UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

ESCOLA DE INFORMÁTICA APLICADA

Descoberta Automática de Palavras-chave Para Classificação de Textos.

Eduardo Moreira Leite

Orientadora

Prof. Kate Revoredo, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

JANEIRO DE 2016

Descoberta Automática de Palavras-chave
Para Classificação de Textos.

Eduardo Moreira Leite

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada por:

_____

Prof. Kate Revoredo, D.Sc. (UNIRIO)

_____

Prof. Fernanda Baião, D.Sc. (UNIRIO)

_____

Prof. Flávia Santoro, D.Sc. (UNIRIO)

RIO DE JANEIRO, RJ – BRASIL.
JANEIRO DE 2016

Agradecimentos

Agradeço primeiramente aos meus pais, Alexandre e Gleidy, pois não seria possível chegar até este momento sem a ajuda deles que nunca deixaram faltar nada para que eu pudesse continuar estudando e não desistir de galgar um futuro melhor, além do fato de sempre me apoiarem em todas minhas decisões.

Agradeço também à minha irmã Camila, que sempre que pode fez o possível pra me ajudar, fosse em minhas tarefas acadêmicas ou com as tarefas diárias de casa.

À minha amiga Mariana, que com sua amizade sempre soube me acalmar nos momentos de nervosismo e sempre me encorajou em tudo na vida, até mesmo quando pensei em desistir e que sem dúvidas sempre adicionou razão as minhas decisões que insistem em pender para um lado emocional.

Agradeço ao meu amigo Tiago Neves que foi meu grande parceiro nessa jornada acadêmica, que sempre me apoiou nos trabalhos e nos estudos para as provas.

À minha amiga Olivia Klejnow que sempre que pode não mediu esforços em me ajudar, na obtenção de artigos e materiais de leitura para que esse trabalho pudesse ser realizado.

Aos meus amigos do grupo Preciosos, que sempre creditaram a mim confiança e sempre me apoiaram, além de que sempre que puderam me ajudaram com algum teste que precisei fazer, fosse de algum código ou até mesmo de alguma configuração.

Agradeço à Universidade Federal do Estado do Rio de Janeiro (UNIRIO) e a Escola de Informática Aplicada (EIA) por me fornecerem uma educação digna e de qualidade para desenvolver meus conhecimentos no universo da Tecnologia da Informação e aguçar minha curiosidade na constante busca pelo conhecimento.

Agradeço também ao CNPQ e à Universidade de Newcastle que me admitiu como aluno do programa Ciências sem Fronteiras, que permitiu em um ano que eu evoluísse não só como aluno, mas como pessoa melhor.

Em especial à minha orientadora, Kate Revoredo, que sem dúvidas teve a maior paciência comigo, e que sempre acreditou em mim e me apoiou sem medir esforços, mesmo antes desse trabalho, além de que sempre se fez disponível pra me ajudar em

qualquer problema que eu tivesse no mundo acadêmico mesmo que não fosse da competência da mesma.

Por fim, e não menos importante, agradeço a Deus por sempre me proteger e tranquilizar e prepara meu caminho adiante para que seja possível trilhar os próximos desafios.

RESUMO

Usuários muitas vezes tentam buscar informações sobre um tema de interesse a partir de múltiplas fontes de informação. Por vezes, a informação que esta sendo buscada pode ser expressa em termos, palavras-chave, em um documento, artigo, publicações e etc.… disponível com o intuito de facilitar o motor de busca a identificar o conteúdo destas.

A grande questão é que nem sempre estas palavras-chaves estão disponíveis, e fazer a inclusão das mesmas de forma manual é algo inviável. Criando-se o problema de como fazer para se identificar as palavras-chaves mais relevantes para um determinado documento, de forma que se seja possível categorizar o conteúdo do mesmo.

Por outro lado, existem algumas técnicas de classificação automática de textos que, através de análises estatísticas, estão sendo amplamente utilizadas para extrair o conhecimento de um conjunto de textos não classificados, onde essas técnicas vêm apresentando bons resultados para expressar o significado do conteúdo interno dos textos analisados.

Neste trabalho, os desafios destas classificações são abordados através da aplicação de algumas dessas técnicas de extração de texto, visando de uma forma automatizada retirar as palavras-chaves de um determinado texto, calculando suas relevâncias, de uma forma que estas expressem o significado desses textos, permitindo assim a classificação dos mesmos. Para isso, essa proposta foi avaliada usando um conjunto de dados do mundo real, implementando a extração de palavras-chave com o intuito de classificar as publicações de um site que aglomera publicações de diversos blogs que abordam conteúdos do universo feminino, preadly.com, onde esse estudo foi realizado em uma base de dados que foi fornecido pelo próprio site.

Palavras-chave: Palavras-chave, Extração Automática, Classificação de Textos, Frequência de termos, Alchemy.

ABSTRACT

Users often try to seek information on a topic of interest from multiple information sources. Sometimes, the information that is being sought can be expressed in terms, keywords, in a document, article, and publications, available in order to facilitate the search engine to identify the contents thereof.

The big problem is that not always these keywords are available, and the manual inclusion of them is something almost impossible. Creating the problem of how to identify the most relevant keywords, for a particular document, so that it is possible to categorize it contents.

On the other hand, there are some text classification techniques, that through statistical analysis, are being widely used to extract the knowledge of a set of unclassified texts, where these techniques have shown good results to express the meaning of the analyzed texts internal content.

In this work, the challenges of these classifications are addressed through the application of some of these text extraction techniques, aiming for an automated way to remove the keywords of a given text, calculating their relevance in a way that they can express the meaning of these texts, allowing them to classification of it content. For this, the proposal was evaluated using a set of real-world data, implementing the extraction of keywords in order to rank the publications from a site that crowd publications of several blogs, that address the feminine universe, called preadly.com, where this study was conducted in a database that was provided by the site itself.

Keywords: keywords, Automatic Keywords Extraction, Text Classification, Term Frequency, Alchemy.

Índice

Índice de Tabelas

Índice de Figuras

1 Introdução

1.1 Motivação

A classificação de documentos ou categorização de documento é um problema na biblioteconomia, ciência da informação e ciência da computação. Esta tarefa consiste em atribuir a um documento uma ou mais classes ou categorias. Isto pode ser feito "manualmente" (ou "intelectualmente") ou através de algoritmos. A classificação intelectual de documentos tem sido o principal problema da biblioteconomia, enquanto a classificação de documentos através de uma abordagem algorítmica é da competência da ciência da informação e da ciência da computação. [5]

No âmbito da ciência da informação, que é a parte que é abordada nesse trabalho, como temos hoje em dia a presença da tecnologia quase que constante em tudo que utilizamos, seja o computador, tablets ou até mesmo o nosso celular, tudo isso gerando um grande volume de dados, criou-se, portanto a necessidade de extrair informações desses dados com o intuito de gerar novos conhecimentos a partir deles.

O avanço da tecnologia permitiu o armazenamento desse grandes volumes de dados e a análise desses dados pode ser útil para as diferentes organizações. O Facebook, por exemplo, processa mais de 500 TB de dados diariamente [2], dados esses que quando processados, podem gerar informações úteis, e que podem ser utilizadas de diversas formas, como por exemplo, com o intuito de se gerar uma propaganda direcionada para um determinado usuário, baseado nas buscas realizadas pelo mesmo no seu celular. Criando uma nova forma de interação com o usuário, muito mais personalizada. Nas palavras do pesquisador Erick Siegel, em seu livro Predictive Analytics, *"os dados que coletamos atualmente nos permitem ver as coisas, que até pouco tempo atrás eram grandes de mais para enxergarmos"*.

Além disso, a análise dos dados pode se beneficiar caso esses dados estejam

categorizados, ou seja, se palavras-chaves, que indicam o principal teor do dado, estiverem associadas a esses dados. Artigos científicos, por exemplo, quando associados a palavras-chave, auxiliam na análise do perfil da conferência naquele ano ou de uma edição de uma revista. A associação de palavras-chave também permite uma busca mais eficiente por um artigo de interesse, já que palavras-chave indicam a ideia principal de um artigo e assim permitem alinhar facilmente o interesse do pesquisador a esse artigo.

Essas palavras chaves são frequentemente informadas pelos próprios autores. E grande parte dos documentos disponíveis na Web ainda não possuem palavras-chave atribuídas a eles. A associação manual de palavras-chave é algo inviável, considerando-se o grande volume de dados disponível atualmente. Sendo assim, existe uma demanda por abordagens que consigam descobrir automaticamente as palavras-chave que melhor descrevem/representam o assunto de um determinado documento, para então associar essas palavras a eles.

1.2 Trabalhos Relacionados

A busca por um método para classificação de texto é algo que é estudado até hoje e que já foi abordado em diversos trabalhos. De acordo com [18] que busca a identificação de tópicos persistentes em uma publicação em redes sociais, por meio da extração de palavras-chave de cada publicação, existem inúmeras pesquisas que tentam descobrir palavras-chaves que podem caracterizar o conteúdo de um documento ou texto. Onde basicamente, uma linha de pesquisa consiste em técnicas de mineração de texto para extrair padrões para descobrir tópicos de um determinado conjunto de documentos, e a outra linha de pesquisa pode ser categorizada como a abordagem para de extração de uma palavra-chave, onde que vários métodos têm sido propostos com base em métodos não supervisionados ou supervisionados.

Em [17], segundo o autor, as evidências experimentais acumuladas ao longo dos últimos 20 anos indicam que os sistemas de indexação de texto com base na atribuição de palavras-chave individuais quando devidamente ponderadas, produzem resultados de recuperação superiores àqueles obtidos com outras representações de texto mais elaborados.

Já em [16], Os autores avaliam os interesses dos usuários de um *microblog* chinês, Sina Weibo, que funciona de forma semelhante ao Twitter, extraindo das mensagens

publicadas pelos usuários, palavras julgadas importantes, e que são descritas palavras-chave. Segundo eles, alguns pesquisadores consideram a extração de palavras-chave como uma tarefa de classificação binária (se é ou não é uma palavra-chave) e aprendem modelos de classificação, utilizando dados de treinamento. Estes métodos supervisionados precisam que dados de treinamento sejam anotados manualmente, o que vem a ser um processo demorado.

O método sem supervisão mais direto que existe para a extração de palavras-chave que está sendo amplamente usado, classificando-as de acordo com as suas frequências, é o TF-IDF que é usado pela API da linguagem Alchemy, utilizada nesse trabalho e explicado na seção 2.4 do Capítulo 2.

Além disso, em [16] por se tratar de uma análise em uma rede social, o uso da linguagem informal no *microblog* é bastante presente, o que segundo eles gerou um desafio para a realização do trabalho e que só foi resolvida através de métodos baseados na frequência de termos, onde foi possível efetuar a retirada de *stopwords*, que em computação são palavras mais comuns a língua, filtradas fora antes ou depois do processo de processamento de linguagem natural, permitindo selecionar-se um melhor subconjunto de palavras-chave, enfatizando a eficácia dessa abordagem.

Portanto como o conteúdo da base estudada neste trabalho tem um assunto principal, que é o universo feminino, que é muito amplo, esse tema possui inúmeras ramificações, levando esse trabalho a concentrar-se em métodos não supervisionados para a extração de palavras-chave, como ocorre em [16].

1.3 Objetivos

Considerando-se o problema apontado de como se extrair automaticamente as palavras-chave associadas a um determinado documento, o objetivo desse trabalho de conclusão de curso, é justamente o de propor um método automático, para a solução de uma forma eficiente, desse problema de classificação de textos. Utilizando-se de conceitos estudados ao longo do curso de Sistemas de Informação, para extrair das publicações em um *site*, palavras-chave que transmitam o conteúdo interno destas publicações, onde a relevância dessas palavras para o texto é feita por meio de análises estatísticas.

Para avaliar essa abordagem foi utilizado um cenário real, no site, www.preadly.com. Onde a organização mantedora do site nos contatou com uma demanda, indicando que possui hoje, uma dificuldade em categorizar suas publicações, devido a diversidade de assuntos abordados por elas, de uma forma que as palavras-chave realmente transmitam o conteúdo presente no texto, sem que isso seja feito de uma forma manual. Uma análise quantitativa e qualitativa foi feita verificando que a abordagem proposta nesse trabalho é promissora.

1.4 Organização do texto

O presente trabalho está estruturado em 5 capítulos e, além desta introdução, será desenvolvido da seguinte forma:

- Capítulo II: introduz os temas que são fundamentais para fazer o melhor proveito da informação tratada nesse trabalho.
- Capítulo III: trata do cenário em que foi aplicado esse trabalho.
- Capítulo IV: aborda a realização trabalho proposto utilizando tudo o que foi estudado.
- Capítulo V: Conclusões – Reúne as considerações finais, assinala as contribuições da pesquisa e sugere possibilidades de aprofundamento posterior.

2 **Fundamentação Teórica**

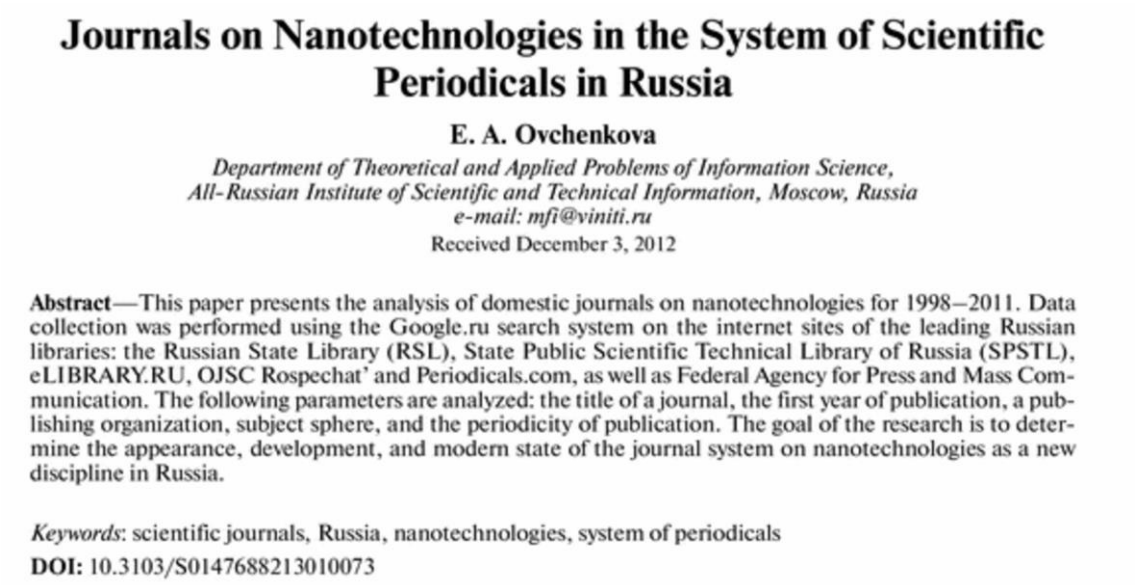
Neste Capítulo são abordados os conceitos importantes e fundamentais para a compreensão do tema proposto. Ao final deste capítulo, o leitor terá conhecimentos dos conceitos de palavras-chave, cálculo da Frequência de termos(TF), Frequência inversa em um documento (IDF), do cálculo do valor da Frequência de termos e frequência inversa do termo no documento, que é a combinação das técnicas de TF e IDF e por último e não menos importante da Linguagem Alchemy..

2.1 Palavras-Chave.

Uma palavra-chave, na recuperação de informação, é um termo que capta a essência do tema de um documento, explicada como “Descritora” por Calvin Mooers, Cientista da computação americano conhecido por seu trabalho na recuperação de informação, em 1948 [3] ; Elas, no ramo da biblioteconomia, compõem um vocabulário controlado para uso em registros bibliográficos. Sendo portanto, usadas no por bibliotecas para o controle da literatura. Já no âmbito de sistemas de informação, elas podem ser usadas como termos para indexar documentos, por exemplo, em um catálogo ou em um motor de busca. Elas são criadas por meio da análise manual de documentos com a indexação de assuntos ou automaticamente com uma indexação automática ou através de métodos mais sofisticados de extração de palavras-chave, que é o objetivo desse trabalho.

As palavras chaves podem ser representadas de diferentes formas, seja através de uma simples lista com o intuito de categorizar um documento, como ocorre em artigos científicos, como podemos ver na figura 1, onde não há distinção de importância entre o peso dessas palavras para o texto, ou através de uma núvem de palavras-chave, ou núvem de *tags* como é comumente referenciada, ou através de uma lista ordenada, que são abordagens mais completas por justamente transmitirem a importância de cada palavra para o conteúdo presente no texto.

Figura 1 Exemplo de Núvem de *Tags*.



Journals on Nanotechnologies in the System of Scientific Periodicals in Russia

E. A. Ovchenkova

Department of Theoretical and Applied Problems of Information Science, All-Russian Institute of Scientific and Technical Information, Moscow, Russia
e-mail: mfi@viniti.ru

Received December 3, 2012

Abstract—This paper presents the analysis of domestic journals on nanotechnologies for 1998–2011. Data collection was performed using the Google.ru search system on the internet sites of the leading Russian libraries: the Russian State Library (RSL), State Public Scientific Technical Library of Russia (SPSTL), eLIBRARY.RU, OJSC Rospechat' and Periodicals.com, as well as Federal Agency for Press and Mass Communication. The following parameters are analyzed: the title of a journal, the first year of publication, a publishing organization, subject sphere, and the periodicity of publication. The goal of the research is to determine the appearance, development, and modern state of the journal system on nanotechnologies as a new discipline in Russia.

Keywords: scientific journals, Russia, nanotechnologies, system of periodicals
DOI: 10.3103/S0147688213010073

Figura 2 Exemplo de Palavras-chave para categorizar um Artigo Científico.

2.2 Frequência de termos (TF).

A primeira forma de frequência de termo(TF) é dada a Hans Peter Luhn (1957) e baseia-se na suposição de Luhn que: "O peso que um termo ocorre em um documento é simplesmente proporcional à frequência do termo" [6].

A Frequência do termo, mede a frequência com que um termo ocorre em um documento. Uma vez que cada documento é diferente de comprimento, é possível que um termo apareça muito mais vezes em documentos longos do que em documentos curtos. Assim, a frequência do termo é frequentemente dividida pelo comprimento do documento (Total de termos em um documento), Como forma de normalização. tf(*t*,*d*) = (número de vezes termo t aparece em um documento) / (número total de termos no documento).

Apesar disso é fácil encontrar, outras formas de calcular essa frequência[7]

- Frequências Booleanas: $\mathrm{tf}(t,d) = 1$ se t occorre em d e 0 caso contrário;
- Frequência escalar logaritma: $\mathrm{tf}(t,d) = 1 + \log f_{t,d}$, ou 0 se $f_{t,d}$ é zero;
- Frequência aumentada para prevenir um viés em documentos mais longos por exemplo. frequência simples dividida pela frequência simples máxima de qualquer termo no documento:

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

2.3 Frequência inversa do termo no documento (IDF).

A Frequência inversa do documento, mede o quão importante é um termo. Enquanto computando TF, todos os termos são considerados igualmente importantes. No entanto, sabe-se que determinados termos, tais como "é", "de", e "que", podem aparecer uma série de vezes, mas que possuem pouca importância.

Isso tende a enfatizar incorretamente documentos que ocorrem esses termos com mais frequência, sem dar peso suficiente para os termos mais significativos. O termo “de”, por exemplo, não é uma boa palavra-chave para distinguir documentos e termos relevantes e não relevantes, ao contrário das palavras menos comuns com "casa" e "carro". Assim, um fator de frequência de documento inverso é incorporado o que diminui o peso de termos que ocorrem frequentemente no conjunto de documentos e aumenta o peso de termos que ocorrem raramente.

Karen Spärck Jones (1972) concebeu uma interpretação estatística de especificidade termo chamado IDF, que se tornou um marco da ponderação de termos:

"A especificidade de um termo pode ser quantificada como uma função inversa do número de documentos em que ocorre"[1].

Assim, precisamos pesar os termos frequentes, enquanto ampliar os raros, calculando o valor IDF da seguinte forma:

A fração logaritmicamente escalada dos documentos que contenham a palavra, obtida pela divisão do número total de documentos pelo número de documentos que contenham o termo, e em seguida, tomando o logaritmo desse quociente.

$$\operatorname{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

onde

- N: Número total de documentos no corpo $N = |D|$
- $|\{d \in D : t \in d\}|$: número de documentos onde o termo **t** aparece (exemplo $\operatorname{tf}(t, d) \neq 0$). Se o termo não está no corpo do texto, isso levará a uma divisão por 0, por isso é comum ajustar o denominador para : $1 + |\{d \in D : t \in d\}|$.

Matematicamente a base da função de log não importa e constitui um fator multiplicativo constante para o resultado global.

2.4 Frequência de termos e frequência inversa do termo no documento (TF-IDF).

TF-IDF é um peso frequentemente utilizado na recuperação de informação e de mineração de texto. Este peso é uma medida estatística utilizada para avaliar o quão importante é uma palavra para um documento em uma coleção ou texto. Onde, a importância aumenta de forma proporcional ao número de vezes que uma palavra aparece no documento, mas é compensada pela frequência da palavra nele. Isso auxilia a distinguir o fato da ocorrência de algumas palavras serem geralmente mais comuns

que outras.

Variações do sistema de ponderação TF-IDF são muitas vezes utilizados pelos motores de busca como um instrumento central na pontuação e classificação de relevância de um documento dado uma consulta do usuário.[10]

Uma das mais simples funções de classificação é calculada pela soma dos TF-IDF para cada termo da consulta; Onde muitas funções de classificação mais sofisticadas são variantes deste modelo simples.

O valor de TF-IDF pode ser utilizado com sucesso como palavras-chave em vários campos, incluindo sumarização de texto e classificação.

TF-IDF é calculado como:

TF-IDF(t,d,D) = tf(t,d) x idf(t,D)

Ou seja, o peso TF-IDF é composto por dois termos: o primeiro calcula a frequência do termo normalizado (TF), o número de vezes que uma palavra aparece em um documento, dividido pelo número total de palavras contidas nesse documento; O segundo termo é a frequência inversa do documento(IDF), que é calculada como o logaritmo do número dos documentos do corpus, dividido pelo número de documentos onde o termo específico aparece.

Com isso, um peso elevado no TF-IDF é atingido por uma alta frequência (no dado documento) e uma frequência baixa de documento do termo em toda a coleção de documentos; Portanto, os pesos tendem a filtrar os termos comuns. Como a proporção dentro função log da IDF é sempre maior do que ou igual a 1, o valor de IDF (e TF-IDF) é maior do que ou igual a 0. Como um termo aparece em vários documentos, a relação dentro do logaritmo se aproxima de 1 , elevando o IDF e TF-IDF mais próximo de 0.

Exemplo Prático:

Considere-se um documento que contém 100 palavras em que a palavra carro aparece 4 vezes. A frequência do termo (TF) para carro é, (4/100) = 0,04. Agora, suponha que temos 10 milhões de documentos e a palavra carro aparece em mil destes. Então, a frequência de documento inversa (IDF) é calculada como o log (10.000.000 / 1000) = 4. Assim, o peso TF-IDF é o produto destas quantidades: 0,04 * 4 = 0,16.

2.5 Linguagem Alchemy

Para que a idea desse trabalho fosse executada, foi utilizada uma API de processamento de linguagem natural em java, da Linguagem Alchemy da IBM, de extração de palavras-chaves para que se pudesse extrair as mesmas automaticamente de cada um das publicações da base de dados do site preadly.

A AlchemyAPI possui 12 funções para análise de texto, cada uma das quais utiliza técnicas de processamento de linguagem natural sofisticadas para analisar o seu conteúdo e adicionar um alto nível informação semântica, dentre as suas funções de análise de texto temos:

2.5.1 Entity Extraction

A função de Extração de Entidades especifica coisas como pessoas, lugares e organizações. Essa API é capaz de identificar pessoas, empresas, organizações, cidades, características geográficas e outras entidades digitadas a partir do seu HTML, texto ou conteúdo baseado na web.

A extração dessas entidades pode adicionar uma riqueza de conhecimento semântica a um conteúdo para ajudar a compreender rapidamente o assunto do texto. É um dos pontos de partida mais comum para a utilização de técnicas de processamento de linguagem natural para enriquecer o conteúdo.

Extração de Entidades da Alchemy é baseada em algoritmos estatísticos sofisticados e tecnologia de processamento de linguagem natural. Ela é a única na indústria, segundo eles, com uma combinação de apoio multilíngue, dados vinculados, desambiguação de entidade sensível ao contexto, suporte abrangente de digitação e extração de citações.

Figura 3 Exemplo de Extração de Entidade.

2.5.2 Sentiment Analys

Sentimento é a atitude ou opinião em direção a algo, como uma pessoa, organização, produto ou localização. Essa função da Análise de Sentimento da Alchemy fornece mecanismos de fácil utilização para identificar o sentimento positivo ou negativo dentro de qualquer documento ou página da web.

O método de análise de sentimento é capaz de computar o nível de sentimento de um documento, o sentimento de metas especificadas pelo usuário, sentimento de nível de entidade, sentimento de nível de citação, sentimento direcional e de sentimento de nível de palavra-chave. Estes múltiplos modos de análise de sentimento provêm uma variedade usos possíveis que variam de monitoramento de mídia social para a análise de tendências.

O Algoritmo de análise de sentimento de AlchemyAPI procura palavras que carregam uma conotação positiva ou negativa, e em seguida, descobre a pessoa, o lugar ou coisa na qual eles estão se referindo. Ele também entende negações (ou seja, "este carro é bom" versus "este carro não é bom") e modificadores (ou seja, "este carro é bom" versus "este carro é realmente bom"). O método de análise de sentimento funciona em documentos grandes e pequenos, incluindo artigos de notícias, publicações, análises de produtos, comentários e até tweets

Figura 4 Exemplo de Análise de Sentimento.

2.5.3 Concept Tagging

O método de Marcação de Conceito emprega técnicas sofisticadas de análise de texto para efetuar a marcação de conceitos em documentos de uma forma semelhante a como

nós humanos identificamos conceitos. Ela é capaz de fazer abstrações de alto nível por entender como os conceitos se relacionam entre si, e pode identificar conceitos que não são necessariamente diretamente referenciados no texto.

Por exemplo, se um artigo menciona CERN(*Conseil Européen pour la Recherche Nucléaire ou Organização Euroéia de Pesquisas Nucleares*) e o bóson de Higgs, Ela vai marcar Grande Colisor de Hádrons, que é o maior acelerador de partículas do laboratório CERN e que é responsável por estudos relacionados a partícula do Bóson de Higgs. Seu principal objetivo é obter dados sobre colisões de feixes de partículas como um conceito, mesmo que o termo não seja mencionado explicitamente na página. Usando o conceito de marcação você pode realizar uma análise de mais alto nível do seu conteúdo de identificação.

Figura 5 Exemplo de Extração de Conceitos

2.5.4 Relation Extraction

Relação é a ligação de um assunto, ação e de objetos dentro de uma sentença. O método de Extração de Relação é capaz de analisar frases e extrair sujeito, ação e objeto de formulário e, em seguida, adicionar informação semântica adicional, como extração de entidade, a extração de palavras-chave, análise de sentimento e identificação de localização.

Essa extração pode ser usada para identificar automaticamente sinais de compra, os principais eventos e outras ações importantes.

Figura 6 Exemplo de Extração de Termos Relacionados

2.5.5 Taxonomy Classification

O método de Classificação de Taxonomia classifica automaticamente um texto, HTML ou conteúdo baseado na web em uma taxonomia hierárquica. Usando estatísticas complexas e tecnologia de processamento de linguagem natural, esse método de classificação de taxonomia pode classificar seu conteúdo em sua categoria temática mais provável até cinco níveis de profundidade.

Níveis mais profundos permitem classificar o conteúdo em subsegmentos mais precisos e lucrativos. Por exemplo, um aplicativo que se concentra na identificação de conteúdo para discutir práticas de empréstimos pessoais pode estreitar a sua classificação em sub temas que têm como alvo as decisões com resolução mais fina.

/ finanças / financiamento / empréstimos / cartões de crédito pessoais

/ finanças / finanças pessoais / empréstimo financiamento / home

/ finanças / financiamento / empréstimos / empréstimos pessoais pessoais

/ finanças / financiamento / empréstimos / empréstimos estudantis pessoais

/ finanças / financiamento / empréstimos de financiamento pessoal / veículo

Ela pode ser usada para filtrar ou agrupar conteúdo antes de realizar uma análise mais aprofundada ou para acompanhar os temas de alto nível dos seus documentos.

2.5.6 Author Extraction

Esse método permite extrair automaticamente informações sobre o autor de artigos de notícias ou publicações. Ela permite que o conteúdo de publicações on-line seja

categorizada por autor.

Combinado com outras funções de análise de texto da API do Alchemy, pode-se gerar uma núvem de *tags*, identificar sentimentos em relação a tópicos, encontrar opiniões e fatos para os autores individuais expressos.

Figura 7 Exemplo de Extração de Autor.

2.5.7 Language Detection

O método de Detecção de Linguagem fornece um recurso de identificação de idioma robusto capaz de detectar o idioma de qualquer texto, HTML ou conteúdo baseado na web.

Com a detecção de idioma, pode-se facilmente classificar ou filtrar qualquer conteúdo baseado na linguagem que foi escrito.

Figura 8 Exemplo de Identificação de Idioma

2.5.8 Text Extraction

O método de Extração de texto pode extrair automaticamente as informações importantes a partir de uma página web, efetuando a remoção de links de navegação, anúncios e outros conteúdos indesejados. Com isso é possível se concentrar apenas no texto-chave para melhorar a indexação do site analisado, aumentando a relevância

contextual para publicidades e simplificar a análise. O método de Extração de texto também pode retornar os links embutidos no conteúdo importante, o que torna possível a utilização para aplicações *web-crawling.*

Figura 9 Exemplo de Texto Sendo Extraído de Um HTML

2.5.9 Microformats Parsing

Esse método faz a análise de micro formatos que estão incluídos no HTML das páginas da web para adicionar informação semântica. Estes micro formatos permitem que a página da web seja mais facilmente verificada e processada automaticamente através de softwares. Eles são normalmente utilizados para informações de contato, coordenadas geográficas, informações de licença e informações semelhantes. Ela pode ser usada para melhorar a categorização página da Web, indexação e executar tarefas de descoberta de conteúdo.

2.5.10 Feed Detection

Feeds são muitas vezes incorporado em sites para permitir que os visitantes e leitores de *feeds* acessem o conteúdo destes sites. Esse método de detecção de *Feed* pode encontrar os *feeds* dentro de páginas web e retornar os *links*. Essa detecção pode ser usada para descobrir novos conteúdos, incluindo blogs, artigos de notícias e fluxos de comentário

2.5.11 Linked Data

A função de Dados Referenciados é um método de exposição, compartilhamento e conexão de dados na web via URLs. Ela visa estender a Web como um bem comum de dados através da publicação de vários conjuntos de dados abertos como RDF(**Resource Description Framework** (**RDF**) ou Framework de descrição de recursos que é uma

linguagem para representar informação na Internet[13]) na Web e definindo ligações RDF entre itens de dados de diferentes fontes de dados. A nuvem de dados referenciados atualmente consiste de mais de 7,4 bilhões de triplas RDF, interligados por 142+ milhões de ligações RDF.

Figura 10 Diagrama da núvem de dados vinculados da alchemy.

Essa API fornece suporte abrangente para RDF e Linked Data, permitindo que qualquer conteúdo possa ser trazido para a web semântica com relativa facilidade. Ela está disponível para extração entidades, o conceito de marcação e de extração relação, e permite que uma quantidade incrível de informação adicional a ser analisado.

Por exemplo, se o Facebook é identificado como uma entidade em seu conteúdo, usando dados referenciados torna-se possível obter a seguinte informação: número de funcionários, o número de locais em todo o mundo do Facebook.

2.5.12 Keyword Extraction

O método de Keyword Extraction é capaz de extrair palavras-chave de arquivos HTML, texto ou conteúdo baseado na web, por isso que a mesma foi escolhida para ser utilizada na execução da ideia proposta neste trabalho.

O procedimento adotado no caso de documentos HTML ou de conteúdo web, é de que o conteúdo do documento é normalizado, ou seja, é feita uma limpeza (remoção de anúncios, links de navegação e outros conteúdos sem importância), onde esse método detecta automaticamente o idioma do conteúdo e, em seguida, executa a análise apropriada, onde emprega algoritmos estatísticos sofisticados baseados no modelo de

frequência de termos citados neste capítulo e tecnologias de processamento de linguagem natural para analisar seus dados, onde extraídos esses meta-dados, eles podem ser retornados em formatos XML, JSON, RDF, e Microformats rel-tag.

Figura 11 Extração de Palavras-Chave

3 Realização do Trabalho

Neste Capítulo é abordada a execução da ideia apresentada no capítulo 1 utilizando como base os conceitos do capítulo 2. Ao final deste capítulo, o leitor aprenderá como colocar em prática os conceitos apresentados no capítulo anterior, bem como verá o resultado desses conceitos aplicados em um cenário real.

3.1 Escopo do Trabalho.

Este trabalho efetua extração automática de palavras-chave em publicações na web, onde é formado um conjunto de palavras-chaves em um formato de uma *string* ordenada por ordem de relevância, onde sua relevância é calculada através de seu peso, que é baseado no conceito de frequência de termos em um documento, ao invés de efetuar a formação de uma figura de palavras-chaves, como é feita na abordagem de nuvem de tags, citada no capitulo 2.

Essa análise foi aplicada em uma parte da base de dados do site preadly com o intuito de se obter uma classificação relevante para cada publicação através das palavras-chave extraídas automaticamente destas publicações, conforme será descrito neste trabalho.

3.2 Procedimento utilizado pra extrair as palavras-chave.

Para que análise descrita no Capítulo 3.1 fosse executada, foi utilizada a o método de extração de palavras-chave, Keyword Extraction, em java da API Alchemy IBM, descrita no capítulo 2 para que fosse feita a extração das palavras-chave automaticamente de cada um das publicações do site preadly. Com isso, todo esse

processo foi dividido em um processo sequencial de 4 partes que são mostrados no diagrama de blocos da figura 12 a seguir:

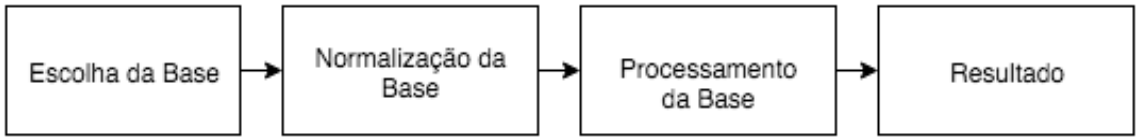


Figura 12 Processo Sequencial Realizado

3.2.1 Escolha da Base

Mas para iniciar o processo de extração dessas palavras-chave é preciso dispor-se de uma base de dados confiável. Que além de conter dados consistentes, ela deve conter preferencialmente o maior número possível de informações úteis que represente o contexto da análise.

A base de dados inicialmente pode estar disponível na forma de um banco de dados relacional, planilha local, ou armazenada na núvem. Mas para que o programa desenvolvido consiga processá-la, independente de qual seja a fonte, ela deve ser exportada para o formato .csv, onde os atributos essenciais que a base de dados deve conter, nas 2 primeiras colunas, são: id da publicação(como chave identificadora única) e URL válida da publicação. Portanto para que a base de dados possa ser processada é necessário ter-se ao menos essas 2 informações para que a aplicação desenvolvida possa processar a mesma.

3.2.2 Limpeza da Base

Nesse passo, ocorre uma avaliação destes dados, visando assegurar que todo o conteúdo selecionado não possui valores estranhos à semântica do contexto. Onde foi assegurado a não ocorrência de ids(chaves-primárias que identificam cada linha na base de dados) de publicação repetidas, para que não seja efetuado o processamento de um mesmo post duas vezes, ou que a coluna responsável pela URL não possuísse dados numéricos ou uma sequência de caracteres que constituíssem uma URL inválida e impedissem o processamento da aplicação.

Uma vez tendo ocorrido o tratamento inicial desses dados, o próximo passo é de submeter a base tratada e exportada em um arquivo no formato .csv para a aplicação em java desenvolvida, que irá efetuar os procedimentos para extrair as palavras-chave ordenadas.

3.2.3 Processamento da Base

Nesse passo, é feito o processamento dos dados pela aplicação desenvolvida para esse trabalho, que é descrita a parte na seção 3.3 deste capítulo, extraindo as palavras-chaves de cada uma das publicações da base.

3.2.4 Saída do Resultado.

Nesse passo é retornado pela aplicação, os resultados da base processada, tendo como saída um conjunto de palavras-chave em uma nova coluna na base, chamada de *Ranked Keywords* (Palavras-chave Ordenadas), e em um arquivo xml para cada arquivo HTML que for processado, onde temos exemplos dessas saídas na tabela 1 e no pedaço de xml a seguir.

Tabela 1 - Exemplo de Saída de Resultado.

Id do Post	Url	Ranked Keywords
55dde2dfec29047060000199	http://www.blogdaleoliveira.com/2015/08/look-jumpsuit.html	Amo estampas;
55dde333b825ae1d4c00008b	http://kkbeauty.com.br/2015/08/26/resenha-4-lips-da-felicitta-looks/	Rua Tijucas; Santa Catarina; marca felicitta;
55dde365b825ae16bd0000b6	http://thiswaybypallesen.com/sailing-in-south-of-france/	St. Tropez; amazing fantastic places; St. Tropez harbour;
55dc9158b6f8eb65700002d5	http://hotlovedrama.com/look-do-dia-say-my-name/	Cropped Oh; Saia Oh;
55dc9158b6f8ebedfa0002d6	http://www.blogdaleoliveira.com/2015/08/look-saia-jeans-com-botoes-frontal.html	Luau;
55dc918cb6f8eb62ce0003ff	http://produzir.me/2015/08/made-of-silver-saia-bandage/	black and grey; today look; saia bandage;

55dc914aec29048f5d000260	http://www.lauraperuchi.com/2015/08/5-mercados-gastronomicos-para-conhecer.html	Chelsea Market; City Kitchen; Gansevoort Market; Fifth Avenue; New York;
55dc3d37ec290419f30002ff	http://www.monalisadebatom.com.br/look-do-dia/print-dress/	pra academia; diário fitness; fotos fitness

Exemplo da saída em xml, obtida em uma das URLs da base de dados do site preadly*:*

?xml version="1.0" encoding="UTF-8" standalone="no"?><results>

<status>OK</status>

<usage>By accessing AlchemyAPI or using information generated by AlchemyAPI, you are agreeing to be bound by the AlchemyAPI Terms of Use: http://www.alchemyapi.com/company/terms.html</usage>

<url>55dc914aec29048f5d000260.html</url>

<totalTransactions>1</totalTransactions>

<language>portuguese</language>

<keywords>

<keyword>

<relevance>0.979938</relevance>

<text>Chelsea Market</text>

</keyword>

<keyword>

<relevance>0.74882</relevance>

<text>City Kitchen</text>

</keyword>

<keyword>

<relevance>0.747818</relevance>

<text>Gansevoort Market</text>
</keyword>
<keyword>
<relevance>0.747378</relevance>
<text>Fifth Avenue</text>
</keyword>
...
<keyword>
<relevance>0.509603</relevance>
<text>Gansevoort St</text>
</keyword>
</keywords>
</results>

3.3 Aplicação desenvolvida.

Utilizando o método de extração de palavras chave da Alchemy, citada no capítulo 2, e como a mesma possuí sua SDK em 9 linguagens de progrmação (Java, Perl, Ruby, Python, PHP, C/C++, C#, Node.js and Android OS), sendo uma delas java, a mesma foi escolhida pela ausência da necessidade de se instalar a aplicação que roda em qualquer sistema operacional, desde que tenha a JVM (Java Virtual Machine) instalada, onde se é necessário utilizar a JDK 8 ou superior do Java para execução dessa aplicação. Onde, dentre os métodos disponíveis por essa API foi utilizado o método *HTMLGetRankedKeywords,* que será descrito adiante, na seção 3.3.4.

A aplicação desenvolvida pode ser dividida basicamente em 5 partes sequenciais, representadas no diagrama de bloco da figura 13 a seguir:

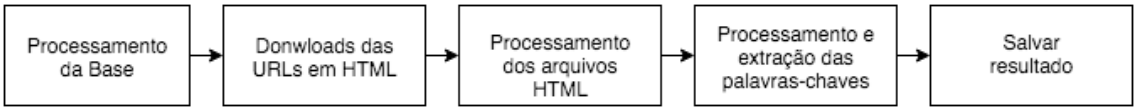


Figura 13 Procedimento Realizado pela Aplicação

3.3.1 Processamento da base

Nessa primeira parte é feita a escolha do documento .csv da base de dados já normalizada de acordo com o que foi descrito na seção 3.2.2 deste capítulo, onde é feita a leitura dessa base, que é transformado pela aplicação para o formato .xls com o intuito de se utilizar uma API, chamada de POI do Apache em java que contém métodos que permitem a melhor manipulação de arquivos .xls ou . xlsx do Microsoft Excel. Utilizando-se dos métodos da POI, a base de dados selecionada é processada, percorrendo-se linha a linha extraindo-se do mesmo uma lista de Urls, para que se possa efetuar o download do HTML das mesmas.

3.3.2 Download das publicações da base de dados em arquivos HTML

Recebida essa lista de URLs é executada uma rotina para efetuar o download de cada página web referente a essas publicações em documentos no formato HTML em uma pasta com o nome *data* no diretório raiz da aplicação. Essa abordagem foi adotada como uma alternativa ao processamento do conteúdo textual das publicações, que foi uma dificuldade encontrada, devido a grande quantidade de caracteres em formatação diferente, formando uma grande quantidade de textos com lixo armazenado na base de dados.

3.3.3 Processamento dos arquivos HTML das publicações

Após ser efetuado o Download desses documentos HTML, através de um procedimento em java, é feita uma leitura dentro da pasta *data* para que sejam extraídos todos os documentos no formato HTML a serem processados pelo servidor da Alchemy.

3.3.4 Processamento e extração das palavras-chave das publicações

Para utilizar a API da Alchemy primeiro é obtida uma instância da API com a chave que foi obtida, para fins de estudo apenas, com a Alchemy, que permite até 1000 chamadas ao servidor diariamente e, vale ressaltar, que cada vez que o método de extração de palavras-chaves é utilizado, cada utilzação conta como uma chamada ao servidor da Alchemy API.

Para se obter as palavras chaves no documentos HTML a serem processados, foi

utilizado o método **HTMLGetRankedKeywords**, que é descrito abaixo:

- ***HTMLGetRankedKeywords:***

O método HTMLGetRankedKeywords é usado para extrair uma lista ordenada (Pela Relevância dos pesos TF-IDF) de palavras-chave de tópicos de um documento HTML postado. AlchemyAPI irá extrair o texto a partir da estrutura do documento HTML e realizar operações para a extração dessas palavras-chave.

3.3.5 Salvar resultado

Para este trabalho, no entanto, foi utilizada parte dos recursos dessa API, onde como parâmetros para o método *HTMLGetRankedKeywords* foram passados apenas o documento HTML no formato de string, e nome do arquivo HTML que representa o id da publicação, e como saída desse método, temos as palavras-chave ordenadas que foram escolhidas a serem salvas no formato .xml na pasta data do diretório raiz e inseridas num arquivo .xls com a base de dados, e uma nova coluna chamada *Ranked Keywords* na pasta output também no diretório raiz.

3.4 Execução na base de dados do site pread.ly

Após o tratamento dos dados ter sido executado de acordo com o especificado e ter se assegurado que as informações estão posicionadas de forma correta, conforme descrito na secção 3.2, é necessário submeter ela na aplicação. Podemos ver um exemplo de como fazer isso na figura 14 a seguir, que mostra como é a tela de seleção da base de dados da aplicação desenvolvida.

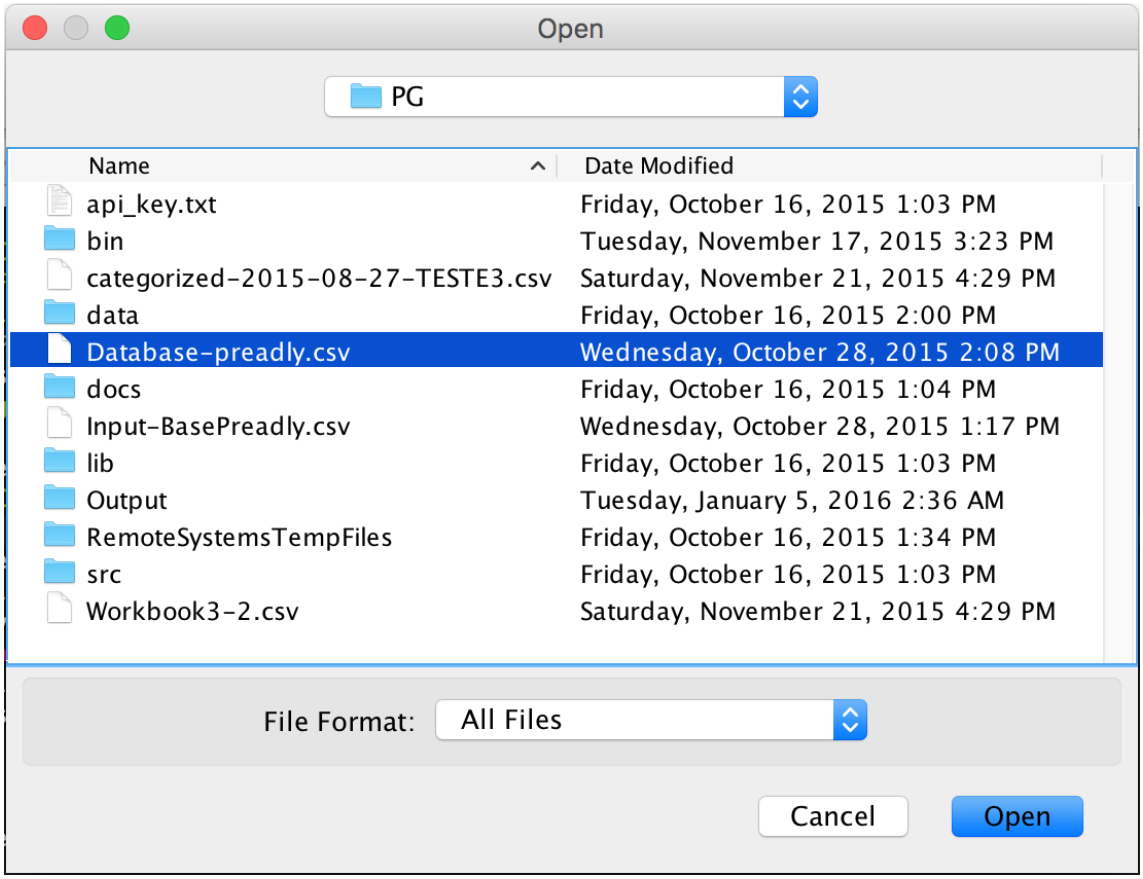


Figura 14 Tela de Seleção da Base de Dados

Com isso a aplicação irá extrair as URLs conforme citado na secção 3.3 desse capítulo e irá realizar o download dos arquivos HTML na pasta "data", onde irá efetuar o processamento dessas URLs usando-se um método da API da Alchemy, que irá retornar a saída em um arquivo .xml identificado pelo nome da publicação na pasta "data" e um arquivo .xls com o mesmo nome do arquivo da base de dados selecionada na pasta “output”.

4 Análise de Resultados

Esta seção visa apresentar e discutir os resultados encontrados, para a geração das palavras-chave. Os problemas encontrados e suas possíveis soluções.

4.1 Palavras-Chave geradas através da base de dados do site Preadly.

Aplicando a abordagem proposta por este trabalho na base de dados da Preadly, foram obtidos os seguintes resultados:

Tabela 2 - Amostra do Resultado Obtido na Base da Preadly

	Url	Palavras-Chave Ordenadas	Relevância Palavras-chave Alchemy (0 a 1)	TagsSlugs
1	http://www.blogdaleoliveira.com/2015/08/look-jumpsuit.html	Amo estampas;	0.999087	malacco, sammydress, c-s buzios
2	http://kkbeauty.com.br/2015/08/26/resenha-4-lips-da-felicitta-looks/	Rua Tijucas; Santa Catarina; marca felicitta;	0.925881; 0.918329; 0.828001;	lipstick
3	http://thiswaybypallesen.com/sailing-in-south-of-france/	St. Tropez; amazing fantastic places; St. Tropez	0.955149; 0.917207; 0.893709	saint-tropez, nice

		harbour;		
4	http://hotlovedrama.com/look-do-dia-say-my-name/	Cropped Oh; Saia Oh;	0.931395; 0.918137;	oh-boy, whoop
5	http://www.blogdaleoliveira.com/2015/08/look-saia-jeans-com-botoes-frontal.html	Luau;	0.975727;	luau, oasap
6	http://produzir.me/2015/08/made-of-silver-saia-bandage/	black and grey; today look; saia bandage;	0.918629; 0.733449; 0.605034;	a-colorida, anne
7	http://www.lauraperuchi.com/2015/08/5-mercados-gastronomicos-para-conhecer.html	Chelsea Market; City Kitchen; Gansevoort Market;	0.979938; 0.74882; 0.747818;	chelsea-market, gansevoort-market, city-kitchen
8	http://www.monalisadebatom.com.br/look-do-dia/print-dress/	pra academia; vontade pra; diário fitness;	0.967305; 0.774762; 0.595571;	arezzo, lez-a-lez

Externados na forma de uma tabela com 4 colunas, onde:

- ***URL** representa o link para o post;*

- *A coluna **Palavras-chave ordenadas** representa as palavras obtidas com a aplicação e é externada de forma ordenada por sua relevância e separada pelo delimitador " ; ".*
- *A coluna **Relevância Keywords Alchemy (0 a 1),** representa o valor de relevância das palavras-chaves da segunda coluna, também de forma ordenada e delimitada por ";"*
- *A coluna **TagSlugs** representa as palavras-chaves já pré-existentes na base da preadly.*

4.2 Análise das Palavras-chave.

Para analisar os resultados obtidos com a aplicação foram feitas 3 análises principais em cima do resultado final apresentado anteriormente, que foram posteriormente externadas em uma tabela com 5 colunas:

- ***URL,** similar a da Tabela 1 representa o link para o post;*
- *A coluna **Ranked Keywords,** similar a da Tabela 1 representa as palavras obtidas com a aplicação e é externada de forma ordenada por sua relevância e separada pelo delimitador " ; ".*
- *A coluna **Relevância Keywords Alchemy (0 a 1),** representa o valor de relevância das palavras-chaves da segunda coluna, também de forma ordenada e delimitada por " ; "*
- *A coluna **Relevância Baseada no conteúdo do texto (0 a 1)** representa o número de palavras-chaves que foram consideradas relevantes pelo Autor ao conteúdo presente na publicação.*
- *A coluna **Relevância Baseada na presença das palavras-chaves sugeridas nas palavras chaves pré-existentes (0 a 1),** verifica o numero de palavras-chaves geradas que estão presentes na coluna **TagSlugs,** descrita anteriormente.*

Url	Palavras-chave ordenadas	Relevância Keywords Alchemy (0 a 1)	Relevância Baseada no conteúdo do texto (0 a 1)	Relevância Baseada na presença das palavras-chaves sugeridas nas palavras chaves pré-existentes (0 a 1)
http://www.blogdaleoliveira.com/2015/08/look-jumpsuit.html	Amo estampas;	0.999087;	1/1=1	0/3=0
http://kkbeauty.com.br/2015/08/26/resenha-4-lips-da-felicitta-looks/	Rua Tijucas; Santa Catarina; marca felicitta;	0.925881; 0.918329; 0.828001;	1/3=0.33	0/1=0
http://thiswaybypallesen.com/sailing-in-south-of-france/	St. Tropez; amazing fantastic places; St. Tropez harbour;	0.955149; 0.917207; 0.893709;	3/3=1	1/2=0.5
http://hotlovedrama.com/look-do-dia-say-my-name/	Cropped Oh; Saia Oh;	0.931395; 0.918137;	2/2=1	0/2=0
http://www.blogdaleoliveira.com/2015/08/look-saia-jeans-com-botoes-frontal.html	Luau;	0.975727;	1/1=1	1/2=0.5
http://produzir.me/2015/08/made-of-silver-saia-bandage/	black and grey; today look; saia bandage;	0.918629; 0.733449; 0.605034;	2/3=0.66	0/2=0
http://www.lauraperuchi.com/2015/08/5-mercados-gastronomicos-para-conhecer.html	Chelsea Market; City Kitchen; Gansevoort Market;	0.979938; 0.74882; 0.747818;	3/3=1	3/3=1

http://www.monalisadebatom.com.br/look-do-dia/print-dress/	pra academia; vontade pra; diário fitness;	0.967305; 0.774762; 0.595571;	2/3=0.66	0/2=0
Média Aritmética	-	0.81	0.79	0.29

Tabela 3 - Tabela com valores da Análise dos Resultados Obtidos na Tabela 2.

Onde A seguir, tem-se a descrição das análises realizadas:

4.2.1 Relevância Keywords Alchemy

Essa primeira análise foi para se obter a média da relevância das palavras extraídas pela aplicação, que ao se somar os 8 resultados obteve-se um total de 15.47 em 19, que totaliza uma média aritmética de 0.81 ou 81%.

Esse valor nos diz que a aplicação foi capaz de retornar palavras-chave com uma relevância média de 81% para as publicações analisadas. E como a designação de palavras-chave é um pouco subjetiva e nem sempre muito bem feita, essa extração automática melhora essa tarefa.

4.2.2 Relevância Baseada no conteúdo do texto.

Essa análise levou em conta a relevância das palavras-chaves geradas pela aplicação para o texto, onde se obteve um total de 15 palavras relevantes dentre um total de 19 palavras geradas.

Com isso, os resultados mostrados na tabela do capítulo 5.1 mostram que as palavras-chave geradas poderiam ser utilizadas para a classificação, tendo em vista que a média obtida foi de 15/19 = 0.79 ou 79%, mostrando uma possibilidade de sugerir em quase 79% das vezes uma palavra-chave relevante ao texto, que muitas vezes em um primeiro momento, um input manual poderia não associar aquela palavra-chave a publicação, ajudando nesse processo de classificação de publicações.

4.2.3 Relevância Baseada na presença das palavras-chaves sugeridas nas palavras chaves pré-existentes.

Essa análise visa identificar se houve a ocorrência ou não das palavras-chave sugeridas pela aplicação em relação com às palavras-chave já pré-existentes na base.

Onde o resultado obtido foi de 5 ocorrências em 17, que deu uma média aritmética de 5/17 = 0.29 ou 29%.

Essa análise mostra que as palavras-chave obtidas com a aplicação não possuem grande similaridade com as palavras-chave pré-existentes na base de dados, o que apesar de ser um número baixo, e parecer uma média ruim, esse valor representa justamente o fato de que as palavras extraídas pela aplicação desenvolvida geram um novo conhecimento para a base, e essas palavras-chave extraídas mostram com a análise efetuada nas seções 4.2.1 e 4.2.1 que a relevância dessas palavras-chave para o texto são muito altas, mostrando portanto um resultado promissor nessa categorização.

5 **Conclusão**

Tendo seguido todos os passos e técnicas abordados anteriormente neste trabalho, a saída do processo consiste em uma forma automática de se classificar um texto através da utilização de palavras-chaves. Este capítulo reserva-se a analisar o produto final.

5.1 Considerações finais.

Palavras-chave são uma alternativa de grande utilidade na categorização de textos. Como foi dito anteriormente, tendo em vista ser inviável a leitura de cada texto de forma manual com o intuito de categorizar o mesmo, Motivado por este fato, a grande contribuição desse projeto final é uma proposta de abordagem para automatizar esse processo de extração.

Para o fim deste trabalho, foi desenvolvida uma aplicação em java, que consegue através de uma base de dados de publicações e suas respectivas URLs, efetuar o carregamento dessas URLs e através da abordagem proposta nesse trabalho, extrair de publicações palavras-chaves que categorizam o conteúdo presente nessas páginas, para solucionar um problema real no âmbito da extração automatizada de palavras-chaves que foi apresentada pelo site www.preadly.com, que alegou não conseguir extrair de uma forma eficiente palavras-chave que categorizem o conteúdo de suas publicações.

Sendo assim foi possível abordar esse problema em uma base de dados real e propor uma alternativa a ele.

5.1 Problemas e limitações enfrentados.

A primeira grande dificuldade apresentada foi na etapa de extração das informações da base. A existência de muitos campos com informações irrelevantes ou com “lixo” precisou ser retirada, o que foi feito na etapa de limpeza da base.

A segunda dificuldade se deu, devido à primeira abordagem idealizada para o trabalho ter de ser abandonada, pois ler os posts diretamente da base não foi possível, devido a uma grande variedade de caracteres especiais salvos na base, ou seja, com uma formatação errada, além de diferente formatações entre eles, não permitindo portanto que a API da Alchemy identificasse a linguagem dos posts, fornecendo resultados não aceitáveis.

A solução encontrada foi a de ao invés de ler as publicações na base, processar as URLs e efetuar o download dos HTMLs para que se pudesse extrair as palavras-chave dos documentos.

5.2 Trabalhos Futuros

Trabalhos futuros incluem o de desenvolver um método que possa ser implementado em conjunto com aprendizado de máquina, para que se possa treinar um algoritmo com um conjunto de palavras relevantes diferentes para cada tema em que o método proposto seja aplicado. Tendo como intuito desenvolver uma ferramenta web, que permita que cada texto escrito gere automaticamente uma lista de palavras-chaves associado a ele, assim categorizando automaticamente cada texto que for escrito.

Para essa proposta futura que visa estender o conceito desse trabalho, pode-se aplicar outras técnicas presentes na literatura, como, incluir uma análise utilizando o algoritmo classificador de Naive Bayes[15] que é um classificador probabilístico simples, baseado na aplicação do teorema de Bayes com fortes (ingênuos) hipóteses de independência entre seus atributos, e é um dos classificadores mais utilizados em *Machine Learning*.

Com esse classificador, seria possível formar um dicionário de palavras baseado no modelo de *bag-of-words* (saco-de-palavras) *[8]*, que é um método de classificação de documentos em que a frequência de ocorrência de cada palavra, é usada como um recurso para a formação de um conjunto classificador, onde em conjunto com o classificador de Bayes, pode ser utilizado para se treinar o algoritmo no contexto em que o texto será aplicado, para efetuar uma melhor classificação dos textos.

Referências Bibliográficas

[1] Bielenberg, K.Zacher, M. Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation. Master, Universität Bremen, 2006.

[2] CNET,. Facebook processes more than 500 TB of data daily - CNET. Disponível em: <http://www.cnet.com/news/facebook-processes-more-than-500-tb-of-data-daily/>. Acesso em: 6 nov. 2015.

[3] Dalkescientific.com,. Calvin Mooers. Disponível em: <http://www.dalkescientific.com/writings/diary/archive/2014/06/19/Calvin_Mooers.html>. Acesso em: 6 jan. 2016.

[4] Frank, EibeBouckaert, Remco R. Naive Bayes for Text Classification with Unbalanced Classes. Lecture Notes in Computer Science, p. 503-510, 2006.

[5] Wikipedia,. Document Classification. Disponível em: <https://en.wikipedia.org/wiki/Document_classification>. Acesso em: 6 nov. 2015.

[6] Luhn, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, v. 1, n. 4, p. 309-317, 1957.

[7] Manning, Christopher D, Raghavan, PrabhakarSchütze, Hinrich. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.

[8] Harris, Zellig S. Distributional Structure. Papers on Syntax, p. 3-22, 1981

[9] Pt.wikipedia.org,. Palavra-chave. Disponível em: <https://pt.wikipedia.org/wiki/Palavra-chave>. Acesso em: 6 nov. 2015.

[10] Russell, Stuart J. (Stuart Jonathan)Norvig, Peter. Artificial intelligence a modern approach. New Jersey: Prentice Hall, 2003.

[11] SPARCK JONES, KAREN. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. Journal of Documentation, v. 28, n. 1, p. 11-21, 1972.

[12] Tf-idf,. A Single-Page Tutorial - Information Retrieval. Disponível em: <http://www.tfidf.com>. Acesso em: 6 nov. 2015.

[13] W3.org,. RDF - Semantic Web Standards. Disponível em: <http://www.w3.org/RDF/>. Acesso em: 6 nov. 2015.

[14] Wikipedia,. Bag-of-words model. Disponível em: <https://en.wikipedia.org/wiki/Bag-of-words_model.>. Acesso em: 6 nov. 2015.

[15] Eibe Frank,Remco R. Bouckaert Naive Bayes for Text Classification with Unbalanced Classes.

[16] Zhiyuan Liu, Xinxiong Chen, Maosong Sun Mining the interests of Chinese microbloggers via keyword extraction Journal, Frontiers of Computer Science in China archive, Volume 6 Issue 1,Pages 76-87, February 2012.

[17] Salton, GerardBuckley, Christopher. Term-weighting approaches in automatic text retrieval. Information Processing & Management, v. 24, n. 5, p. 513-523, 1988.

[18] Shin, Yongwook, Chuhyeop Ryo, and Jonghun Park. "Automatic Extraction Of Persistent Topics From Social Text Streams". *World Wide Web* 17.6 (2013): 1395-1420. Web. 8 Jan. 2016.