CENTRO UNIVERSITÁRIO FEEVALE

LUCAS BRANDT

EXTRAÇÃO DE INFORMAÇÕES EM BLOGS

Novo Hamburgo
2009

LUCAS BRANDT

EXTRAÇÃO DE INFORMAÇÕES EM BLOGS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação pelo Centro Universitário Feevale.

Orientador: Rodrigo Rafael Villarreal Goulart

Novo Hamburgo
2009

AGRADECIMENTOS

Agradeço a Deus e a todas as pessoas que me guiaram e ajudaram tornar possível a realização deste grande sonho.

Aos meus pais que são grandes pessoas, sempre dispostas a realizarem meus sonhos.

Ao meu professor Rodrigo Goulart, pela confiança a mim depositado pela oportunidade e paciência para o desenvolvimento deste trabalho.

LUCAS BRANDT

Trabalho de Conclusão do Curso de Sistemas de Informação, com título Extração de Informações em Blogs, submetido ao corpo docente do Centro Universitário Feevale, como requisito necessário para obtenção do Grau de Bacharel em Sistemas de Informação.

Aprovado Por:

_____
Rodrigo Rafael Villarreal Goulart

_____
José Garibaldi de Carvalho

_____
Paulo Roberto Pasqualotti

RESUMO

Esse trabalho apresenta a necessidade atual do ser humano em comunicar-se e estar informado. Atualmente a comunicação é realizada de diversas maneiras, entretanto a que mais se destaca é a comunicação eletrônica, que ocorre através de emails, mensagens instantâneas, blogs entre outros. Este meio de comunicação, utilizado por número cada vez maior de pessoas, faz com que sejam geradas inúmeras quantidades de informações não estruturadas. Tal cenário possibilita o desenvolvimento de diversas ferramentas para estruturar e organizar estes dados. Sendo que este trabalho propõe uma ferramenta capaz de extrair informações em blogs, devido que a mesma atualmente é utilizada não somente como entretenimento, mas como ferramenta de apoio ao ensino e disseminação do conhecimento. Para idealizar o desenvolvimento desta ferramenta, será necessária a utilização de técnicas de PLN e extração de informações. Sendo assim, o presente trabalho descreve a metodologia que será empregada de ambas as áreas, bem como propõem o desenvolvimento de um sistema para a extração de informações em blogs indexados pela API Technorati.

Palavras-Chave: PLN, Extração, Informação, Comunicação, Blog.

ABSTRACT

This paperwork presents the need of the human being to communicate and be informed about everything around. Nowadays the communication is done in many ways, however the electronic communication stands out from the others, this communication occurs through emails, messengers, blogs and others. This way of being in touch, used by increasing numbers of people, generates a huge amount of unstructured information. This scenario enables the development of different tools to structure and organize data. Base on it, this paperwork proposes a tool capable to extract information from blogs and it is not only used as entertainment but also as educational support and dissemination of knowledge. In order to envision the development of this tool will be necessary to use NLP techniques and data mining, so this paperwork describes the methodology to be employed in both areas, and proposes the development of a system to extract information from blogs indexed for Technorati's API.

Keywords: NLP, Extraction, Information, Comunication, Blogs.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PLN	Processamento de Linguagem Natural
NLP	Natural Language Processor
MIT	Massachusetts Institute of Technology
HTML	Hyper Text Markup Language
PDF	Portable Document Format
XML	Extensible Markup Language
RSS	Rich Site Summary
EPC-P	Extrator de Palavras-Chave por freqüência de Padrões
EPC-R	Extrator de Palavras-Chave por freqüência de Radicais
POS	Part-Of-Speech
ID	Identificador Único
TF	Term-Frequency
IDF	Inverse Document Frequency

SUMÁRIO

INTRODUÇÃO

Atualmente percebe-se que estamos inseridos em um mundo altamente globalizado, cuja principal arma para a sobrevivência é a informação. A mesma é de suma importância para sobrevivermos contra a concorrência, sabermos aonde ir, estarmos informados sobre o mercado e demais área de interesse pessoal.

Entretanto não temos a informação pronta de maneira fácil em nossas mãos. O ambiente no qual vivemos está repleto de informações, estamos imersos em um mar de dados não estruturados; vivenciamos uma era onde temos uma quantidade enorme informações em nossa volta.

Com o passar dos anos e subseqüente ao desenvolvimento da sociedade, os dados e informações tem seu crescimento multiplicado em escalas extraordinárias. Então como podemos sobreviver a essa imensidão? Como filtrar esta enorme diversidade e garantir para nós somente o que é importante? Eis que surge a informática para sobrevivermos a tal desafio imposto. Apesar do surgimento da informática, para nos auxiliar a lidar com esta problemática das grandes massas de dados existentes, a mesma acabou facilitando para uma geração de um grande volume ainda maior de dados. Graças o surgimento da internet, e de sua popularização, as massas de dados foram aumentando exponencialmente, não apenas em tamanho, mas se disseminando em diversos lugares, impondo sobre nos mais um obstáculo. Este gigantesco mar que é a internet é perceptível também uma redundância de informações, muitas vezes algumas delas estão incorretas, entretanto cabe a nós sabermos escolher e identificar o que é válido.

Outro ponto importante a se destacar em nossas vidas atualmente, além da necessidade de estarmos constantemente informados, é a comunicação, assim como os dados, os meios de comunicação também vêm sofrendo grandes modificações, tanto pelo desenvolvimento tecnológico ou pelo surgimento de novas maneiras de comunicação entre as pessoas. Por exemplo, há um tempo atrás, podíamos nos comunicar apenas por carta, após um tempo passamos a nos comunicar por telefone, e hoje temos várias opções, como ligações telefônicas via internet, sites de relacionamento, mensageiros instantâneos, chats e entre outros disponíveis gratuitos ou não. Focalizando na comunicação via escrita como alguns

exemplos citados anteriormente, podemos ter uma grande fonte de dados. Este meio de comunicação tem se consagrado muito entre as pessoas, que os utilizam para finalidades múltiplas, como para a realização de reuniões de negócio, disseminação do conhecimento, entretenimento e até mesmo estudos.

Focalizando os blogs existentes na internet, os mesmos exercem um importante papel, no que se diz respeito a compartilhamento de informações. Por exemplo, os blogs que tem como tema principal a área da informática. Atualizados constantemente pelos seus autores, todas as pessoas que os acessam, acabam absorvendo o conhecimento exposto. Além disso, os blogs permitem que todos os visitantes possam participar de um determinado assunto, e em conseqüência, acabam interagindo com o autor do mesmo.

Através da diversidade de blogs existentes na internet, é notável que os mesmo armazenem uma quantidade enorme de informações. Podemos definir estas informações como semi-estruturadas, devido à grande diversidade de blogs que abordam o mesmo tema. A possibilidade de se estruturar as informações de uma forma mais organizada é possível somente com o desenvolvimento de uma ferramenta capaz de trabalhar com blogs indexados. Sendo este o principal foco deste trabalho, possível somente com a utilização das técnicas ligadas de PLN (Processamento de Linguagem Natural), técnicas de extração de informação e o indexador de blogs da Technorati.

Para o entendimento completo da ferramenta proposta por este estudo, o capítulo I é dado uma introdução referente à área de PLN, bem como a história de seu surgimento e evolução. No Capítulo II, é descrito o funcionamento de algumas ferramentas já existentes, conseqüentemente o Capítulo III aborda a arquitetura destas ferramentas. O conceito, as regras de geração de um corpus, estão descritos no Capítulo IV. O Capítulo V, aborda algumas metodologias para extração de informação, descrevendo e exemplificando alguns métodos para a extração da informação. E por final o Capítulo VI descreve a ferramenta, desde a obtenção dos dados, bem como sua arquitetura e funcionamento.

1 PLN

Desde o seu surgimento até os dias atuais o PLN vem abrangendo cada vez mais, diversas e complexas áreas do conhecimento e da informação. Com a intenção de remover algumas barreiras impostas entre a relação do ser humano com máquina, como por exemplo, a incapacidade do computador compreender a linguagem natural, isto é, a língua que falamos, a mesma vem abstraindo a necessidade de digitação de códigos complexos e muitas vezes incompreensíveis para operarmos um computador. Atualmente existem algumas ferramentas que diminuem essa complexidade; tomamos como exemplo a interface gráfica de um sistema operacional, que torna a usabilidade mais simples e eficiente ao usuário. Entretanto a apesar dessa facilidade, o computador ainda é incapaz de compreender nossa língua, a Linguagem Natural. Existem diversos estudos na própria área do PLN, que tentam acabar com esse problema, entretanto esses estudos vêm caminhando a passos curtos, devido ao desafio imposto, que é fazer o computador compreender nossa linguagem.

Neste capítulo, será exposto um breve histórico de como surgiu o PLN, a metodologia envolvida, uma breve introdução aos diferentes tipos de conhecimento lingüístico para o tratamento da linguagem natural, arquiteturas que possibilitam a interpretação e geração de línguas naturais, e ainda uma introdução de processamento de análise sintática automática.

1.1 PLN: Historia e metodologia.

Como descrito no capítulo inicial, os meios comunicação tem evoluído constantemente nas últimas décadas, isso graças ao desenvolvimento das tecnologias envolvidas, e principalmente desde a introdução do computador em nossa sociedade, no inicio dos anos 40, além de trazerem avanços substanciais na área do conhecimento científico, os mesmos possibilitaram o início e desenvolvimento de pesquisas em diversas outras áreas.

Devido a sua praticidade em auxiliar os pesquisadores nos estudos e descobertas, o computador, desde que inserido em nosso cotidiano tem modificado muito nossas vidas.

Apesar de inúmeros benefícios, no âmbito de auxílio a desenvolvimento de tarefas complexas ao ser humano, os computadores impuseram um grande paradigma à sociedade, desde seu surgimento: Como fazer com que eles possam entender as instruções por nos impostas, para serem executadas? A partir deste grande enigma que surgiram as linguagens de programação, como uma possibilidade imediata para a solução "parcial" deste problema.

Entretanto esta solução tem estado presente nos dias de hoje, embora tenha sido uma solução inicial para o problema, as mesmas vêm evoluindo constantemente. Para os envolvidos no campo de desenvolvimento de programas para computador, sua vida atualmente tem sido muito facilitada, com as linguagens de alto nível. Para a solução inicial do enigma, surgiram às linguagens de programação de baixo nível, neste cenário podemos citar a linguagem *Assembler*, sendo que a mesma exígia do programador, inúmeras e complexas linhas de instruções, no padrão da linguagem, para então executar uma pequena tarefa. Essa complexidade acabou virando uma necessidade para os programadores, devido que as linguagens de baixo nível, além de muito complexas, exigiam muito tempo de desenvolvimento dos algoritmos e muitas vezes o código era entendível somente ao autor do mesmo. Atualmente vemos inúmeras linguagens de programação que facilitam muito a vida dos desenvolvedores, como por exemplo, a linguagem Java.

Apesar desta grande evolução das "linguagens de máquina", presenciado por nós através de sua facilidade de aprendizagem e inteligibilidade, as mesmas ainda não solucionaram completamente a problemática da relação entre o homem e a máquina. Justificando-se ao fato de que para um algoritmo funcionar, independente da linguagem de programação, é necessário seguir um padrão de desenvolvimento imposto pelos fabricantes, por exemplo: Caso desenvolvermos um algoritmo e não escrevemos corretamente os comandos, o compilador simplesmente não executará o script.

Atualmente muito tem se ousado para resolver este desafio; eis que surge o PLN. Devido à preocupação em facilitar a comunicação entre o computador e o ser humano, vários estudos nesta área vem sendo realizados. Muitos pesquisadores têm ousado, em fazer com que os computadores sejam capazes de aprender a linguagem natural. Sendo que um dos primeiros

experimentos foi realizado na tradução automática de textos, outros mais ousados ainda criaram programas capazes de interagir verbalmente com o usuário, como por exemplo, os *ChatterBots*.

Hoje perante a ousadia dos pesquisadores, aliada a vastos estudos e experimentos bem ou mal sucedidos nesta área, fazem com que o PLN apresente-se como um campo de estudo bem desenvolvido, porém muito fragmentado, com muito material teórico e diversas idéias de aplicação para a mesma.

Dentre as experiências realizadas na área, muito presente e utilizado nos dias de hoje, podemos destacar a tradução automática de texto. Conforme dito em assuntos anteriores, a mesma pode ser considerada como um dos marcos iniciais na utilização do computador para a investigação de linguagens naturais.

Os primeiros estudos em PLN começaram pela década de 50, através da distribuição de uma carta escrita por Warren Weaver, que era um grande estudioso na área de criptografia de dados, sendo que esta carta era um convite as universidades e empresas para desenvolverem algo que ele chamava de "tradução mecanizada". Apesar da preocupação que a carta passava sobre o estudo nesta área, apenas nos dois anos iniciais à sua distribuição é que as universidades começaram a investir em estudos nesta área (Silva et al. 2007, p8).

A primeira reunião sobre o assunto ocorreu na MIT (Massachusetts Institute of Technology), em 1953, e a primeira demonstração ao público ocorreu dois anos depois. Esta demonstração consistia em um programa que era capaz de traduzir do Russo para o inglês, apenas 50 frases retiradas de um texto sobre química. O programa possuía um dicionário de consulta de apenas 250 palavras e seis regras para a gramática. Após a demonstração, a ferramenta foi considerada um sucesso, e muito foi investido pelos financiadores do projeto em aperfeiçoar e aumentar a capacidade da mesma (Silva et al. 2007, p8).

Os sistemas desenvolvidos e existentes nesta época, apesar de terem sido reconhecidos como um grande sucesso, estes programas apresentavam uma tradução de péssima qualidade, devido que o programa apenas listava uma série de possibilidades de tradução para cada palavra do texto, sendo que não havia nenhuma análise gramatical. Apesar

de facilitar, o texto traduzido exigia muito das pessoas que o utilizavam, devido à ausência de concordância entre as palavras(Silva et al. 2007, p10).

Esse falso sucesso do sistema teve seu primeiro crítico em 1964, um grande pesquisador sobre o tema, Bar-Hillel, devido que o mesmo incentivou constantemente a divulgação de um relatório produzido pelo Comitê Assessor de Processamento Automático das Línguas Naturais entre os pesquisadores. O relatório continha uma avaliação completamente negativa ao sistema de tradução atual, relatando que até agora não tinha se conseguido efetuar a tradução automática de textos científicos de forma alguma, pelo fato de que o empreendimento necessitava de pessoas para a pré e pós tradução dos textos(Silva et al. 2007, p15).

Este relatório foi um acontecimento catastrófico não somente ao sistema de tradução em questão, mas também a todos os estudos que se iniciavam na área de PLN, sendo que após a sua divulgação, vários dos investidores diminuíram, ou até mesmo cortaram os investimentos aos projetos(Silva et al. 2007, p17).

Após este desastre, os estudos na área se reiniciaram apenas na década de 70, através de Winograd, que desenvolveu um sistema em PLN, que consistia basicamente de um programa que movimentava um braço mecânico, através de comandos em linguagem natural, digitados por uma pessoa(Silva et al. 2007, p22).

Para promover uma demonstração resumida da evolução do PLN, o quadro 1 proposto por Silva et al. (2007) lista os principais acontecimentos, descobertas e desenvolvimento na área em cada década.

Década	**Evento**
Década de 50	Tradução Automática
Década de 60	Novas Aplicações e criação de Formalidades
Década de 70	Retomada e consolidação dos estudos em PLN
Década de 80	Sofisticação dos sistemas existentes
Década de 90	Sistemas baseados em “representações do conhecimento”

Quadro 1 - Evolução do PLN (Silva et al. 2007, p25).

2 FUNCIONAMENTO DOS SISTEMAS DE PLN

Antes de descrever a estrutura dos sistemas de PLN, esta seção apresenta como estes sistemas obtêm informações e conhecimentos para tratarem as línguas naturais.

Em PLN o material de entrada para o processamento, são os textos; sendo que para a idealização da análise, o mesmo é divido em partes menores, para então facilitar a detecção dos fenômenos da língua natural.

2.1 Estrutura Lingüística

Conforme descrito anteriormente, o processador lingüístico de um sistema de PLN recorta o texto em diversas partes, sendo que estes recortes, recebem o nome de sentenças(S), ou conhecidos por nós como frases ou orações. Como por exemplo: “O carro segue pela estrada”.

Tendo o texto recortado em sentenças, é possível descrever a sua constituição na língua natural. Através da análise estrutural da frase é possível dividi-la em entidades menores, denominado constituintes, sendo possível organizá-las hierarquicamente. Esta organização recebe o nome de estrutura interna, podendo ser representada em forma de estrutura arbórea. Na figura 1 é demonstrado o exemplo de uma estrutura arbórea.

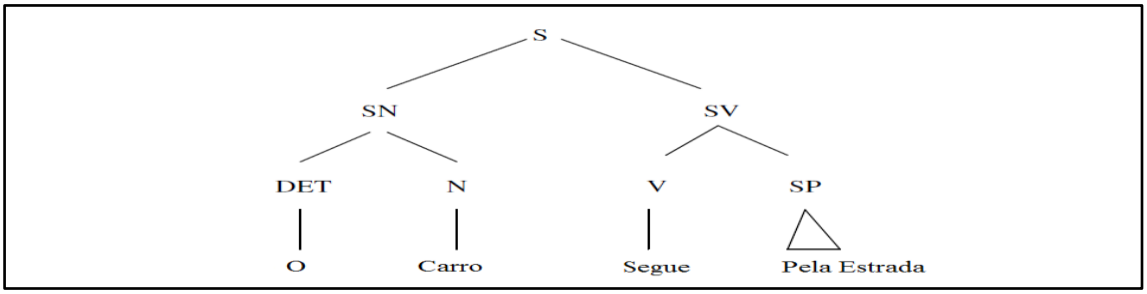

Figura 1 – Estrutura Arbórea (Silva et al. 2007, p16).

Na imagem anterior podemos notar que cada item da hierarquia da sentença foi indicado por SN, SV, SP, DET, N e V. Os elementos intermediários desta estrutura são expressões que fornecem informações necessárias para composição da sentença, denominados Sintagmas e as de nível baixo definem qual a classe que a palavra se enquadra.

Este esquema estrutural aqui exemplificado é válido para qualquer língua natural, mediante itens.

Segundo Silva et al. (2007):

> Os sintagmas são grupos de palavras organizados em torno de um núcleo sintático que o denomina. Assim, quando o núcleo de um sintagma é um nome (substantivo, adjetivo ou pronome substantivo) falamos de sintagma nominal (SN); quando é um verbo (ou locução verbal), há sintagma verbal (SV); preposição, sintagma preposicional (SP) e advérbio, sintagma adverbial (SAdv). As categorias gramaticais ou sintáticas, por sua vez, refletem as classes nas quais as palavras da língua são organizadas: determinante (DET), nome (N), verbo (V), advérbio (Adv), preposição (P) e assim por diante.

Ainda, Segundo Silva et al. (2007):

> Para a caracterização de cada um dos sintagmas e das categorias gramaticais da estrutura é necessária a compreensão da função de cada elemento na sentença. Para isso o sistema de processamento é alimentado pelos itens lexicais que carregam toda sorte de informações pertinentes para a sua operacionalização, isto é, aos itens lexicais (palavras da língua) são associadas informações de natureza fonético-fonológica, morfológica, sintática, semântica e pragmático-discursiva. A maneira como essas informações são combinadas para a disposição dos itens lexicais na sentença é dada através das diversas regras de tratamento lingüístico das quais falaremos em outras seções.

2.2 Níveis de processamento das palavras.

Conforme apresentado na seção 2.1, as palavras podem ser caracterizadas de várias maneiras, de acordo com o estatuto da descrição lingüística.

De acordo com Silva et al. (2007), podemos definir uma palavra pelo seu estatuto, devido ao fato que as palavras possuem propriedades de natureza distinta, sendo que seu comportamento é refletido quando combinadas com outras palavras, na atividade comunicativa, isto caracteriza o sentido. Abaixo segue a forma pela qual podemos definir o estatuto das palavras:

- Fonético-fonológico: O sentido sonoro que a palavra implica na sentença.
- Morfológico: Isolamento das entidades mínimas, as palavras, para compreender a formação e flexão das mesmas.
- Sintático: Compreender a função desempenhada das palavras em uma sentença.
- Semântico: Possibilita a capacidade de se identificar objetos do mundo.
- Pragmático-discursivo: quando a força expressiva das palavras possibilita além da identificação de objetos, podendo-se identificar para quem a mesma está se referenciando.

Para maioria dos sistemas de PLN cada um dos níveis acima apresentados, cuja função, possibilita descrever as palavras, constitui um modulo lingüístico, ou seja, a etapa para o processamento da língua natural. Sendo que em cada um desses módulos as informações contidas são processadas de forma a se obter o melhor tratamento lingüístico, não apenas para reconhecimento, mas também para produção de sentenças.

Entretanto Silva et al. (2007), enfatiza a necessidade das informações de que tratam cada um dos módulos, para que as mesmas sejam armazenadas juntamente às formas lingüísticas correspondentes.

3 ARQUITETURA DE UM SISTEMA DE PLN

Para o desenvolvimento de um sistema de PLN, não existe um modelo definido a se seguir, o modelo pode variar de acordo com as necessidades da aplicação(Silva et al. 2007, p23). Para exemplificar, nesta seção será demonstrada a estrutura de um sistema de tradução automática.

3.1 Arquitetura proposta a um sistema de Tradução Automática

Abaixo segue o que um sistema de tradução automática deve ser capaz de fazer:

- Reconhecer as palavras das sentenças da língua existente no texto de origem:
- Ser capaz de associar os atributos e funções sintáticas correspondentes a cada palavra da sentença, isto é, capacidade de análise sintática.
- Representar a sentença em uma forma intermediária entre a língua origem e destino, ou seja, uma língua que o sistema agregue além das palavras as suas devidas características.
- Identificar o significado geral da sentença, ou seja, através das várias palavras existentes na mesma, conseguir identificar qual o significado; capacidade de análise semântica.
- Mapear e associar o significado extraído da linguagem intermediária.
- Transformação da língua intermediária na língua de destino.

Podemos perceber que um sistema de tradução automática é bastante complexo. Simplesmente pelo fato da fase de interpretação, isto é, a passagem da língua original em uma

intermediaria, e a fase de geração, que é realizada a passagem da língua intermediária na língua destino. A maioria dos sistemas de PLN apresenta somente uma delas, um exemplo a ser dado, é o sistema de consulta a base de dados, ao fato de que na interface, o sistema apenas retorna as perguntas realizadas pelo usuário.

Enfatizando um sistema de tradução automática, Silva et al. (2007) demonstra sua arquitetura na figura 2, em seguida é fornecida uma breve explicação de cada um dos módulo que o constituem.

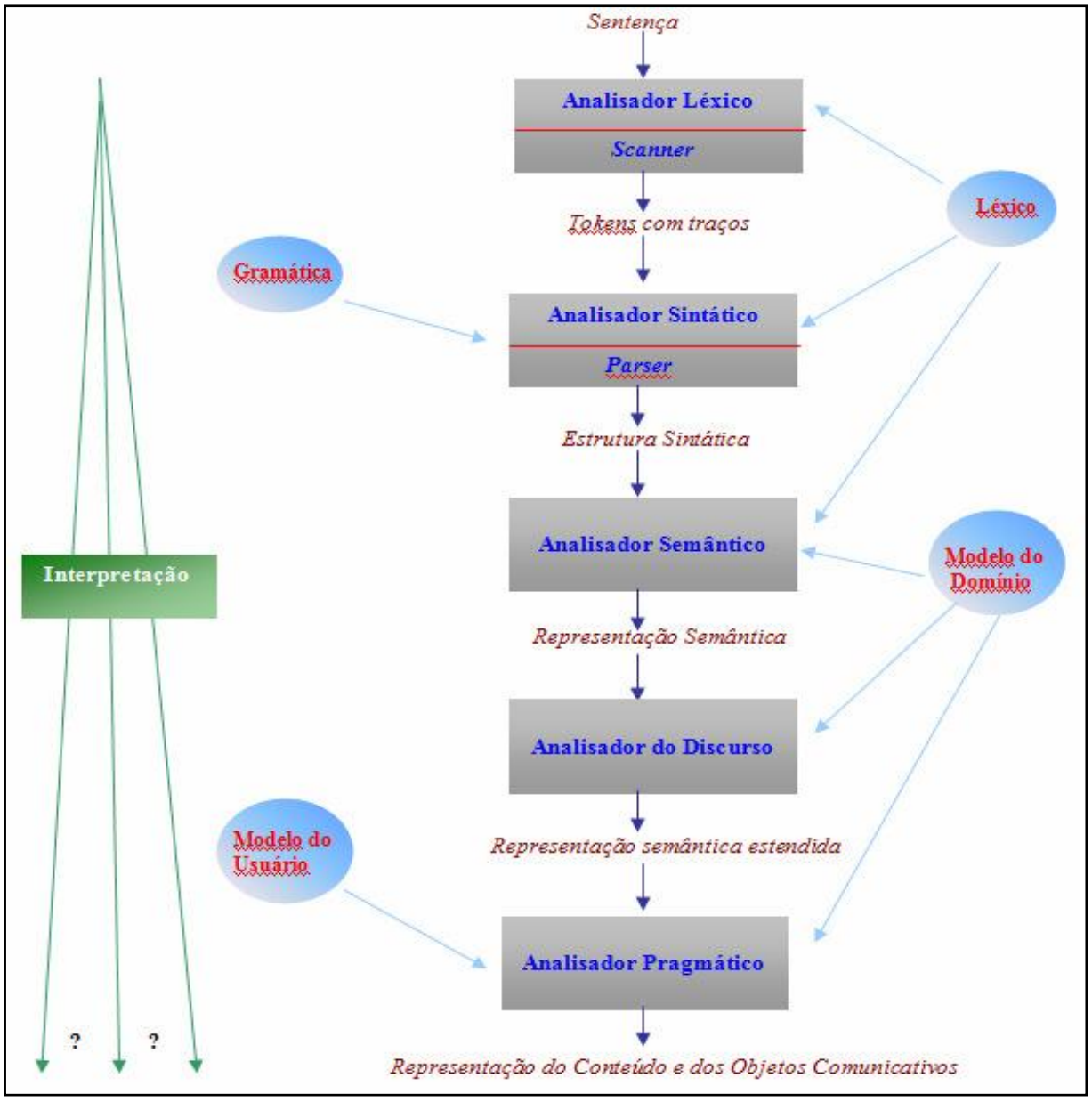

Figura 2 – Estrutura Sistema de Tradução Automática (Silva et al. 2007, p25).

- Analisador léxico (scanner): Processo comumente chamado de *tokenização*, este processo consiste em identificar e separar os componentes significativos da sentença, envolvendo desde as palavras, bem como a associação de atributos as mesmas. Este processo pode alcançar níveis de complexidade avançados, dependendo somente do resultado que se deseja do mesmo.

- Analisador Sintático (parser): Processo responsável por gerar ou recuperar a estrutura sintática de uma sentença, construído a estrutura baseando-se na gramática da língua de origem; geração da árvore sintática.

- Analisador semântico: Processo responsável pela interpretação da sentença em si ou de sentido global, perante as demais sentenças.

- Analisador de discurso: Processo responsável a dar sentido geral ao texto, isto é, faz com que o sentido do texto traduzido apresente uma maior fluidez, ao fato de que ele forma uma ligação de sentidos entre as sentenças antecedentes e procedentes.

- Analisador pragmático: Processo responsável pela identificação do contexto de uma sentença, por exemplo, o mesmo pode identificar se uma sentença interrogativa exige ou não uma resposta.

Apesar de a figura definir certa seqüência de execução dos processos, Silva et al. (2007), defende a idéia de que nem sempre os processos são executados um após o outro.

Silva et al. (2007) exemplifica:

> Considere, p.ex., a sentença "É o pote creme de molho inglês?" (exemplo extraído de Rich and Knight, 1993, p.437). durante sua análise sintática, é preciso decidir qual é o sujeito e qual é o predicado, dentre os três substantivos da sentença *(pote, creme e molho*) e dar a ela o formato "É x y?". Lexicamente, todas as seguintes delimitações da frase *pote creme de molho inglês* são possíveis: o pote, o pote creme, o pote creme de molho, o pote creme de molho inglês, creme de molho inglês, molho inglês, inglês. Entretanto, o processador sintático será incapaz de decidir quais, dentre essas formas, correspondem a estruturas sintáticas válidas, se não contar com algum modelo de mundo em que certas estruturas fazem sentido e outras não. Caso esse modelo exista no sistema automático, é possível obter-se uma estrutura que permita p.ex., a interpretação o pote de cor creme contém molho inglês, e não o pote é creme de molho inglês. Desse modo, as decisões sintáticas

> dependem da análise do discurso ou do contexto de uso e, portanto, os processos representados na Figura 4.1 interagem entre si. Não é difícil notar que a execução seqüencial dos processos de interpretação simplifica sobremaneira o projeto do sistema, se considerarmos que o resultado de uma fase constitui a entrada para a fase subseqüente. Neste caso, os processos se tornam modulares e, portanto, o controle é menos complexo. As decisões sobre a seqüencialização ou combinação dos processos dependem das características do projeto particular que se tem em mente.

3.2 Considerações Finais do Capítulo

Nesta seção foi relatado de forma simplificada e resumida a historia do PLN, bem como a uma demonstração de uma estrutura básica de um sistema de tradução automática.

Na seção 4 será abordado um dos grandes desafios impostos no desenvolvimento deste trabalho, a construção de um corpus para análise de blogs. Abordando todas as etapas necessárias para a elaboração e compilação do mesmo.

4 CORPUS

Para a realização deste trabalho, é necessário criar um corpus, para que o sistema proposto possa ser desenvolvido. Sua concepção é de suma importância para o desenvolvimento deste trabalho. Devido ao fato de que não existe nenhum corpus específico para a análise de blogs. Entretanto antes de descrever uma metodologia de criação de corpus, primeiramente o mesmo deve ser conceitualizado.

4.1 Conceito de Corpus

Segundo Almeida e Aluísio(2006, p.157) podemos definir corpus como: um conjunto finito de enunciados tomados como objeto de análise. Mais precisamente, conjunto finito de enunciados considerados característicos do tipo de língua a estudar, reunidos para servirem de base à descrição e, eventualmente, à elaboração de um modelo explicativo dessa língua. Trata-se, pois, de uma coleção de documentos quer orais (gravados ou transcritos) quer escritos, de acordo com o tipo de investigação pretendido. As dimensões do corpus variam segundo os objetivos do investigador e o volume dos enunciados considerados como característicos do fenômeno a estudar. Um corpus é chamado exaustivo quando compreende todos os enunciados característicos. Também pode ser chamado de seletivo quando compreende apenas uma parte desses enunciados.

Atualmente existem diversos corpus disponíveis para uso, na internet, entretanto nem sempre os corpus disponíveis, poderão abranger as necessidades de um projeto. Existem inúmeras questões a serem avaliadas ao se utilizar um corpus já existente, como por exemplo, a língua e o assunto que o mesmo foi compilado. Um caso mais especifico, se dá ao próprio desenvolvimento deste trabalho, que necessita de um corpus direcionado a língua portuguesa, e que possa tratar os inúmeros desafios impostos pela linguagem, como o coloquialismo e gírias, usados com freqüência em blogs.

Essa necessidade de mudanças para a concepção de corpus, segundo Almeida e Aluisio(2006, p.157) deve-se a lingüística de corpus, tida como uma abordagem que se ocupa da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador .

Podemos dizer que um conjunto de livros, revistas artigos e notícias de jornais, possam ser considerados um corpus. Porém esta idéia não é correta, devido ao fato que os dados lingüísticos contido neles, não estão em formato eletrônico, ou seja, não podem ser processado por um computador.

Tendo essa noção de o que é um corpus e como ele deve estar apresentado, Almeida e Aluísio (2006, p.157), definem 4 características fundamentais:

- Amostragem e representatividade: o corpus deve ter uma amostragem ou variedade da língua ao qual se quer analisar, suficiente para se obter um grau satisfatório de representatividade;

- Tamanho finito: o corpus deve ter uma quantidade finita, por exemplo, 500 mil palavras, entretanto o Corpus-Monitor é uma exceção, ao fato de que ao receber mais textos, o nível de amostragem aumenta.

- Formato eletrônico: O mesmo deve estar em formato eletrônico, ou seja, em arquivo eletrônico, para ser processado em maior velocidade pelo computador;

- Referencia padrão: O corpus deve conter uma referência, para que o mesmo alcance um grau de representatividade maior no assunto em que o mesmo foi compilado. Por exemplo: um corpus compilado com textos eruditos e poéticos, não terá grau de representatividade satisfatório ao se analisar textos técnicos na área da medicina;

Em relação às quatro características descritas anteriormente podemos destacar a questão do formato eletrônico, que não basta estar apenas gravado em arquivo, o mesmo deve estar compilado e arquivado, de forma que os sistemas possam utilizá-los de forma rápida, tendo fácil acesso aos dados nele contido. Neste caso é possível citar a Internet, que apesar de estar em formato eletrônico, à mesma ainda se apresenta em formato de artigos ou meramente textos, sendo assim precisariam de tratamento para tornar-se um corpus.

O formato eletrônico do corpus é um caso a parte, devido ao fato de que os anos 90, ter sido uma década de grande evolução na questão de estudos sobre PLN. Tendo uma grande evolução, principalmente no surgimento de novas ferramentas para processamento de diversas línguas.

Tal evolução no desenvolvimento das ferramentas, e principalmente no surgimento do corpus, podemos ter uma noção quase que precisa do comportamento lingüístico existentes nos textos, não necessitando de uma pré e pós avaliação da análise por pessoas. Sobretudo por meio de um corpus podemos observar informações morfológicas, semânticas, sintáticas e discursivas; que são peças fundamentais em uma pesquisa. Além disso, podemos explicar o emprego das palavras nos textos, o sentido que as mesmas exercem, além de descobrir fatos novos da língua em questão, que são de difícil intuição ao ser humano.

4.2 Questões Relevantes para o Desenvolvimento de um Corpus.

Para o desenvolvimento do corpus a ser utilizado neste projeto, faz-se necessário a observação de alguns pontos para sua compilação, para que o mesmo atinja um nível de confiabilidade necessário para sua utilização na pesquisa. Dentre os pontos a se observar podemos citar os seguintes: autenticidade, representatividade, balanceamento, amostragem, diversidade e tamanho, cujos quais terão seus detalhes descritos nos seguintes itens:

- Autenticidade: os textos que formarão o corpus devem ser textos em linguagem natural, além de que devem ser escritos por pessoas de linguagem nativa com a que se esta desenvolvendo o corpus;

- Representatividade: Os textos utilizados deverão ter uma representatividade com o assunto a que se destina a pesquisa, no caso o desenvolvimento deste trabalho, será necessário utilizar textos, os *posts* dos blogs relacionados a um determinado tema. E como já comentado anteriormente, não se tem um nível satisfatório no sistema, ao se utilizar um corpus formado por poesias e textos épicos, para pesquisa em áreas técnicas como a medicina, por exemplo.

- Balanceamento: Muito parecido com a representatividade, o balanceamento reforça a questão de se escolher o assunto que o corpus abrange, por exemplo, seria confuso misturar diversas áreas em um corpus, sendo assim poderia se confundir o sistema de PLN, no caso de avaliação de uma determinada palavra ser utilizada de certa maneira em um tema, e de um jeito completamente diferente em outro.

- Amostragem: Apesar de a construção do corpus necessitar focar um tema, o mesmo não impede que os textos sejam retirados apenas de uma fonte, as mesmas podem variar, desde livros, notícias, revistas e artigos acadêmicos.

- Diversidade: Apresenta muita semelhança com a amostragem; a diversidade reforça a idéia de se retirar os textos de meios de propagação da informação e conhecimentos, devido ao fato de enriquecer o corpus pelo uso diferenciado na colocação das palavras nos textos.

- Tamanho: O corpus que se deseja desenvolver deve ter uma quantidade adequada para a realização de pesquisa na área em que se focar. Esta quantidade não se resume apenas a quantidade de palavras ou tipos da mesma. O corpus deve contemplar uma quantidade diversificada de textos, de diversas fontes, bem como autores, tipos discursivos e datas. Dependendo do tipo de pesquisa que se deseja realizar, por exemplo, existem corpus de 100.000 palavras que atingem um ótimo nível de processamento. Entretanto

> existem outras pesquisas, que necessitam de um número maior de palavras. Atualmente o corpus de maior diversificação é o da *Bank of English*, com mais de 530 milhões de palavras.

Para Almeida e Aluisio (2006, p.157) a elaboração de um corpus é um processo que avança em ciclos: iniciado-se a partir da escolha de textos baseada em critérios externos culturalmente aceitos (tipologia de gêneros e tipos de textos, por exemplo), depois se prossegue com investigações empíricas da língua ou variedade lingüística sob análise (também denominados critérios internos) e, finalmente, procede-se com a revisão de todo o projeto.

4.3 Etapas para a Compilação de um Corpus

Para o desenvolvimento de um corpus é necessário que sejam seguidas algumas etapas, lembrando que as etapas abaixo descritas, não são necessariamente obrigatórias, ou que estão formalizadas em algum site de estudos na área de PLN, instruindo que a compilação do corpus deva seguir a metodologia apresentada nas subseções seguintes deste capítulo.

4.3.1 Seleção do texto

Para a compilação de um corpus, a primeira etapa a ser executada é a seleção de textos, sendo que esta etapa deve se ter conhecimentos em mente, as questões relevantes para o desenvolvimento de corpus, conforme descrito na seção 4.2. Tendo conhecimentos das questões anteriores, torna-se possível selecionar os textos que farão parte do corpus e passar ao próximo passo descrito na subseção 4.3.2.

4.3.2 Compilação e manipulação

Segundo Almeida e Aluisio(2006, p.160), a compilação consiste no armazenamento em arquivos pré-determinados de todos os textos selecionados.

Os textos que irão formar o corpus poderão ser provenientes de diversas partes, como por exemplo, revistas, jornais, artigos e livros; sendo que estes deverão ser digitalizados por meio de um scanner com OCR. Existe também a possibilidade de se obter textos da internet, sendo que os mesmos devem passar por programas que eliminem as TAGS em HTML (Hyper Text Markup Language).

A manipulação do corpus consiste em tratar os textos selecionados, por exemplo, textos contidos em arquivos PDF (Portable Document Format), DOC dentre outros. Os mesmos devem passar por uma transformação de sua extensão de arquivo atual para arquivos de texto puro. Tendo os textos, devemos organizá-los em arquivos, seguindo um padrão de nomeação dos mesmos, para facilitar a sua manipulação.

4.3.3 Direitos autorais

No caso da compilação de um corpus pessoal, com algum já existente, devemos ter a autorização para a utilização do mesmo, junto aos seus criadores.

4.3.4 Anotação

A anotação consiste basicamente em duas representações de informação: a anotação estrutural e a anotação lingüística.

Almeida e Aluisio (2006, p.160) definem a *anotação estrutural*:

> [...] compreende a marcação de dados externos e internos dos textos. Como dados externos entendemos a documentação do corpus na forma de um cabeçalho que inclui os metadados textuais (ou dados estruturados sobre dados), isto é, dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, a tipologia textual e informação sobre a distribuição do corpus. Como dados internos temos a anotação de segmentação do texto cru, que envolve: a) marcação da estrutura geral – capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras, e b) marcação da estrutura de subparágrafos – elementos que são de interesse lingüístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e ênfases tipográficas do tipo negrito, itálico, sublinhado, etc.

Tendo estas informações, fica fácil no momento do processamento, selecionar um trecho de um determinado autor ou data de publicação entre outras possibilidades.

Ainda, Almeida e Aluisio(2006, p.160) definem também a *anotação lingüística*:

> [...] pode ser em qualquer nível que se queira, isto é, nos níveis morfossintático, sintático, semântico, discursivo, etc., sendo inserida de três formas: manualmente (por lingüistas), automaticamente (por ferramentas de Processamento de Língua Natural – PLN) ou semi-automaticamente (correção manual da saída de outras ferramentas). Essa última é comprovadamente mais eficiente, pois revisar é mais rápido e gera dados mais corretos do que anotar pela primeira vez.

Tendo em mente o conceito e as formas de estrutura de anotação de um corpus, a subseção 4.4 demonstrará dois tipos de etiquetas morfossintáticas para a anotação de um corpus. Cabendo ao pesquisador devera optar por uma delas, para a anotação do corpus a ser gerado.

4.4 TagSets

TagSets, são um conjunto de etiquetas utilizadas no processo de anotação de um corpus, nesta seção será apresentado dois tipos de TagSets, disponíveis para a língua portuguesa.

4.4.1 NILC

Tagset proposto por AIRES, (1998) para o português, possui várias formas de classificação para as funções sintáticas, facilitando o processo de parsing, devido ao fato de fornecerem tags, para a anotação sintática. No quadro 2, algumas Tags são demonstradas.

Classificação	**Tag**
Adjetivo	ADJ
Advérbio	ADV
Artigo	ART
Número Cardinal	NC
Número Ordinal	ORD
Outros Números	NO
Substantivo Comum	N
Nome Próprio	NP
...	...

Quadro 2 – Lista de TagSet da NILC (AIRES, 1998)

4.4.2 VISL

O conjunto de *tags* do VISL, proposto por Bick, (2000), apresenta um grande nível de abrangência, proporcionando uma ótima classificação individual das palavras. Além das

tags principais há outras que possibilitam fornecer informações morfológicas sobre a palavra, tal como gênero, número, pessoa, dentre outras possibilidades. No quadro 3, estão demonstradas algumas das *tags* deste conjunto.

Classificação	TAG
$	Pontuação
Adj	Adjetivo
Adv	Advérbio
Det	determinador (artigo+pronome)
In	Interjeição
Kc	conjunção coordenativa
Ks	conjunção subordinativa
N	Substantivo
Num	Numeral
Pers	pronome pessoal
...	...

Quadro 3 – Lista de TagSet – VISL (Bick, 2000)

4.4.3 Exemplo

No processo de anotação de um corpus, cada um dos termos constituintes, recebe uma etiqueta, a mesma identifica qual o tipo que a palavra se encaixa, ou seja, através da etiqueta, é possível identificar se o termo é um verbo, artigo, preposição, entre outro tipo. Na figura 3, é exibido um exemplo da frase "Um revivalismo refrescante", anotada pelo tagset VISL.

```
Um revivalismo refrescante
A1
UTT:np
>N:art('um' <arti> M S) Um
H:n('revivalismo' M S)  revivalismo
N<:adjp
=H:adj('refrescante' M S)   refrescante
</s>
</t>
<p>
<s>
```

Figura 3 - Frase anota pelo Tagset VISL TagSet – VISL (Bick, 2000)

5 EXTRAÇÃO DA INFORMAÇÃO

Esta seção tem como objetivo, demonstrar algumas das heurísticas, que serão utilizadas para a extração de informações. As mesmas terão seus conceitos descritos nos subseções seguintes desta seção.

5.1 Método TF/IDF

Esta subseção tem por objetivo descrever o método estatístico, TF/IDF(term-frequency / inverse document frequency), sendo que esta metodologia auxilia na obtenção de resultados mais concretos, ao fato que o método, é próprio para a análise de termos em diversos documentos. Nas subseções 5.1.1 a 5.1.4, o método em questão é descrito para facilitar a compreensão de sua utilização neste trabalho.

5.1.1 Objetivo

Seu principal objetivo na ferramenta proposta, descrita na seção 6, é verificar os termos resultantes dos processos realizados sobre o corpus formado a partir dos textos obtidos nas consultas realizadas na API Technorati. Através desta análise em todos os documentos é possível verificar se o termo é relevante ou não, ao assunto cujo qual foi realizado a pesquisa. Nas subseções 5.1.2 a 5.1.4, são descrito cada um dos cálculos, para a obtenção da informação a partir dos métodos em questão.

5.1.2 TF(Term Frequency)

Através deste método é possível realizar a proporção da quantidade de ocorrências de um determinado termo em relação ao termo de maior ocorrência. Para a realização do calculo a seguinte formula é proposta por Wojciechowski et al. (2003, p.6):

$$tf^* = \frac{QtdeKeyword}{QtdeMaxKeyword}$$

Figura 3.1 – Fórmula para calculo de TF. Wojciechowski et al. (2003, p.6).

Onde:

- QtdeKeyword: é o valor da freqüência do termo em um documento.
- QtdeMaxKeyWord: é a quantidade do termo de maior freqüência em um documento.

5.1.3 IDF (Inverse document Frequency)

Este cálculo, mede a proporção inversa de um determinado termo, em relação ao seu aparecimento em outros documentos da pesquisa. Obtendo o resultado do cálculo, é possível avaliar se o termo pode ser considerado ou não como informação na pesquisa executada. Wojciechowski et al. (2003, p.6). O mesmo é dado pela seguinte formula:

$$idf = \log \frac{N}{ni}$$

Figura 3.2 – Fórmula para calculo de TF. Wojciechowski et al. (2003, p.6).

Onde:

log: é o logaritmo de base 10.

N: É a quantidade total de todos documentos.

ni: É a quantidade de documentos que o termo aparece.

5.1.4 TF/IDF

Conforme objetivo descrito na subseção 5.1.1, o cálculo de TF/IDF é dado pela multiplicação do resultado de TF, descrito na subseção 5.1.2, com o resultado de IDF descrito na subseção 5.1.3. A partir deste resultado, é possível avaliar se um determinado termo possui algum valor como informação na pesquisa realizada.

5.2 Extração de informação a partir de métodos estatísticos

As subseções 5.2.1 e 5.2.2 descrevem dois algoritmos, EPC-P(Extração baseada em freqüência de padrões) e EPC-R (Extração baseada em freqüência de Radicais) propostos por Nunes et al. (2000), para a extração de palavras-chave.

Nunes et al. (2006, p.160) defendem de que a extração de palavras chaves:

> [...]podem ser úteis em diversas aplicações computacionais, em especial aquelas que necessitam indexar documentos para buscas posteriores. [...] Técnicas extrativas baseadas na seleção de frases do texto são consideradas simples se comparadas a técnicas que incluem compreensão de texto (e, portanto, muito complexas), mas podem ser sofisticadas com algum conhecimento lingüístico e assim alcançarem índices razoáveis de eleição de palavras-chave.

5.2.1 EPC-P (Extração baseada em freqüência de padrões)

Tendo delimitado a língua e o gênero dos textos que serão utilizados na análise, a eficácia deste método se torna mais satisfatório, ao fato que temos alguns padrões pré-estabelecidos. O método pode ser chamado de EPC-P, sendo que o mesmo efetua uma busca de palavras que casam com padrões morfossintáticos. Resumindo sua funcionalidade, o método encontra todas as frases que apresentem uma ligação, com os padrões pré-definidos, tendo uma listagem das frases, são utilizados métodos estatísticos, filtrando somente a que possuírem uma maior relevância.

5.2.2 EPC-R (Extração baseada em freqüência de Radicais)

Este método também conhecido como EPC-R, apresenta uma grande semelhança ao método descrito anteriormente.

Baseado no algoritmo de Extractor proposto por Turney(1999 apud AVILA, 2006), tendo seu funcionamento é considerado simples. O mesmo utiliza somente a freqüência de radicais no texto, não importando se os mesmos se encaixam nos padrões predeterminados. Entretanto este método se diferencia, pelo uso constante de listas de *Stopwords*, para remoção das palavras irrelevantes, e também pela formação de três listas de palavras; a primeira para palavras simples, a segunda para as duplas e a terceira as triplas e juntamente a estas listas são armazenadas as freqüência que cada uma apareceu no texto.

6 A FERRAMENTA

O presente capítulo aborda a ferramenta proposta para a extração de informações em blogs, sendo que será abordado o objetivo de seu desenvolvimento, bem como e onde os dados para testes foram coletados, as heurísticas utilizadas, ambos demonstrados de formas descritivas e ilustrativas.

6.1 Objetivo:

O objetivo da ferramenta conforme mencionado no início desta sessão é a extração de informações em blogs. Descrevendo de uma maneira minuciosa, o corpus formado, a partir de consultas realizadas em um site de indexação de blogs, será processado por heurísticas de tratamento de textos e métodos estatísticos.

As heurísticas de tratamento de texto têm a função de analisar o corpus por um dicionário com base na língua a qual, foi realizada a pesquisa; para então serem extraídas as palavras mais freqüentes nos textos, mas que não estejam presentes no dicionário.

Após este primeiro método, os termos restantes, passam por uma lista de *stopwords*, para a remoção das palavras que não tenham sentido relevante para o objetivo deste trabalho, como por exemplo, palavras de classe fechada (pronomes, conjunções, artigos e preposições).

Ao final do processo de tratamento do corpus, os resultados obtidos, podem ser compostos por termos freqüentemente utilizados na Internet, como por exemplo, marcas, nomes de produto e gírias. Esta lista de termos será processada por métodos estatísticos para fim de obtermos informações do processamento do corpus construído. Dentro os métodos estatísticos, que a lista de termos resultantes da análise textual, o método TF/IDF é que definirá a relevância do termo no texto. Devido à alta eficiência no processamento de textos que este método apresenta. Todos os métodos, superficialmente descritos, terão suas características e funcionalidades, detalhadas nas subseções 6.2 à 6.6.

6.2 Coleta de textos:

Conforme mencionando no capítulo 4 subseção 4.3.1; para a compilação de um corpus é necessário a coleta de textos. Como o objetivo deste trabalho é a extração de informações em blogs, a principal fonte de textos é proveniente de uma API disponível na internet, cujas características e funcionalidades serão descritas na seção 6.4.

6.3 O assunto:

A ferramenta possibilita que se possa fazer qualquer tipo de busca, entretanto para a realização dos testes optou-se por utilizar o assunto “Smartphone”, devido à utilização do tema em vários trabalhos relacionados ao PLN desenvolvidos, inclusive, na Feevale.

6.4 A API Technorati

Technorati é um motor de buscas, com a especialidade de efetuar buscas em blogs. Criado em 2002 por David Sifry, a empresa é concorrente direta ao Yahoo e a Google, que também disponibilizam ferramentas similares. Porém a Technorati é muito forte na área, devido a gama de API gratuitas disponíveis para a realização de buscas(SIFRY, David. 2002) .

Como mencionado anteriormente, esta API de consulta na internet disponibiliza uma serie de ferramentas de pesquisas em blogs. Apesar desta diversidade, a ferramenta escolhida foi a Search, devido à possibilidade de se realizar pesquisas, nos blogs indexados pelo site, em

um determinado assunto. Nas seções 6.4.1 será descrito como a API foi utilizada para a obtenção dos textos para a compilação do corpus.

Para a utilização da ferramenta é necessário ser cadastrado no site. Em conseqüência desta exigência do site, foi necessário realizar o cadastro, para então ter acesso à mesma.

Cada consulta realizada à ferramenta será retornado um arquivo em formato XML contendo os POSTS dos blogs indexado pelo site.

6.4.1 Formação da URL:

Para efetuar uma consulta na ferramenta, é necessário compor uma URL, formada por uma série de parâmetros que especificam a necessidade da pesquisa, como por exemplo, o assunto e língua que se deseja realizar a pesquisa. Na página principal da ferramenta, cada um dos parâmetros que formam a URL, possui uma descrição detalhada de sua finalidade na consulta. A API define que alguns parâmetros são mandatórios e outros são opcionais. Tendo somente os parâmetros obrigatórios, é possível realizar as buscas, entretanto os parâmetros opcionais nos oferecem algumas facilidades a mais, como por exemplo, a quantidade de registros retornados pela busca.

6.4.1.1 Parâmetros Mandatórios:

Abaixo são listados os parâmetros mandatórios, especificado pela a API:

- Key: Para utilizar cada uma das ferramentas disponíveis é necessário informar uma chave, obtida mediante cadastro no site, conforme descrito na seção 6.4.

- Query: Neste parâmetro é informada a palavra ou as palavras "chave" que se deseja realizar a consulta. Entretanto, na API está especificado que caso no de consultas com várias palavras, as mesmas devem estar separadas pelo sinal de adição "+", conforme exemplo: "smartphone+samsumg".

Tendo especificado os parâmetros obrigatórios conforme a necessidade, já podemos compor a seguinte URL:
http://api.technorati.com/search?key=[CHAVE]&query=smartphone.

6.4.1.2 Parâmetros Opcionais:

Parâmetros opcionais são os parâmetros não obrigatórios, para a realização de uma consulta. Entretanto os mesmo oferecem a possibilidade de refinar a consulta, abaixo os mesmo estão listados e detalhadamente descritos.

- Format: Especificando este parâmetro, podemos escolher entre o XML e RSS(Rich Site Summary) como retorno da consulta. Em conseqüência este parâmetro aceita somente os valores "XML" ou "RSS".

- Language: Neste parâmetro podemos especificar o idioma em que deverá ser executada a busca. O valor a ser posto neste parâmetro, deve ter dois caracteres, conforme especificado na ISO 639-1. Nos testes efetuados com a ferramenta se utilizou o idioma Português, em conseqüência foi informado o valor “pt” ao parâmetro.

- Authority: Este parâmetro possibilita retornar somente os registros, que possuem uma quantidade variável de ligações. A própria API realiza este calculo de autoridade. Para este parâmetro deve ser definido um dos seguintes valores:

    - “n” : todos os resultados;

    - “a1”: pelo menos uma ligação;

    - “a4”: resultados com quantidade razoável de ligações;

    - “a7”: resultados com centenas de ligações;

- Start: Sendo declarado este parâmetro podemos selecionar a página da busca realizada. Devido ao fato que a API, pagine os resultados de uma busca.

- Limit: Este parâmetro possibilita limitar o número de registro a ser retornado em cada consulta, por padrão é retornado 20 registros, entretanto podemos definir no máximo 100 registros por consulta.

- Claim: Ao definir este parâmetro com o valor “1”, se existir, a API retornará alguma informação vinculada ao autor do blog.

6.4.1.3) A URL

Depois de especificados cada um dos parâmetros, conforme a necessidade é possível compor a URL para executar uma consulta. Abaixo é exibido um exemplo de URL utilizado para o desenvolvimento deste trabalho.

http://api.technorati.com/search?key=[CHAVE]&query=smartphone&format=xml&language=pt&authority=n&start=1&limit=100

6.4.2 O Corpus

Após a formação da URL, para se obter o resultado é necessário realizar uma chamada HTTP, que pode ser realizada, em qualquer navegador de internet, como por exemplo, o Mozilla Firefox. Como resultado é exibido um arquivo XML, contendo os POSTS de todos os blogs indexados pelo site, na figura 4.1 é exposto o resultado retornado pela URL composta.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- generator="Technorati API version 1.0" -->
<!DOCTYPE tapi PUBLIC "-//Technorati, Inc.//DTD TAPI 0.02//EN" "http://api.technorati.com/dtd/tapi-002.xml">
<tapi version="1.0">
<document>
  <result>
    <query>smartphone</query>
    <querycount>2725</querycount>
    <querytime>1.603</querytime>
    <rankingstart>2</rankingstart>
    <language>pt</language>
</result>
  <item>
    <weblog>
      <name>SeidiMobile</name>
      <url>http://www.seidimobile.com.br</url>
      <rssurl>http://www.seidimobile.com.br/feed/</rssurl>
      <inboundblogs>1</inboundblogs>
      <inboundlinks>1</inboundlinks>
      <lastupdate>2009-10-10 01:00:16 GMT</lastupdate>
    </weblog>
    <title>Quando um smartphone tem de ser Fashion!</title>
    <excerpt>por Seidi Mobile, em 9-outubro-2009, Ã s 8:50 pm  Lembram-se dos rumores de um smartphone Giorgio Armani a ser lanÃ§ado pela Samsung e
    <created>2009-10-09 23:50:44 GMT</created>
    <permalink>http://www.seidimobile.com.br/2009/10/09/quando-um-smartphone-tem-de-ser-fashion/</permalink>
  </item>
```

Figura 4.1 – Exemplo de arquivo com o resultado obtido em uma consulta realizada a API Technorati.

Tendo realizado a consulta junto a API, obtemos os textos necessários para a formação de um corpus. Na seção 4.2 estão definidas algumas questões relevantes para o desenvolvimento de um Corpus. O uso da API supriu todas as questões de maneira razoável. Por exemplo, a questão autenticidade é suprida pelo fato de que os textos estão em linguagem natural, ou seja, são compostos por textos escritos por pessoas. O Balanceamento, amostragem e diversidade são supridas ao fato de que todas as consultas realizadas necessitam de um tema, devido que a API exige, por meio de um parâmetro na URL, conforme descrito na seção 6.4.1, que seja informado uma ou mais "palavras-chave", para a realização da consulta. Salientando apenas a questão da amostragem, a mesma é satisfeita somente com os resultados provenientes dos blogs indexados pela API, entretanto essa carência não afeta de maneira que impeça a formação do corpus. Apesar de a amostragem vir de apenas uma fonte, a questão tamanho não foi afetada. Ao fato que esta API é de grande difusão na Internet, recebendo diariamente novos cadastros, favorecendo a quantidade de textos.

6.5 Processando o Corpus.

Conforme especificado na seção 6.4, tendo realizado a busca através da URL composta, temos como resultado um arquivo XML. Na figura 4.2, é exibida a estrutura básica da ferramenta, para uma compreensão visual de seu funcionamento. Nas seções 6.5 e 6.6, será descrito como o resultado obtido será processado pela ferramenta proposta.

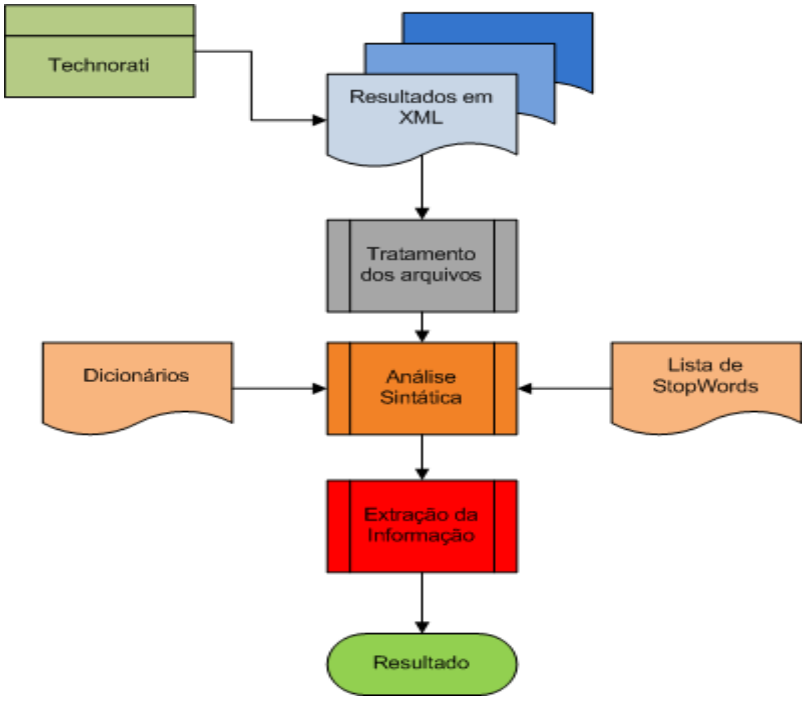

Figura 4.2 – Arquitetura da ferramenta proposta

6.5.1 Armazenando o resultado

Até o momento sabemos que o resultado obtido de uma consulta esta em formato XML. Entretanto cada consulta retorna no máximo 100 registros. Sendo assim é necessário submeter várias vezes a URL composta, alterando o valor do parâmetro “start”. Na figura 4.1,

é notável a TAG <querycount>, armazenando a quantidade total de registro que a pesquisa retornou. Porém para obter todos os registros é necessário calcular o valor total da paginação. O calculo é dado pela divisão da quantidade de registros retornados na consulta, pela quantidade de registros por arquivo, parametrizado na URL. Um exemplo é descrito na seção 6.5.1.1.

6.5.1.1 Exemplo:

Seja tomada como exemplo uma pesquisa que retorne 5000 registros. Sendo que ao aplicarmos este valor na formula proposta na seção 6.5.1, é obtido como resultado o valor 50. Tendo o resultado obtido, a URL composta será submetida novamente por 50 vezes, alterando o valor do parâmetro "start", com o valor correspondente. No exemplo da URL abaixo, é exemplificado a composição das URLs:

http://api.technorati.com/search?key=[CHAVE]&query=smartphone&format=xml&language=pt&authority=n&start=**1**&limit=100

http://api.technorati.com/search?key=[CHAVE]&query=smartphone&format=xml&language=pt&authority=n&start=**2**&limit=100

...

http://api.technorati.com/search?key=[CHAVE]&query=smartphone&format=xml&language=pt&authority=n&start=**50**&limit=100

Tendo executado as 50 vezes, são obtidos 50 arquivos em formato XML. Após a obtenção, os mesmos serão processados pelas etapas seguintes, descritas nas seções 6.5.2 até a 6.6.

6.5.2 Tratamento do texto:

Esta seção descreve como é realizado o tratamento do corpus, composto a partir dos arquivos XML, obtidos na busca realizada junto a API Technorati.

O texto relevante para o processamento localiza-se nas TAGS, <title> e <excerpt>, do arquivo XML resultante da consulta, conforme apresentado na figura 4.1. Estas TAGS armazenam o título e o conteúdo de cada POST do blog indexado pela API. Cada consulta realizada na API, com o assunto escolhido para a formação do corpus, conforme descrito na seção 6.3, a mesma retorna em média 9000 palavras por arquivo XML, sendo que esta quantidade é variável conforme o tamanho dos POSTS.

Após a seleção dos textos do corpus formado, cada sentença proveniente do mesmo, precisa ser quebrada em partes menores, para então ser processado sintaticamente. Em PLN, este procedimento é conhecido como *Tokenização*.

SOARES, Fabio de A. (2008), define o processo de *tokenização* como:

> O primeiro passo de uma operação de Pré-processamento é a *tokenização* ou **atomização** e sua execução tem como finalidade seccionar um documento textual em unidades mínimas, mas, que exprimam a mesma semântica original do texto. O termo *token* é utilizado para designar estas unidades, que em muitas vezes correspondem a somente uma palavra do texto, porém, nem sempre estas unidades textuais não podem ser consideradas palavras ou apresentam mais de uma palavra: “21/10/2007”, “PM”, “R$100,00” e “couve-flor”.

Cada POST (titulo e conteúdo) é separado em *tokens*. Sendo este processo necessário para a realização do processamento sintático, descrito na seção 6.5.3.

6.5.3 Processamento Sintático:

Esta subseção tem por objetivo descrever os passos para a execução do processamento sintático executado pela ferramenta. Nas subseções 6.5.3.1 à 6.5.3.3 cada um dos passos realizados serão descritos.

6.5.3.1 StopWords

Conforme descrito na seção 6.1, para evitar um resultado exagerado, com palavras que não possuem relevância para processamento e a extração de informações, como por exemplo, as preposições, pronomes e artigos, o sistema possui uma lista com estes termos irrelevantes para o processamento. Sendo assim esta lista recebe o nome stopwords. No Quadro 4, é exemplificado algumas das palavras que o formam esta lista.

Termo
A
De
Da

Quadro 4 - Exemplo de uma lista de *Stopwords*

Na figura 4.3, é exibido um exemplo do processo de *tokenização* seguido pela remoção dos termos pela lista de *stopwords* (SOARES, Fabio de A. 2008, p44).

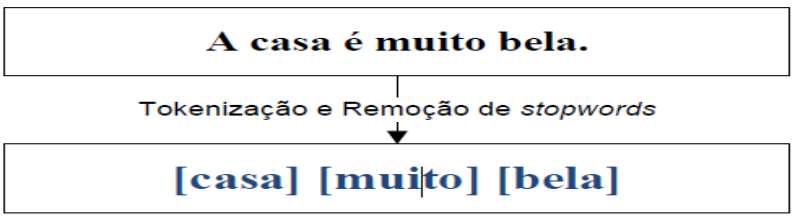

Figura 4.3 - Processo de tokenização seguido por remoção de stopwords (SOARES, Fabio de A. 2008, p44).

6.5.3.2 Dicionário:

Nesta Etapa, cada um dos *tokens*, sofre uma comparação com o dicionário correspondente ao idioma da consulta realizada, para então formar o resultado do processamento sintático. O *token* que não for encontrado no dicionário será adicionado em uma lista para ser processada estatisticamente, conforme descrito na seção 6.6. Para a realização desta etapa, existe a necessidade de se obter um dicionário de tradução. Este dicionário foi obtido das traduções do sistema operacional Linux. Sofrendo atualizações constantes, este dicionário possibilita um processamento mais confiável.

6.5.3.3 Resultado

O resultado obtido após as passagens dos *tokens*, na lista de Stopwords, para a remoção dos termos irrelevantes ao processamento sintático, e na comparação com o dicionário, descritos até o momento, os mesmos formam uma lista de palavras que serão processadas estatisticamente, para então obtermos alguma informação na pesquisa realizada. O processamento estatístico, na subseção 6.6, será detalhadamente descrito.

6.6 Processamento Estatístico

Após a passagem dos textos pelos métodos descritos nas seções 6.5.2 e 6.5.3, as informações geradas pelos mesmos serão processadas por métodos estatísticos, propostos e descritos na seção 5. Sendo assim, é possível se avaliar os resultados da ferramenta.

Para exemplificar este processo, será necessário utilizar as tabelas 1, 2 e 3, como exemplo. O primeiro passo, para o processamento estatístico, é o calculo da freqüência dos termos em cada documento. Este passo consiste em contar a quantidade de vezes, que o termo foi mencionado no documento. Por exemplo, o termo “Touchscreen”, exibido pela Tabela 4, apareceu 35 no Documento A, sendo o termo de maior freqüência neste documento. Entretanto somente o valor da freqüência, não possibilita saber se o termo possui algum valor informativo.

Documento A				
Termo	**Freqüência**	**TF**	**IDF**	**TF/IDF**
Touchscreen	35	1,000	0,477	0,477
Megapixel	29	0,829	0,176	0,146
Iphone	10	0,286	0,000	0,000

Tabela 1 – Exemplo de termos e informações estatísticas obtidas no processamento dos textos.

Documento B				
Termo	**Freqüência**	**TF**	**IDF**	**TF/IDF**
Camera	56	1,000	0,477	0,477
Megapixel	44	0,786	0,176	0,138
Iphone	21	0,375	0,000	0,000

Tabela 2 - Exemplo de termos e informações estatísticas obtidas no processamento dos textos.

Documento C				
Termo	**Freqüência**	**TF**	**IDF**	**TF/IDF**
Wi-fi	9	1,000	0,477	0,477
Presentão	7	0,778	0,477	0,371
Iphone	3	0,333	0,000	0,000

Tabela 3 - Exemplo de termos e informações estatísticas obtidas no processamento dos textos.

Para um termo ser relevante como informação nesta ferramenta, será utilizado como base o valor obtido no cálculo do método estatístico TF/IDF, descrito na seção 5.1.

Após os cálculos das freqüências dos termos nos documentos, o próximo valor a ser obtido é o TF e em seqüência o valor de IDF, ambos descritos na seção 5.1.2 e 5.1.3. Os termos que apresentarem maior valor de TF/IDF, serão considerados mais relevantes na pesquisa.

Para exemplificar o resultado gerado pela ferramenta, as tabelas 1, 2 e 3, cada uma deve ser considerado um documento. Apesar das mesmas apresentarem termos de freqüência de valor elevado, os mesmo não são considerados relevantes. Devido ao fato que aparecem em todos os documentos, conseqüentemente não possui uma importância significativa perante a pesquisa. Sendo assim o termo "Camera" possui uma maior relevância na pesquisa, pois o mesmo possui um valor alto de TF/IDF. Como resultado da análise gerada pela ferramenta, o quadro 5, exibe os resultados gerados pela mesma, ordenados pelo valor de TF/IDF, ou seja, ordena os termos de maior relevância.

Termo	TF/IDF
Camera	0,477
Touchscreen	0,477
Wi-fi	0,477
Presentão	0,371
Megapixel	0,142
Iphone	0,000

Quadro 5 - Resultado gerado pela ferramenta

Ao final dos processos descritos até o momento, é possível realizar uma análise dos termos de maior relevância para a pesquisa. Sendo que estes termos resultantes possam ser considerados como termos comuns na internet, marcas ou gírias, constantemente usados nos blogs.

CONCLUSÃO

Neste presente trabalho foi realizado um estudo e análise do material bibliográfico adquirido durante as pesquisas realizadas ao longo do ano de 2009. Este estudo proporcionou um grande conhecimento sobre PLN, bem como a área de extração de informações.

Sem dúvida o PLN é uma área admirável e impressionante, pelo fato como a mesma pode tratar a linguagem natural, proporcionando uma série de soluções para o que se deseja desenvolver, no que se diz respeito ao Processamento da Linguagem Natural. Neste trabalho pode-se destacar que um dos maiores desafios superados, é o desenvolvimento de um corpus adaptado para os blogs, sendo que o mesmo é inédito, ao caso que não foi localizado nada semelhante em núcleos de pesquisa sobre PLN.

A ferramenta foi desenvolvida, seguindo cada um dos passos descritos neste trabalho; mostrando-se muito útil e confiável. Devido à utilização de métodos próprios para análise de textos e de processamento estatístico.

Analisando os resultados gerados pela ferramenta, chegasse à conclusão de que a ferramenta é muito útil para empresas, bem como a grupos de estudos direcionados a lingüística. Enfatizando seu uso pelas empresas, a mesma as auxilia em descobrir se o nome de sua marca é comentado em determinados assuntos, quais produtos apresentam maior destaque nos comentários, bem como descobrir as preferências descritas pelos seus consumidores e avaliadores. Na área de estudos lingüísticos, a ferramenta possibilita analisar, a maneira como as pessoas tem se expressado nos blogs. Possibilitando a visualização do emprego de coloquialismo, bem como o uso de palavras estrangeiras.

Ao desenvolver este trabalho obtive a grande gratificação da adição de mais uma área de conhecimento, cuja mesma é de grande valia atualmente, ao caso que a mesma promove uma evolução muito grande na interação entre o ser humano e máquina.

REFERÊNCIAS BIBLIOGRÁFICAS

AIRES, Rachel. Nilc TagSets. 1998. Disponível em: http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html. Acesso em 18 de junho de 2009.

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística.** Dezembro de 2006. Disponível em:< http://www.icmc.usp.br/~taspardo/NILC-TR-09-08.pdf >. Acesso em 30 de maio de 2009.

AVILA, Christiano M. O. **Desenvolvimento de um Sistema de Recomendação de Artigos Científicos e Avaliação de Métodos de Extração de Palavras-Chave.** Dezembro 2006. Disponível em:<http://ppginf.ucpel.tche.br/TI-arquivos/2006/ChristianoAvila/PPGINF-UCPel-TI-2006-2-01.pdf> Acesso em 17 de outubro de 2009.

BICK, Eckhard. SymbolSet Manual. 2009. Disponível em: http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html. Acesso em 18 de junho de 2009.

HERNANDES, Carlos A. M.; SANTANA Roberto A. S.; FALCÃO Sérgio D. Sobre o uso do chat como ferramenta auxiliar de ensino e aprendizagem no curso de Mestrado em Informática da Universidade Católica de Brasília. Universidade Católica de Brasília. Disponível em:
< http://carlosmamede.org/Artigo%20sobre%20chat%20na%20UCB%20-%20publicado.pdf >. Acesso em 15 de Março de 2009.

Nunes, C. & Barros, F. A., **ProdExt: a knowledge-based wrapper for extraction of technical and scientific production in Web pages, Proc. of the International Joint Conference.** IBERAMIA-SBIA 2000 - Open Track, pp. 106-115.

SIFRY, David. **About Technorati.** 2002. Disponível em: < http://technorati.com/about-technorati/>. Acesso em 10 de outubro de 2009.

SILVA, Bento C. D.; MONTILA, Gisele; RINO, Lucia H. M.; SPECIA, Lucia; NUNES, Maria das G. V.; OLIVEIRA JR, Osvaldo N.; MARTINS, Ronaldo T.; PARDO, Thiago A. S., **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. Núcleo Interinstitucional da Lingüística Computacional: Agosto de 2007. Disponível em:
< http://www.icmc.usp.br/~taspardo/NILCTR0710-DiasDaSilvaEtAl.pdf >. Acesso em 12 de Fevereiro de 2009.

SOARES, Fabio de A. **Mineração de Textos na Coleta Inteligente de Dados na Web.** Pontifícia Universidade Católica do Rio de Janeiro: Setembro de 2008.
Disponível em: < http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0621324_08_pretextual.pdf> Acesso em 12 de Agosto de 2009.

TURNEY, P. **Learning to Extract Keyphrases from Text, Tech. Report Number**

NRC−41622, National Research Council Canada, Institute for Information Technology, 1999.

WOJCIECHOWSKI, Jaime et al. **Recuperação de informações em base de e-mails.** Fevereiro 2007. Disponível em: <http://www.jaimewo.com.br/files/RecuperacaoDeInformacoesEmBaseDeEmails.pdf>. Acesso em 15 de outubro de 2009.