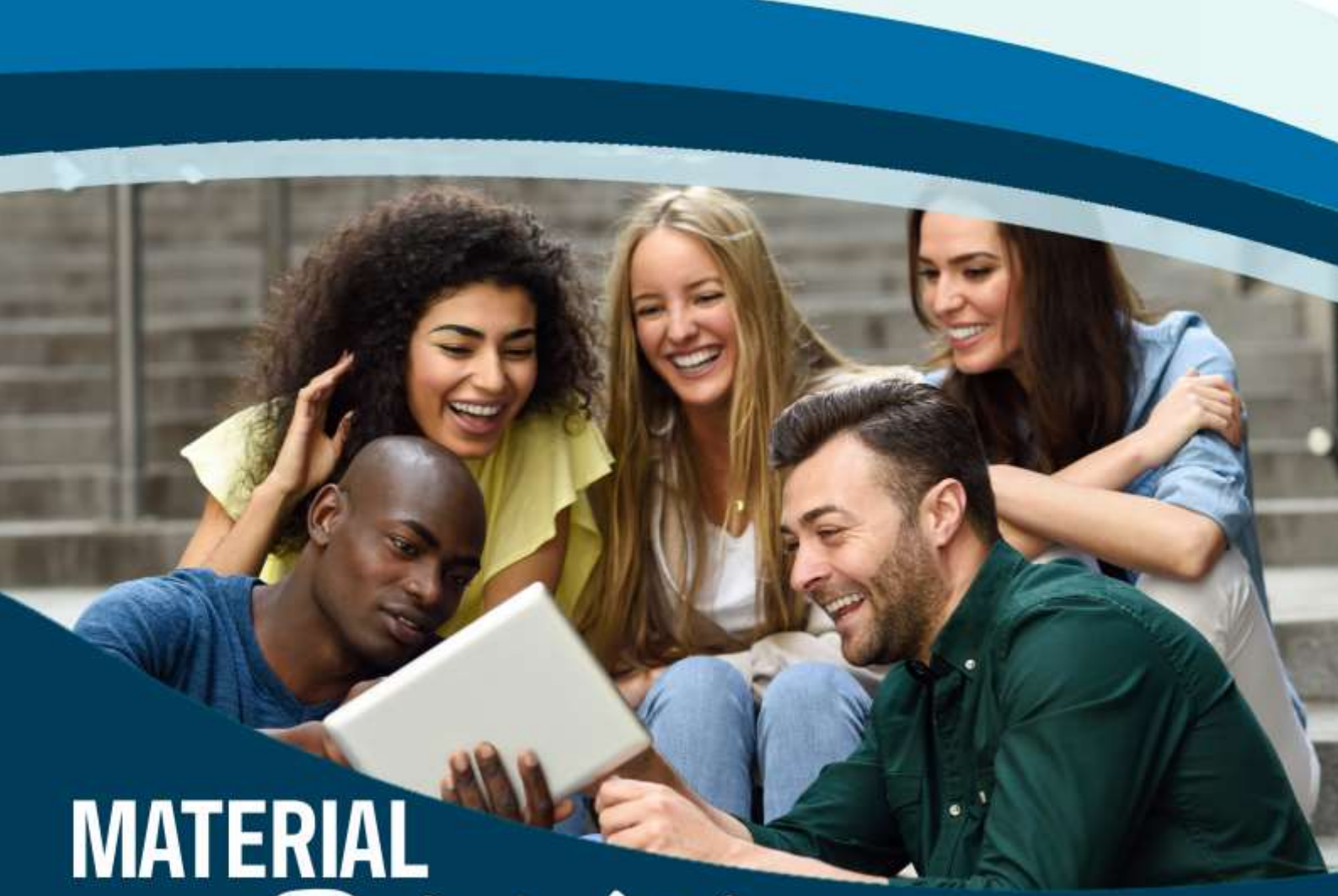


MATERIAL

Didático

MINERAÇÃO DE TEXTOS EM R

FACUMINAS

CREDENCIADA JUNTO AO MEC PELA PORTARIA
N 3.455 DO DIA 19/11/2003

www.facuminasead.com.br 31 3842-3838

FACUMINAS

SUMÁRIO

FACUMINAS

A história do Instituto Facuminas, inicia com a realização do sonho de um grupo de empresários, em atender a crescente demanda de alunos para cursos de Graduação e Pós-Graduação. Com isso foi criado a Facuminas, como entidade oferecendo serviços educacionais em nível superior.

A Facuminas tem por objetivo formar diplomados nas diferentes áreas de conhecimento, aptos para a inserção em setores profissionais e para a participação no desenvolvimento da sociedade brasileira, e colaborar na sua formação contínua. Além de promover a divulgação de conhecimentos culturais, científicos e técnicos que constituem patrimônio da humanidade e comunicar o saber através do ensino, de publicação ou outras normas de comunicação.

A nossa missão é oferecer qualidade em conhecimento e cultura de forma confiável e eficiente para que o aluno tenha oportunidade de construir uma base profissional e ética. Dessa forma, conquistando o espaço de uma das instituições modelo no país na oferta de cursos, primando sempre pela inovação tecnológica, excelência no atendimento e valor do serviço oferecido.

ÁREAS DE CONHECIMENTO EM MINERAÇÃO DE TEXTOS: RECUPERAÇÃO DE INFORMAÇÃO

Recuperação de Informação (RI) é uma área da computação que estuda como armazenar e recuperar dados, geralmente textos e, de forma automática.

MOOERS (1951) cunhou o termo recuperação da informação, destacando que ele "engloba os aspectos intelectuais da descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação". O processo de recuperação da informação baseia-se em técnicas tradicionais da ciência da computação e, em recursos óbvios dos dados para criar estruturas de índices, organizar e recuperar de forma eficiente as informações. Essa estrutura de índice permite identificar, no conjunto de documentos (corpus) de um sistema, quais atendem à necessidade de informação do usuário.

As técnicas de recuperação da informação estão intimamente relacionada com a mineração de textos, principalmente no processo de Indexação, em que são montados filtros para eliminar palavras de pouca significação (stop words) , além de normalizar os termos reduzindo-os a seus radicais, processo conhecido como stemming

Um sistema tradicional de Recuperação de Informação pode ser estruturado conforme ilustrado na Figura 1.

Figura 1 – Componentes de um sistema de Recuperação de Informação

O Processo de Indexação cria estruturas de dados ligadas à parte textual dos documentos, por exemplo, as listas invertidas. As listas invertidas são ordenadas por tipo , em que as chaves são termos do vocabulário da coleção e os valores são listas de referências para documentos.

O Processo de Especificação da Busca na maioria dos casos é uma tarefa difícil. "Há frequentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada". Essa distância é gerada pelo limitado conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta (BAEZA-YATES, 1999).

O Processo de Recuperação gera uma lista de documentos respondendo à consulta formulada pelo usuário. Os índices especificados para aquela coleção de documentos são utilizados para acelerar o procedimento.

APRENDIZAGEM DE MÁQUINA

Aprendizado de Máquina estuda a criação de modelos algoritmos probalísticos capazes de “aprender” através da experiência. O aprendizado se dá através de métodos dedutivos para extração de padrões em grandes massas de dados Chakrabarti (2002). ML (do inglês, Machine Learning) tem sido muito utilizado no processo de classificação automática de textos.

Segundo Mitchell (1997) o aprendizado de máquina estuda como os algoritmos computacionais são capazes de automaticamente melhorarem a execução de tarefas através da experiência. Os algoritmos desenvolvidos sobre a aprendizagem de máquina se baseiam na estatística e probabilidade para aprender padrões complexos a partir de algum corpus.

Na literatura encontramos diversos trabalhos interdisciplinares que aplicaram os algoritmos de aprendizado de máquina. Com exemplo, o trabalho de correção ortográfica Schmid (1994) e o de diagnóstico de doenças Bair e Tibshirani (2003). A aplicabilidade desse algoritmo rendeu bons resultados, muito devido a sua utilização em diversas áreas de natureza diferenciadas. Os casos de sucesso do aprendizado de máquina nos leva a acreditar que esse tipo de algoritmo é fortemente relacionado à área de mineração de textos, estendendo-se a classificação automática de textos.

PROCESSAMENTO DE LINGUAGEM NATURAL

Devido a escalabilidade dos gerenciadores de bancos de dados em armazenar informações, diversos sistemas conseguiram manter disponíveis textos em formato de documentos sem problemas de demanda, acesso e disponibilidade

dos dados. Todavia, com o aumento exponencial de documentos circulando em diversos tipos de sistemas, mesmo os computadores modernos podem não comportar essa massa de dados, tendo que restringir a representação à um conjunto limitado de termos.

Além disso, o que os usuários necessitam é representado por uma expressão de busca, que pode ser especificada em linguagem natural. Mas, esse mecanismo de expressão de busca traz dificuldades para a maioria dos usuários, pois eles têm que predizer as palavras ou expressões que satisfaçam sua necessidade.

O Processamento de Linguagem Natural (PLN) surge então, para resolver problemas relacionados à recuperação da informação, ao observar que os documentos e as expressões de busca são apenas objetos linguísticos. Através dessa observação, criou-se várias técnicas dentro da PLN para análisar textos em um ou mais níveis linguísticos, com intuito de emular o processamento humano da língua.

O PLN é uma área de Ciência da Computação que estuda como os computadores podem analisar e/ou gerar textos em linguagem natural (PERNA; DELGADO; FINATTO, 2010). Turban et al. (2010) descreve que o processamento da linguagem natural pode ser vista como a forma de comunicação entre o homem e a máquina, sendo essa comunicação em qualquer linguagem que se fale. Os autores ainda dizem que:

> Para entender uma consulta em linguagem natural, o computador precisa ter conhecimento para analisar e interpretar a entrada de informação. Isso pode significar conhecimento lingüístico de palavras, conhecimento sobre áreas específicas, conhecimentos gerais e até mesmo conhecimento sobre os usuários e seus objetivos. No momento em que o computador entende a informação, ele pode agir da forma desejada (TURBAN et al., 2010).

Para Lopes (2002) O PLN não é uma tarefa trivial devida a natureza ambígua da linguagem natural. Essa diversidade faz com que o PLN difere do processamento das linguagens de programação de computador, as quais são fortemente definidas para evitar a ambiguidade. Ainda este autor classifica as técnicas de PLN conforme o nível linguístico processado: fonológico, morfológico, lexical, sintático, semântico e

pragmático. Estes níveis precisam ser entendidos e diferenciados. Especificamente, o morfológico que trata das palavras isoladamente, o léxico que trabalha com o significado das palavras, o sintático que se refere a estrutura das frases, o fonológico que lida com a pronúncia, o semântico que interpreta os significados das frases (LIDDY, 2001).

Todas essas técnicas podem ser utilizadas em um processo de PLN, contudo, para o presente trabalho, o nível fundamental é o morfológico. O analisador morfológico tem o propósito de selecionar as palavras e expressões que encontra-se de maneira isolada no texto. É importante ressaltar que existem técnicas dentro da PLN que não são aplicáveis a Mineração de Textos, como exemplo, as correções ortográficas e a tradução automática de textos (JUNIOR, 2007).

MINERAÇÃO DE TEXTOS

Graças ao desenvolvimento da Internet e das redes de computadores os documentos virtuais se transformaram no principal método de armazenamento de informações, principalmente as informações comerciais, que segundo estimativas armazena cerca de 85% de suas informações em documentos de texto. Porém, os paradigmas mais tradicionais de desenvolvimento de software não são capazes de entender o relacionamento confuso e geralmente ambíguo que existe nos documentos de texto virtuais (MACHADO et al., 2010).

O principal objetivo da Mineração de Textos (MT) consiste na extração de características em uma grande quantidade de dados não estruturados. Segundo Weiss et al. (2010) as técnicas utilizadas na mineração de textos são semelhantes as utilizadas na mineração de dados, ou seja, fazem o uso dos mesmos métodos de aprendizagem, independente se uma técnica utiliza-se de dados textuais (MT) e a outra com dados numéricos (MD).

A mineração de textos é um paradigma de programação criado para resolver este problema, sendo capaz de entender a linguagem natural dos documentos de texto e conseguindo lidar com a sua imprecisão e incerteza. A mineração de textos envolve várias áreas da informática, como mineração de dados, aprendizado de máquina, recuperação de informação, estatística e linguagem computacional, para conseguir transformar o texto em algo que um computador consiga entender (MACHADO et al., 2010).

O principal objetivo da mineração de textos é encontrar termos relevantes em documentos de texto com grande volume de dados e estabelecer padrões e relacionamentos entre eles com base na frequência e temática dos termos encontrados (SERAPIÃO, 2010). A tecnologia de mineração de textos não é um mecanismo de busca, pois a mineração ajuda o usuário a descobrir informações previamente desconhecidas, enquanto na busca o usuário já sabe o que deseja procurar. Além disso, a mineração também é diferente de robôs de conversação (chatterbot), pois ela não tenta simular o comportamento humano (ARANHA; PASSOS, 2006).

O processo de minerar dados é composto por um conjunto de técnicas baseadas em modelos capazes de encontrar padrões, sumarizar dados, extrair novos conhecimentos ou realizar predições com o objetivo de descobrir informações com base em grandes volumes de dados (SILVA et al., 2013). Mineração de textos, ou em em inglês TextMining, pode ser considerada como uma extensão ou sub-área da Mineração de Dados (RODRIGUES, 2016). Segundo Tan et al. (1999), refere-se a um processo de extração de conhecimento, padrões interessantes e não triviais de documentos de texto.

CLASSIFICAÇÃO DE TEXTOS

Um dos motivos para o crescente interesse no estudos sobre a área da mineração de textos, especificamente na ténica de classificação, é devido ao

crescimento e a disponibilidade de documentos na internet, sobretudo pelas redes sociais.

A técnica dominante para este problema é baseada na aprendizagem de máquina, ou seja, um processo indutivo cria automaticamente um classificador por "aprendizado", a partir de um conjunto de dados classificados previamente. A vantagem dessa abordagem é a independência de domínio (SEBASTIANI, 2002).

Pode-se dizer então que a tarefa de classificar um texto automaticamente é uma derivação da aprendizagem de máquina com o propósito de atribuir rótulos pré-definidos a documentos textuais. Categoricamente Sebastiani (2002) assegura que a classificação de textos consiste em determinar se um documento di, (de um conjunto de documentos D) é pertencente ou não a uma categoria cj (de um conjunto de categorias C), consistentemente com o conhecimento das categorias corretas para um conjunto de documentos de treinamento.

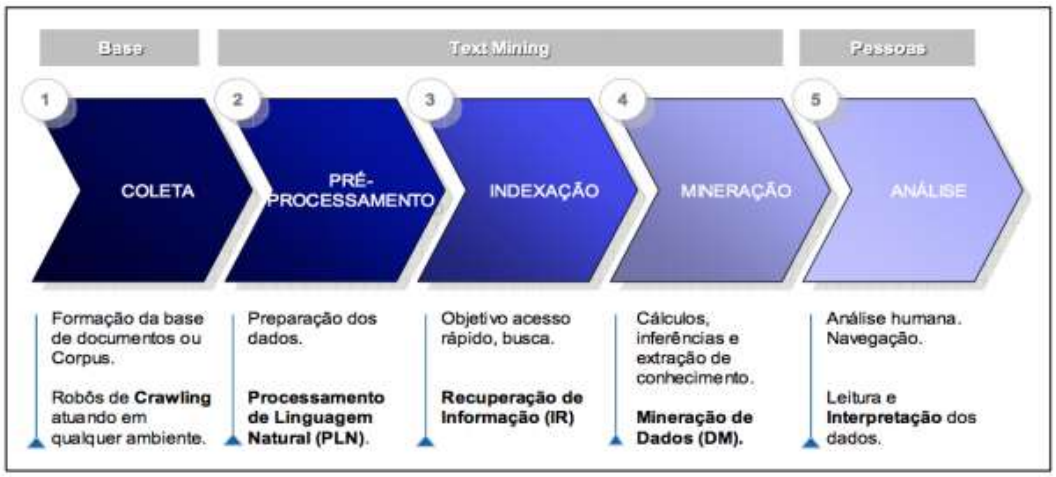
O objetivo principal da classificação é atribuir uma determinada classe à uma conjunto de documentos e, no caso da análise de sentimentos, trata-se de classificar automaticamente um conjunto de dados às classes positivas e negativas.

ETAPAS DA MINERAÇÃO DE TEXTOS

Neste capítulo apresentaremos a metodologia proposta por Aranha e Vellasco (2007) para Mineração de Textos. Em seu trabalho, Aranha descreve como sendo um modelo completo para adquirir conhecimentos a partir de um corpus textual. O objetivo deste capítulo é detalhar todas as etapas e técnicas desta metodologia, uma vez que este processo é o que melhor se enquadra no presente trabalho.

A Figura 2 apresenta segundo Aranha, Vellasco e Passos (2007), um modelo de processo de Mineração de Textos, do início ao fim, muito comum entre vários trabalhos da literatura. Tal processo é semelhante com os apresentados na seção anterior.

Figura 2 – Processo de Mineração de Texto



Fonte: Aranha, Vellasco e Passos (2007)

Mathiak e Eckstein (2004), detalha cada uma dessas etapas. A etapa de coleta é basicamente onde são coletados os dados que serão analisados. Esses dados podem ser coletados em sites, sistemas corporativos, redes sociais entre outros. O pré-processamento tem por finalidade melhorar a qualidade dos dados já disponíveis e organizá-los. A indexação é a fase onde são extraídos os conceitos dos documentos por meio da análise de seu conteúdo. Na fase de mineração é onde são aplicadas técnicas para a extração do conhecimento, tal como clusterização. Por fim, é realizada a análise e interpretação dos dados pela pessoa responsável, procurando padrões e/ou analisando de forma manual os resultados da etapa de mineração.

Extração

Na mineração de textos, quando estamos diante de um problema de classificação automática de documentos, faz-se necessário obter um conjunto de

dados para treinamento Aranha e Vellasco (2007). Portanto, a etapa de extração e coleta de dados tem como função a criação de uma base de dados textual.

Segundo Manning et al. (2008) a coleta poderá ser realizada utilizando-se de crawlers. Crawler é um software que percorre sítios da internet como intuito de coletar automaticamente os dados destes. Após a recuperação destes dados pretendidos para a análise, é possível criar um corpus que servirá de base para aplicar as técnicas de mineração de textos.

Um corpus nada mais é que uma coleção de textos, que representa uma ou um conjunto de linguagens naturais e, a criação deste conjunto de treino revela-se uma tarefa custosa, uma vez que na maioria dos casos exige-se processos manuais à base expert judgment (INDURKHYA; DAMERAU, 2010).

Pré-Processamento

Pré-processamento é a etapa executada imediatamente após a coleta dos dados. Pré-processar textos é, na maioria das vezes, uma etapa muito onerosa, uma vez que utiliza-se diversos algoritmos que consomem boa parte do tempo do processo de extração de conhecimento e, por não existir uma única técnica que possa ser aplicada em todos os domínios de aplicações.

O principal objetivo de pré-processar um texto, consiste na filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias para o conhecimento que se deseja extrair (GONÇALVES et al., 2006).

Mineração

Na etapa de Mineração são aplicadas técnicas direcionadas ao aprendizado de máquina (Machine Learning - ML) para obtenção de novos conhecimentos Witten e Frank (2011). Nessa etapa escolhemos qual tarefa de acordo com a necessidade do usuário. Por exemplo, se a necessidade for verificar o grau de similaridade e a formação de grupos naturais, então a tarefa a ser escolhida é clusterização. Por outro lado, se estes grupos já estão formados, seja por conhecimento prévio do especialista ou pela execução de algoritmos, então a orientação de aonde um novo documento deve ser “rotulado” é conseguida através de algoritmos de classificação.

No contexto deste trabalho, as técnicas aplicadas na etapa de mineração, devem ser capazes de identificar as características que diferenciam documentos pertencentes a diferentes classes e, realizar o processo de classificação.

Interpretação

A etapa da interpretação dos dados é onde será validada a eficiência do processo como um todo, analisando os dados obtidos após aplicação dos algoritmos na etapa anterior. Em outras palavras, é nesta etapa que avaliamos se o objetivo de descobrir um novo conhecimento foi adquirido, a partir de uma base de dados

Por fim, vale ressaltar que este processo é cíclico. Ao final de cada uma das etapas, os resultados devem ser analisados individualmente e caso não apresentem resultados satisfatórios, deve-se realizar alterações no processo para a realização de um novo ciclo.

TÉCNICAS UTILIZADAS

As principais técnicas utilizadas para fazer a mineração de textos são:

Processamento de Linguagem Natural: É um método que procura utilizar computadores para melhorar o entendimento da linguagem natural através de técnicas para processar textos rapidamente, utilizando-se de manipulação de strings até linguagem natural de inquéritos (MACHADO et al., 2010).

Recuperação de Informação: Utiliza métodos e medidas estatísticos ou semânticos para automaticamente processar o texto de documentos para encontrar quais documentos possuem a resposta para a questão (mas não a resposta em si). Embora já fossem utilizadas técnicas deste tipo de forma primitiva em 1975, este método só ganhou notoriedade com a popularização da Internet(MACHADO et al., 2010).

Extração de Informação: Possui como principal objetivo buscar partes relevantes de um texto em um documento e extrair informações específicas destas partes. Possui um conceito mais limitado da compreensão da linguagem natural(MACHADO et al., 2010).

Estas técnicas são vastamente utilizadas na mineração de dados, principalmente em redes sociais e em processos de ensino a distância (MACHADO et al., 2010).

Clusterização: Clusterização é uma técnica de mineração de dados que consiste na divisão dos dados em grupos de objetos com características semelhantes, ou que estejam próximos uns dos outros. A ideia é que objetos que estão no mesmo grupo sejam mais similares em comparação a objetos de grupos diferentes. Entre os principais algoritmos de clusterização estão: K-means (WU, 2008) e o DBSCAN (ESTER et al., 1996).

A técnica de clusterização pode ser usada em diversos contextos, como, por exemplo: a descoberta de perfis de clientes de uma empresa. Cada perfil pode ser um cluster. Os perfis podem ser formados por clientes de uma mesma faixa etária, e que costumam comprar o mesmo tipo de produto, por exemplo. A partir desse conhecimento uma empresa pode realizar campanhas publicitárias e/ou promoções voltadas para um grupo específico de clientes.

DBSCAN: O DBSCAN é um algoritmo baseado em densidade, não sendo necessário que o usuário defina a priori quantos clusters irá ter como resultado, o próprio algoritmo que define os clusters agrupando os dados que estiverem mais próximos para formarem um cluster. Cada região de densidade maior ou igual a um threshold dado de entrada se tornará um cluster(ESTER et al., 1996).

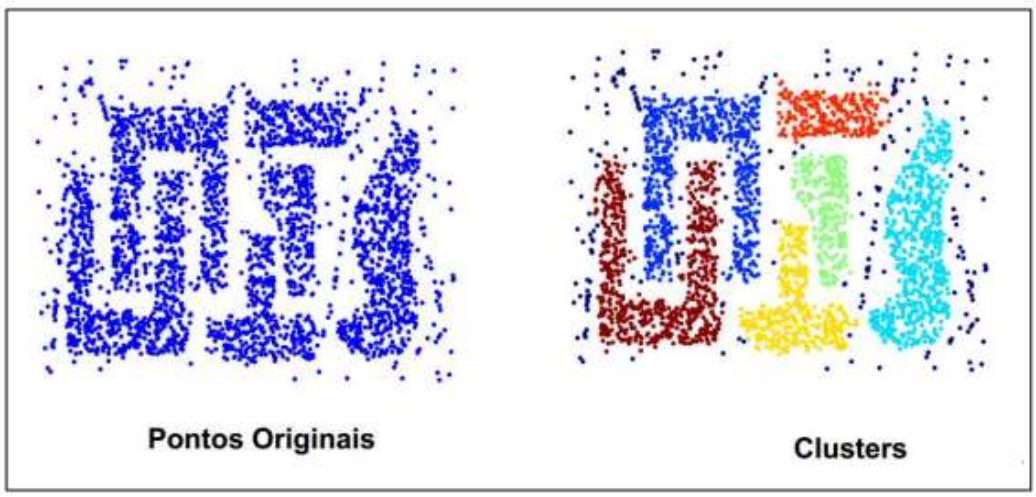
Para explicar o funcionamento do DBSCAN é necessário conhecer algumas definições:

- **eps:** A distância máxima entre dois pontos para serem considerados vizinhos.

- **minPts:** O número minimo de vizinhos que um ponto deve ter para ser considerado um corePoint.
- **corePoint:** É um ponto onde seu número de vizinhos é maior ou igual a minPts.
- **borderPoint:** Possui o número de vizinhos menor que minPts, mas é vizinho de um corePoint.
- **noisePoint:** Possui o número de vizinhos menor que minPts, e não possui um corePoint em sua vizinhança.

O Algoritmo inicia escolhendo um ponto aleatório X, analisa sua vizinhança e se possuir uma vizinhança maior que minPts, um novo cluster C é criado. Depois, o algoritmo percorre todos os pontos vizinhos de X, se algum desses pontos tiver uma vizinhança de pelo menos minPts esses pontos da vizinhança são inseridos no cluster C, e esse processo se repete até que não tenha mais pontos à adicionar neste cluster C. Depois de finalizado um cluster, o algoritmo pega um ponto que ainda não foi visitado e recomeça o processo. A Figura 3 mostra um exemplo de como o DBSCAN divide os dados em clusters.

Figura 3 – Divisão dos dados em clusters



Fonte: Tan et al. (2006)

Medida de Similaridade

KASZNAR e GONÇALVES (2009) falam que o processo de clusterização requer medidas de “proximidade” ou “similaridade”. Medidas de similaridade são métricas capazes de analisar um conjunto de características com o objetivo de comparar objetos (RODRIGUES, 2016).

A técnica utilizada neste trabalho é a similaridade de Jaccard, pois é uma medida simples e que pode ser aplicada no contexto desse trabalho. Tal medida já foi utilizada em trabalhos similares como em Rodrigues (2016) e Yin et al. (2012). Segundo Russell (2013), a similaridade de Jaccard expressa a similaridade entre dois conjuntos como a interseção dos conjuntos dividida pela união deles. Matematicamente, a similaridade de Jaccard é escrita como:

$$J(p^L, b^L) = \frac{|p^L \cap b^L|}{|p^L \cup b^L|}$$

A similaridade de Jaccard resulta em um valor entre 0 e 1. Quanto mais próximo a 1 o resultado de uma comparação, mais semelhantes são os conjuntos. Sendo assim, considerando um conjunto A com 6 elementos, e um conjunto B com 8 elementos, e considerando que eles apresentam 4 elementos em comum, a similaridade de Jaccard(A,B) = 4/10.

TÉCNICAS DE PRÉ-PROCESSAMENTO EM MINERAÇÃO DE TEXTOS

Quando trabalha-se com base textuais, existe uma grande quantidade de termos e atributos para sua representação, resultando assim, em uma denotação esparsa, em que grande parte dos atributos são nulos. Dessa forma, as técnicas aplicadas no Pré-Processamento são importantes para resolver problemas em que dados textuais estão envolvidos. Portanto, uma boa amostragem dos dados é

aquela que, identifica os melhores atributos que representam o conhecimento e, que consiga reduzir drasticamente a quantidade destes sem perder as características principais da base de dados.

Tokenização

A identificação de tokens, ou tokenização é uma importante etapa do Pré-Processamento para extrair unidades mínimas de textos. Cada unidade é chamada de token e, normalmente, corresponde à uma palavra do texto, podendo estar relacionado também a símbolos e caracteres de pontuação, com exemplo “ ”, “.”, “! (MANNING et al., 2008).

O termo Token será bastante utilizado nesta dissertação, visto que, em alguns momentos, ele poderá possuir o mesmo sentido de “palavra". De fato, na maioria das vezes um token representa uma palavra no corpus. Como exemplo, a frase: “Amanhã iremos para Belo Horizonte!”, esta frase poderá ser dividida em seis tokens. Conforme mostra o exemplo abaixo.

“Amanhã iremos para Belo Horizonte!”

[Amanhã] [iremos] [para] [Belo] [Horizonte] [!]

Na geração de tokens o “espaço” é sempre descartado, como pode ser visto na transformação acima.

Entretanto, é importante ressaltar que em algumas línguas não se utiliza o espaço como delimitador, como exemplo, japonês e o chinês. Neste caso, Indurkhya e Damerau (2010) divide a tokenização em duas abordagens: uma para as línguas em que o espaço é o delimitador e outra para para aquelas que não utilizam o espaço como delimitador

Outro problema gerado pelos tokens é a dimensionalidade, uma vez que a divisão do texto em palavras, leva à criação de um grande número de dimensões para análise. Na subseções (2.6.3). serão aprensentadas algumas técnicas de redução de dimensionalidade. Por fim, o principal objetivo de criar tokens é a

tradução de um texto em dimensões possíveis de se avaliar, analisar, para obtenção de um conjunto de dados estruturados (JACKSON; MOULINIER, 2007).

Remoção de stopwords

Em um corpus criado para minerar textos, devemos utilizar na atividade de Pré-Processamento, uma técnica conhecida na literatura como Remoção de Stopwords. Pois, ao manipular uma base textual, encontra-se muitos tokens que não possuem valor para o contexto, sendo úteis apenas para a compreensão geral do texto.

Uma lista de stopwords, também conhecida como stoplist é constituída por palavras que adicionam pouco valor à análise. Normalmente, correspondem aos artigos, preposições, pontuação, conjunções e pronomes de uma língua.

Segundo Wives e Loh (1998), uma stopword é considerada como “palavra vazia”, além de não colaborarem para à analise da polaridade de um texto, elas aparecem em praticamente todos os documentos, ou na maioria deles. São exemplos de stopwords em português “a”, “e”, “de”, “da”, dentre outras. A remoção dessas palavras traz um ganho de performance do sistema como um todo e reduz consideravelmente o tamanho final do léxico.

A criação de uma stoplist se dá através de tabelas de contigência que, depois, dão suporte para remoção das stopwords (MANNING et al., 2008). Geralmente, define-se a stoplist por um especialista no domínio da aplicação e, após essa definição, a remoção poderá ser realizada de forma automática, através da frequência de aparição das palavras no léxico.

É importante ressaltar que, para aplicar a técnica de remoção das stopwords, deve-se analisar o que deseja manter do texto original, pois palavras importantes para a aplicação podem ser consideradas stopwords, ou dependendo do contexto, palavras que geralmente não compõem uma lista de stopwords podem ser adicionadas a ela. Existem várias listas de stopwors disponiveis na internet, o que elimina a necessidade de construir uma lista manualmente, entretanto, para este trabalho construiremos uma que atenda o domínio da aplicação.

Redução do léxico

Conforme mencionado na seção(2.6.1), um dos problemas relacionado ao processamento de linguagem natural é o grande número de tokens que não possuem valor para análise. Pois, se considerarmos que cada token em um texto será mapeado para uma classe, gerando assim uma estrutura de dados de grande porte, que, por conseguinte, demandaria um elevado poder de processamento da máquina. Neste sentido, a redução de dimensionalidade torna-se muito importante em processos de classificação automática, não somente para determinar os melhores atributos para modelagem, mas também para aspectos de escalabilidade dos modelos resultantes (KIM; STREET; MENCZER, 2000).

Yu e Liu (2004) descrevem que a quantidade excessiva de atributos causa lentidão no processo de treinamento, bem como na qualidade do conhecimento extraído. Dessa forma, a redução de atributos assume uma papel importante para o sucesso do processo, na medida em que os textos apresentam grande dimensionalidade e variabilidade de termos.

Normalização de palavras

Normalização é a etapa da Redução do Léxico que identifica e agrupa palavras que possuem relação entre elas. Em geral, a aplicação das técnicas de Normalização introduz uma melhora significativa nos sistemas de Mineração de Texto. Está melhora varia de acordo com o escopo, o tamanho da massa textual e o que se pretende obter como saída do sistema (JUNIOR, 2007).

Segundo Manning et al. (2008) existem diversas técnicas para normalizar os dados e, essas técnicas vão de acordo com a necessidade da aplicação. Dentre as várias técnicas para realizar a normalização do dados, destacam-se os processos de Stemming e Seleção de Características.

Stemming

Após a retirada das stopwords, pode-se realizar a técnica de stemming para reduzir cada palavra do léxico, originando assim os “termos”. A raiz de uma palavra é encontrada, na maioria das vezes, eliminando os prefixos,

sufixos que indicam variação na forma da palavra, como plural e tempos verbais.

Em geral, por se tratar de um processo heurístico que simplesmente corta as extremidades das palavras na tentativa de alcançar o objetivo pretendido, os algoritmos utilizados nesta técnica, não se preocupam com o contexto no qual a palavra se encontra. Bem elaborado, o processo de Stemming traz benefícios no pré-processamento, sendo possível reduzir drasticamente o tamanho do léxico e também o esforço computacional, aumentando assim, a precisão dos resultados, exceto quando a retirada de prefixos e sufixos mudam a essência original da palavra.

Extração de características

Quando utiliza-se um corpus de textos extraído de micro blogs, é natural que se tenha dados de alta dimensionalidade. Para aumentar a precisão do algoritmo na etapa de mineração, fazse necessário a implementação de filtros, que selecionam os termos mais representativos no conjunto de características e, com isso, ter um Ganho de Informação (MITCHELL, 1997).

O Ganho de Informação (Information Gain - IG) é um algoritmo que cria um qualificador dos atributos dos dados através da medicão de um peso para cada uma deles. Este peso é baseado pela distribuição dos diferentes valores do atributo em cada uma das classes. Sendo assim, atributos que apresentam cada um de seus valores presentes em somente uma das classes são melhores classificados (LIU, 2012).

Classificação automática de documentos

A classificação automática de textos reporta-se ao procedimento no qual um algoritmo classificador determina a qual classe um documento é pertencente. O principal objetivo da classificação é atribuir uma classe a um conjunto de documentos Prabowo e Thelwall (2009) Especificamente no caso deste trabalho, trata-se de distribuir automaticamente um conjunto de documentos às classes positivas e negativas.

A classificação pode ser dividida em um nível conhecido como aspectos, que trabalha em uma análise de maior granularidade dos documentos, quando a tarefa consiste em classificar cada característica do documento, e a classificação em nível de sentença (MARTINS, 2003).

Existem diversas estratégias para classificar um documento textual e, para este projeto, utilizaremos um classificador baseado em um modelo estatístico que trabalha com métodos indutivos, através de uma abordagem de aprendizado supervisionado, no qual um novo documento é classificado de acordo com as características aprendidas por este classificador, construído e treinado a partir de dados rotulados (MARTINS, 2003). O algoritmo em questão será o Naïve Bayes, que através dos dados de treinamento irá estimar a probabilidade de um documento pertencer a uma determinada classe.

Naive Bayes

O classificador Naive Bayes é um gerador probalístico para textos, sendo um dos mais utilizados em Machine Learning, devido a sua abordagem simplista com que trata as características dos modelos. O termo (Naive) que, em português significa ingênuo, faz referência ao fato deste modelo assumir que existe uma independência condicional dos atributos, ou seja, a distribuição conjunta dos atributos é igual ao produtório da distribuição de cada um deles (CHAKRABARTI, 2002).

Esse classificador é baseado no Teorema de Bayes, criado por Thomas Bayes no século XVIII, sendo este, considerado o mais eficiente na precisão e rotulação de novas amostras (CHAKRABARTI, 2002). Os Classificadores Naive Bayesianos partem da hipótese que todos os atributos são independentes, dado a variável classe, e sua representação gráfica.

Performance de classificadores

Avaliar a performance do classificador é muito importante na classificação de textos, pois, com as métricas é possível averiguar o quão este classificador é capaz de caracterizar um novo exemplo, quando lhe é apresentado. Essa avaliação deverá

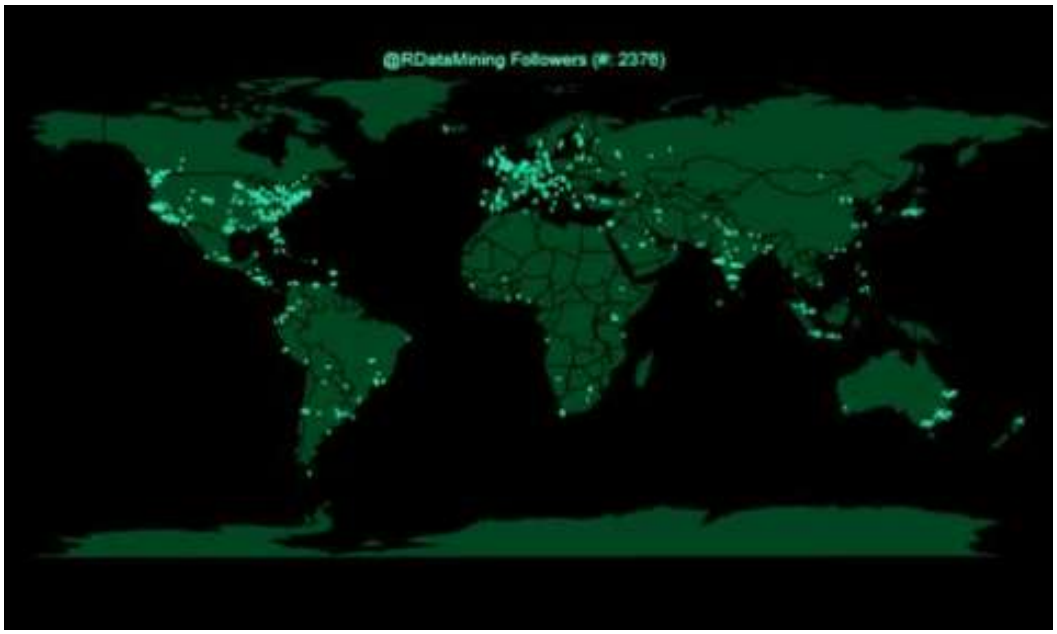
ser realizada logo após a submissão do corpus ao treinamento, utilizando-se do resultado da classificação do conjunto de teste.

TEXT MINING NO R

Além dos diversos pacotes e técnicas disponíveis no Software R, a análise de Text Mining possui ainda funções como a capacidade de mostrar em mapa mundial a quantidade de seguidores que usaram determinada hashtag (#, usada para falar de determinado assunto com destaque). Também é possível demonstrar em gráfico os usuários mais influentes para determinado estudo de tweets, ou seja, as pessoas que mais escreveram sobre determinado assunto, e também se foram copiadas (retweets).

Gráficos de linhas com o máximo número de tweets retweetados (tweets repetidos de alguém que escreveu antes) por tempo, seja por dia, mês ou ano também são feitos na análise de Text Mining atuais. Esses exemplos são para análises feitas com dados do Twitter. A Figura 4 mostra um exemplo de gráfico que mostra, com pontos verdes em tom mais claro, os locais de onde as pessoas tweetaram sobre @RDataMining (ZHAO, 2016).

Figura 4: Mapa de pessoas que tweetaram sobre @RDataMining (ZHAO, 2016).



Software R

O Software R pode ser baixado diretamente na internet pelo site 'https://www.r-project.org/'. Além disso, o Software R tem amplo conjunto de funções e pode ser aperfeiçoado com o uso de novos pacotes, ou seja, é um programa bastante poderoso quanto a análises estatísticas.

Também é possível acessar o Software R remotamente por um Server que pode ser acessado pela internet, tendo assim cada conjunto de usuários uma máquina virtual protegida por senha. Cada usuário teM seu login e senha que dará acesso a máquina e também limitará suas operações que são controladas por um administrador. Isso facilita muito a interação entre as pessoas e troca de informações, de forma que todos podem ver e ajudar seus colegas nas análises. O programa One Book One Chicago conta com uma máquina virtual poderosa onde são armazenadas suas análises, isso tudo para ter um trabalho seguro e ao mesmo tempo compartilhado com os participantes.

O pacote 'tm' é o pacote do R que possui as principais funções utilizadas no Text Mining, sendo o mais utilizado nas análises. Alguns pacotes que serão utilizados nesse trabalho: 'RColorBrewer', 'fpc', 'wordcloud', 'topicmodels' e 'ggplot2'

que servem para colocar cores em mapas temáticos, auxílio na análise de cluster, nuvem de palavras, modelagem por tópicos e construção de gráficos, respectivamente.

Novidades de Text Mining no R

Alguns pacotes para Text Mining foram agregados ao R nos últimos tempos. Os métodos e novos pacotes que serão utilizados no trabalho estão a seguir:

- **Pacote 'stringr'** – pacote que ajuda com o manuseio de strings. Utilização nesse trabalho para excluir tweets que continham palavras indesejáveis.
- **Pacote 'RColorBrewer'** – este pacote serve para colocar cores para mapas temáticos.
- **Pacote 'fpc'** – pacote para auxílio de métodos, validação e estimação de clusters. Será usado nesse trabalho para fazer análise de Cluster around medoids.
- **Pacote 'topicmodels'** – pacote utilizado para modelagem por tópicos, que é o processo de dividir os textos por tópicos. É possível também fazer gráficos com cores dividindo os diferentes tópicos.
- **Pacote 'cluster'** – pacote que contém funções para análises de cluster. Utilização nesse trabalho para fazer gráfico 'cusplot'.

Além desses pacotes, o processo Wordclouds por cluster, que proporciona melhor visualização e entendimento das informações textuais, é também novidade para a análise de Text Mining no R. Serão feitas nuvens de palavras para cada cluster encontrado na análise de k-means e para cada tópico encontrado na modelagem por tópicos.

PROCESSO DE MINERAÇÃO DE DADOS

A mineração de textos pode conter várias etapas, mas quatro delas são básicas em todos os processos: coleta de documentos, pré-processamento, extração de conhecimento e avaliação e interpretação dos resultados (MARTINS et al., 2003). A coleta de documentos objetiva conseguir documentos relacionados ao tipo de conhecimento que se deseja obter. Pode-se utilizar de várias fontes, como livros, emails, fóruns de internet, etc. Existem técnicas como o Processamento de Linguagem Natural e Recuperação de Informação que podem ser utilizadas nesta etapa (MARTINS et al., 2003).

Uma característica importante deste passo é a limitação quanto ao uso de informações externas. O conhecimento de mundo e de especialistas não são utilizados pois os algoritmos de agrupamento dos documentos aprende de forma não supervisionada como extrair o conhecimento dos textos, e desta maneira, categorizar os documentos de forma a facilitar os próximos passos do processo de mineração de textos (CORRÊA et al., 2012).

No pré-processamento os documentos adquiridos na primeira etapa, escritos em linguagem natural, passam por uma formatação para estrutura-los de maneira padronizada, mas sem perder suas características naturais, para que os algoritmos que serão utilizados nas próximas etapas sejam capazes de manipular todos os documentos da mesma maneira. Ao final deste processo obtém-se uma estrutura que representa o grupo de documentos fonte, geralmente uma tabela atributo-valor (MARTINS et al., 2003).

Neste segundo passo também são definidos os termos que serão utilizados para a extração para conseguir um grupo pequeno e representativo dos termos presentes nos grupos de textos. Para isto, eliminam-se as stopwords, termos sem significado relevante para a pesquisa, como artigos, advérbios e pronomes. Além disso, são identificadas variações morfológicas e sinônimos dos termos, utilizando técnicas como stemming e thesaurus, a fim de diminuir ainda mais o conjunto de pesquisa. Esta redução permite a diminuição do custo computacional das próximas etapas do processo (CORRÊA et al., 2012).

Na terceira etapa, a extração do conhecimento, aplica-se algoritmos de extração automática de conhecimento para buscar informações desconhecidas até o momento, mas que possam ser úteis para o domínio da questão (MARTINS et al., 2003). Estes algoritmos buscam agrupar objetos similares através de uma medida de proximidade, ao mesmo tempo em que tenta formar grupos com características dissimilares entre si. A análise por agrupamento também é chamada de análise exploratória de dados ou aprendizado por observação (CORRÊA et al., 2012).

Por último, na avaliação e interpretação dos resultados, utiliza-se a ajuda de um usuário para descobrir se os resultados obtidos são satisfatórios, e em caso negativo, que etapas poderiam ser refeitas para melhorá-los (MARTINS et al., 2003).

APLICAÇÕES DA MINERAÇÃO DE TEXTOS

A mineração de textos possui aplicações nas mais variadas áreas científicas e comerciais. Ela pode ser usada, por exemplo, na medicina, pois a quantidade de informação de texto gerada nesta área é enorme (prontuários, registros hospitalares, fichas de pacientes, etc). Estes documentos podem ser avaliados com técnicas de mineração de dados para auxiliar os profissionais da medicina a diagnosticar doenças ou buscar tratamentos. Uma ferramenta que utiliza este conceito é o Medline, um software que trabalha coma base de dados bibliográficos da Biblioteca Nacional de Medicina dos Estados Unidos da América (CARRILHO JUNIOR, 2007).

Além disso, a mineração de textos pode ser utilizada para a análise de sentimentos em pesquisas de opinião pública. Muitas vezes estas pesquisas são feitas com questionários com perguntas fechadas, ou seja, os entrevistados podem escolher somente opções pré-determinadas. O problema é que muitas vezes isto não reflete a realidade, pois as perguntas podem exigir uma respostas mais elaborada. Se em fez deste tipo de questionário for utilizada uma entrevista com

respostas abertas, de maneira que o entrevistado possa escrever sua resposta em linguagem natural, é possível analisar os resultados com uma ferramenta de mineração de textos (CARRILHO JUNIOR, 2007).

Esta área da mineração de textos é conhecida como “Análise de Sentimentos”, e visa identificar como o autor de um texto expressa seus sentimentos de forma escrita e categorizar a satisfação (favorável ou desfavorável) em relação ao assunto abordado(CARRILHO JUNIOR, 2007). Por último, a mineração de textos pode ser utilizada para ajudar empresas grandes que trabalham com atendimento ao cliente. Muitas vezes, um produto ou serviço apresenta algum defeito e o cliente precisa entrar em contato com algum especialista da empresa para resolver o seu problema. É comum nestes casos a requisição do cliente ser transferida de setor para setor e demorar muito tempo até chegar ao seu destino final (CARRILHO JUNIOR, 2007).

A proposta da mineração de textos para solucionar este problema é analisar textualmente a requisição do cliente e enviá-la de maneira automática diretamente para o especialista no assunto, removendo a intervenção humana do processo e aumenta o tempo de entrega da requisição (CARRILHO JUNIOR, 2007)

A mineração de textos possui potencial para ser muito bem explorada comercialmente, não apenas pela grande variedade de informações comerciais que são armazenadas em forma de texto, mas também por causa da sua capacidade de ser aplicada em várias áreas diferentes do conhecimento.

Além disso, a tecnologia da mineração de dados pode fornecer automação para vários serviços que atualmente são feitos por seres humanos e, consequentemente, a diminuição de custo e maior agilidade nestes processos. Exemplos disto são a análise automática de sentimentos em pesquisas de opinião pública, que quando realizadas por pessoas demoram muito mais tempo do que se realizadas por um algoritmo de mineração de dados.

Por fim, é importante salientar que mesmo com toda a automação fornecida pela mineração de dados, ainda é necessário alguém ao final do processo para avaliar os resultados obtidos na mineração.

REFERENCIAS

ARANHA, Christian; PASSOS, Emmanuel. **A Tecnologia de Mineração de Textos**. 2006. Disponível em: . Acesso em: 01jun. 2015.

AGGARWAL, C. C.; ZHAI, C. **Mining text data**. [S.l.]: Springer Science & Business Media, 2012.

ARANHA, C. N.; VELLASCO, M.; PASSOS, E. **Uma abordagem de pré-processamento automático para mineração de textos em português:** sob o enfoque da inteligência computacional. Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, 2007.

BARROSO, L. P.; ARTES, R. **Análise multivariada**. Lavras: Ufla, 2003.

CORRÊA, Geraldo Nunes et al. **Uso da mineração de textos na análise exploratória de artigos científicos**. 2012. Disponível em: . Acesso em: 01 jun. 2015. CARRILHO

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. et al. **A density-based algorithm for discovering clusters in large spatial databases with noise**. In: Kdd. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FILHO, J. A. C. **Mineração de textos:** análise de sentimento utilizando tweets referentes à copa do mundo 2014. TCC (Graduação em Sistemas de Informação) - Universidade Federal do Ceará, 2014.

JOHNSON, R. E.; RUSSO, V. **Reusing object-oriented designs**. [S.l.]: Department of Computer Science, University of Illinois at Urbana-Champaign, 1991.

JUNIOR, João Ribeiro. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 2007. Disponível em: . Acesso em 02 jun. 2015.

LEITE, J. L. A. **Mineração de textos do twitter utilizando técnicas de classificação.** TCC (Graduação em Sistemas de Informação) - Universidade Federal do Ceará, 2016.

LEONEL JUNIOR, R. d. A.; JUNIOR, J. H. F.; JORGE da S. T.; SILVA, T. L. da; MAGALHãES, R. P. **Mineração em Dados Abertos In: IV Jornada Científica de Sistemas de Informação.** [S.l.]: Parnaíba, PI, 2014. MATHIAK, B.; ECKSTEIN, S. Five steps to text mining in biomedical literature. In: Proceedings of the second European workshop on data mining and text mining in bioinformatics. [S.l.: s.n.], 2004. p. 43–46.

MATTSSON, M.; BOSCH, J. **Stability assessment of evolving industrial object-oriented frameworks.** Journal of Software Maintenance: Research and Practice, Wiley Online Library, v. 12, n. 2, p. 79–102, 2000.

MACHADO, Aydano P. et al. **Mineração de Texto em Redes Sociais Aplicada à Educação a Distância.** 2010. Disponível em: . Acesso em: 01 jun. 2015.

MARTINS, Claudia Aparecida et al. **Uma Experiência em Mineração de Textos Utilizando Clustering Probabilístico Clustering Hierárquic**o. 2003. Disponível em: . Acesso em: 01 jun. 2015.

RECUERO, R.; ZAGO, G. **Em busca das "redes que importam":** redes sociais e capital social no twitter. LÍBERO. ISSN impresso: 1517-3283/ISSN online: 2525-3166, n. 24, p. 81–94, 2016.

RODRIGUES, P. R. F. **Dinamica de temas abordados no twitter via evoluca de clusters**. TCC (Graduação em Sistemas de Informação) - Universidade Federal do Ceará, 2016.

RUSSELL, M. A. **Mining the Social Web:** Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. [S.l.]: "O'Reilly Media, Inc.", 2013.

SILVA, T. L. da; SOUSA, F. R.; MACÊDO, J. A. F. de; MACHADO, J. C.; CAVALCANTE, A. A. **Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem**. [S.l.]: Quixadá, 2013. 48 SIMOS, G. C. How Much Data Is Generated Every Minute On Social Media? 2015. Disponível em:

SERAPIÃO, Paulo Roberto Barbosa et al. **Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia.** 2010. Disponível em: . Acesso em: 01 jun. 2015.

TAN, A.-H. et al. **Text mining:** The state of the art and the challenges. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. [S.l.: s.n.], 1999. v. 8, p. 65–70.

TAN, P.-N. et al. Introduction to data mining. [S.l.]: Pearson Education India, 2006. VIANA, Z. L**. Mineração de textos:** análise de sentimento utilizando tweets referentes às eleições presidenciais 2014. TCC (Graduação em Sistemas de Informação) - Universidade Federal do Ceará, 2014.

WU, F.-x. **Genetic weighted k-means algorithm for clustering large-scale gene expression data.** BMC bioinformatics, BioMed Central, v. 9, n. 6, p. 1, 2008. WU, X.; ZHU, X.;

WU, G.-Q.; DING, W. **Data mining with big data. IEEE transactions on knowledge and data engineering**, IEEE, v. 26, n. 1, p. 97–107, 2014.

YIN, J.; LAMPERT, A.; CAMERON, M.; ROBINSON, B.; POWER, R. **Using social media to enhance emergency situation awareness**. IEEE Intelligent Systems, v. 27, n. 6, p. 52–59, 2012.