hither
hopefully
how
howbeit
however
i
ie
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into

inward
is
it
its
itself
j
just
k
keep
keeps
kept
know
knows
known
l
last
lately
later
latter
latterly
least
less
lest
let
like
liked
likely
little
look
looking
looks
ltd
m
mainly
many
may
maybe
me
mean
meanwhile
merely
might
more
moreover
most
mostly
much
must
my
myself
n
name
namely
nd
near
nearly
necessary
need
needs
neither
never
nevertheless
new
next
nine
no
nobody
non
none
noone
nor
normally
not
nothing
novel
now
nowhere
o
obviously
of
off
often
oh
ok
okay
old
on
once
one
ones
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
own
p
particular
particularly
per
perhaps
placed
please
plus
possible
presumably
probably
provides
q
que
quite
qv
r
rather
rd
re
really
reasonably
regarding
regardless
regards
relatively
respectively
right
s
said
same
saw
say
saying
says
second
secondly
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sensible
sent
serious
seriously
seven
several
shall
she
should
since
six
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
take
taken
tell
tends
th
than
thank
thanks
thanx
that
thats
the
their
theirs
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
theres
thereupon
these
they
think
third
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
to
together

too
took
toward
towards
tried
tries
truly
try
trying
twice
two
u
un
under
unfortunately
unless
unlikely
until
unto
up
upon
us
use
used
useful
uses
using
usually
uucp
v
value
various
very
via
viz
vs
w
want
wants
was
way
we
welcome
well
went
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
willing
wish
with
within
without
wonder
would
would
x
y
yes
yet
you
your
yours
yourself
yourselves
z
zero
about
all
am
an
and
are
as
at
be
been
but
by
can
cannot
did
do
does
doing
done
for
from
had
has
have
having
if
in
is
it
its
of
on
that
the
they
these
this
those
to
too
want
wants
was
what
which
will
with
would
page