Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciência
Instituto de Matemática e Estatística

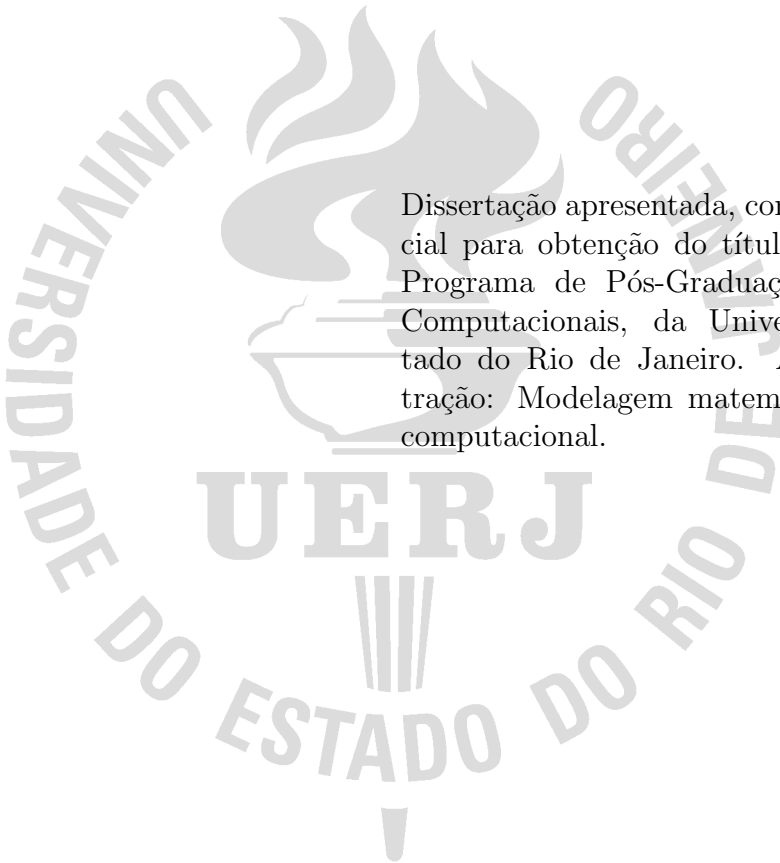
Fabio Mascarenhas Loureiro

Uma Análise das Principais Técnicas de Agrupamento de Dados, aplicadas a Coletâneas Textuais recuperadas de Páginas Web

Rio de Janeiro
2016

Fabio Mascarenhas Loureiro

Uma Análise das Principais Técnicas de Agrupamento de Dados, aplicadas a Coletâneas Textuais recuperadas de Páginas Web



Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro. Área de concentração: Modelagem matemático-estatístico-computacional.

Orientador: Prof.ª Dra. Célia Martins Cortez Silva
Coorientador: Prof. Dr. Alan Freitas Machado

Rio de Janeiro
2016

CATALOGAÇÃO NA FONTE

UERJ / REDE SIRIUS / BIBLIOTECA CTC-A

L892 Loureiro, Fábio Mascarenhas.
Uma análise das principais técnicas de agrupamento de dados, aplicadas a coletâneas textuais recuperadas de páginas Web/ Fábio Mascarenhas Loureiro. – 2016.
81 f. : il.

Orientadora: Célia Martins Cortez Silva
Coorientador: Alan Freitas Machado
Dissertação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística.

1. Mineração de dados (Computação) - Teses. 2. Banco de dados – Teses. 3. HTML (Linguagem de marcação de documento) - Teses. I. Silva, Célia Martins Cortez. II. Machado, Alan Freitas. III. Universidade do Estado do Rio de Janeiro. Instituto de Matemática e Estatística. IV. Título.

CDU 004

Patricia Bello Meijinhos CRB7/5217 -Bibliotecária responsável pela elaboração da ficha catalográfica

_____ _____

Assinatura Data

Fabio Mascarenhas Loureiro

Uma Análise das Principais Técnicas de Agrupamento de Dados, aplicadas a Coletâneas Textuais recuperadas de Páginas Web

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro. Área de concentração: Modelagem matemático-estatístico-computacional.

Aprovada em 03 de Fevereiro de 2016.
Banca Examinadora:

Prof.ª Dra. Célia Martins Cortez Silva (Orientador)
Instituto de Matemática e Estatística – UERJ

Prof. Dr. Alan Freitas Machado (Coorientador)
Instituto de Física – UERJ

Prof. Dr. Leandro Augusto Justen Marzulo
Instituto de Matemática e Estatística - UERJ

Prof.ª Dra. Jéssica Kubrusly
Universidade Federal Fluminense

Rio de Janeiro
2016

DEDICATÓRIA

Ao meu avô/pai, que tornou possível trilhar os caminhos que me trouxeram até este momento. (in memoriam)

AGRADECIMENTOS

Ao meu avô, que é minha grande referência e me auxiliou a construir as bases para que hoje pudesse finalizar mais esta etapa.

Aos meu orientadores, pela compreensão e tempo dedicado à elaboração deste trabalho.

A Instituição, por proporcionar meu desenvolvimento acadêmico.

Um anão sobre os ombros de um gigante pode ver
mais longe que o próprio gigante.
Robert Burton

RESUMO

LOUREIRO, F. M. L. *Uma Análise das Principais Técnicas de Agrupamento de Dados, aplicadas a Coletâneas Textuais recuperadas de Páginas Web*. 2016. 81 f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística , Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

Nos últimos tempos, o volume de dados vem crescendo exponencialmente e, portanto, devemos buscar soluções apropriadas para extrair conhecimento dessa enorme massa. Uma metologia adequada para lidar com grandes conjuntos de dados é a Descoberta de Conhecimento em Bases de Dados (KDD). Esta metodologia é constituída por várias etapas, possuindo como motor principal a Mineração de Dados. Contudo, não se deve empregar as tarefas relacionadas à Mineração de Dados diretamente no banco, visto que, frequentemente, as tarefas que compõem esta etapa não lidam nativamente com objetos textuais. A estes conjuntos de técnicas utilizadas para interpretar e analisar dados textuais foi denominado Descoberta de Conhecimento em Textos (ou Mineração de Textos). Este campo, ainda emergente, destina-se ao recolhimento de informações significativas a partir de textos em linguagem natural, extraindo conhecimento a partir de documentos textuais. Neste quadro, este trabalho busca apresentar uma visão panorâmica das fases do KDD e seu relacionamento com a etapa de Mineração de dados, exemplificando alguns de seus principais métodos através de artigos e pesquisas científicas que os empregaram. Posteriormente, são detalhadas as etapas da Mineração de Textos, apresentando, principalmente, suas tarefas de tratamento e redução dos termos da coletânea. Para a primeira etapa, destinada a coleta de documentos, é sugestão deste trabalho o desenvolvimento de rastreadores web focados na recuperação de conteúdos a partir de páginas HTML, armazenando-os em um formato facilmente processável pelas etapas seguintes. Após a apresentação teórica, foi aplicada a metodologia estudada com a finalidade de criar um rastreador web específico para a seção de economia do jornal O Globo (edição digital), seguido pela execução de técnicas de agrupamento a fim de intuir sobre os principais assuntos abordados em um determinado período e verificar a possibilidade de agrupamento das postagens por tema e autor.

Palavras-chave: Mineração de Dados. Mineração de Textos. Análise de Agrupamento. Rastreadores Web.

ABSTRACT

LOUREIRO, F. M. L. *Uma Análise das Principais Técnicas de Agrupamento de Dados, aplicadas a Coletâneas Textuais recuperadas de Páginas Web*. 2016. 81 f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística , Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

The volume of information is growing exponentially and, therefore, we should search appropriate solutions to extract knowledge from this huge mass. A suitable methodology for the handle large data sets is the Knowledge Discovery in Databases (KDD). This methodology consists of several steps and has Data Mining as main motor. However, we shouldn't employ the data mining tasks directly in the base, because, quite often, the tasks that make up this step doesn't deal with textual objects natively. To these sets of techniques used to interpret and analyze textual data it was called Knowledge Discovery in Texts (or Text Mining). This field, still emerging, intended for gathering meaningful information from texts in natural language, extracting knowledge from text documents. In this context, this research presents an overview of estages of KDD and its relationship with the Data Mining step, ilustrating some of its key methods through articles and scientific researches that employed them. In sequence, are detailed the steps of Text Mining, presenting their treatment and term's reduction tasks. For the first stage, the colection of documents, this work suggests the development of focused web crawlers in recovering contents from HTML pages, storing them in an easily processable format. After the theorical presentation, was applied the methodology studied to create a specific web crawler to the O Globo (digital edition) newspaper business section followed by the execution of clustering techniques to intuit about the main issues addressed in a certain period and check the possibility of grouping posts by topics and author.

Keywords: Data Mining. Text Mining. Clustering. Web Crawler.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SECOM	Secretaria de Comunicação Social da Presidência da República
KDD	Knowledge Discovery in Databases
DM	Data Mining

SUMÁRIO

INTRODUÇÃO

Segundo pesquisa realizada pela Secretaria de Comunicação do Brasil (SECOM) (SECOM, 2014), no ano de 2015 cerca de 37% da população brasileira (estimada em 205 milhões) fazia uso diário da internet. Em 2014 a EMC[1] divulgou que existiam disponíveis quase 1 septilhão de bits de informação, grandeza equivalente às estrelas conhecidas no céu, segundo a Agência Espacial Européia. Esta mesma empresa estima que, até 2020, o número de dados armazenados em computadores, servidores, celulares e tablets, seja no mínimo multiplicada por seis; uma massa tão gigantesca que os pesquisadores passaram a mensurá-la em termos da distância da Terra à Lua (GLOBO, 2014).

Diante desta realidade, é inegável que retirar informações desta enorme massa de dados manualmente é impraticável. Neste contexto, surje a necessidade de desenvolver (ou refinar) metodologias adequadas para extrair deste "bojo" algo útil, relevante, para uma tomada de decisão. Uma alternativa para lidar com este tema é a aplicação de técnicas da Descoberta de Conhecimento em Bases de Dados (em inglês, *Knowledge Discovery in Databases*), também conhecido como KDD.

O KDD pode ser entendido como um processo, constituído de várias etapas, com a finalidade de descobrir padrões e tendências através da análise de grandes conjuntos de dados, tendo, como principal etapa, o processo de mineração de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Entretanto, técnicas desta natureza, em sua maioria, são aplicadas apenas a dados numéricos, o que é bastante problemático, visto que a maior parte destes dados se encontram de forma não estruturada e, em sua maioria, em formato textual. Este fato é corroborado quando se observa os próprios hábitos dos indivíduos, onde facilmente nota-se que a maior parte das informações que são trocadas diariamente estão em algum formato linguístico, e não apenas de forma numérica. Obviamente, tratar este tipo de dado exige um olhar diferenciado e, portanto, quando deseja-se utilizar um processo de mineração de dados para coletâneas textuais, é necessário o emprego de alguns ajustes. Estes refinamentos, no entanto, são tão específicos que podem ser classificados como uma área de estudos a parte, denominada Descoberta de Conhecimento em Textos ou *Text Mining*(Mineração de Textos) (CHAKRABARTI, 2002).

A Mineração de Textos é um campo ainda emergente que objetiva recolher informações significativas a partir de textos em linguagem natural, podendo ser definido como o processo de extrair padrões ou conhecimento, interessantes ou não-triviais, a partir de documentos textuais. Este processo engloba, além de técnicas da KDD, métodos

[1] Empresa multinacional norte-americana líder no fornecimento de sistemas de infraestrutura de informação, armazenamento de dados, software e serviços.

de outras áreas da computação, como Recuperação da Informação e Processamento de Linguagem Natural (WITTEN, 2005).

A primeira etapa para a utilização dos métodos de Mineração de Textos é a seleção e coleta do acervo que será estudado. Em particular, quando deseja-se obter informações disponíveis na Web, é indispensável a utilização de uma ferramenta que auxilie na coleta e tratamento destes documentos. Esta problemática é tratada através de algoritmos e sistemas de busca e recuperação de informação que trabalham na Web, denominados rastreadores web (REIS, 2013). Um rastreador web - também chamado de: *web crawler*, *web spider*, *web robot*, *web wanderer* ou *worm* - pode ser entendido como um programa construído para buscar e recuperar informações da Web de forma sistemática e pré-definida (PAES, 2012). Vale ressaltar que algoritmos deste tipo são utilizados tanto para busca na *World Wide Web* (WWW) quanto em intranets, o que é extramamente benéfico quando se quer recuperar arquivos em servidores particulares ou analisar dados provenientes de bancos de dados próprios.

Neste ambiente, este trabalho buscou explicitar uma visão geral sobre o processo de KDD e seu relacionamento com a Mineração de Dados; apresentando as principais tarefas ligadas a esta metodologia exemplificando-as através de aplicações em artigos e pesquisas científicas. Além disto, foram apresentadas formas de desenvolver rastreadores web para buscar e armazenar coletâneas textuais a partir de páginas WWW, escritas em HTML.

Após o desenvolvimento teórico, foi aplicado o conhecimento obtido para analisar uma coletânea de dados reais. Para isso, foi construído um rastreador web em Python para recuperar informações de *posts* da seção de economia da jornalista Míriam Leitão, alocado no portal do jornal O Globo (edição digital). Os dados recuperados foram estruturados em um formato manipulável a fim de facilitar análises futuras. O desenvolvimento da mineração de dados deu-se no software R e foram aplicados métodos inerentes à tarefa de agrupamento, da mineração de dados, para analisar o conteúdo exposto pelos colunistas, buscando extrair informações interessantes a partir das matérias publicadas.

Objetivos

O objetivo primário deste trabalho é elucidar todo o processo de Descoberta de Conhecimento Textual, apresentando como proposta o desenvolvimento de rastreadores web para a coleta e armazenamento dos objetos. Uma vez montada a coletânea textual, tenciona-se apresentar algumas tarefas da Mineração de Dados que podem ser empregadas em dados desta natureza (Mineração de Texto), em especial, àquelas de agrupamento. Ao final do trabalho, realizar-se-á uma aplicação do conhecimento adquirido para analisar *posts* da seção de economia da jornalista Míriam Leitão, alocada no portal do jornal O

Globo (edição digital). Com isto, deseja-se que a lógica empregada sirva de base para futuras aplicações mais robustas.

Para alcançar este objetivo, foram desdobradas metas secundárias: o trabalho deverá apresentar as atividades de um processo de Descoberta de Conhecimento Textual; terá que apresentar a arquitetura de funcionamento e desenvolvimento de rastreadores web, realizando um *overview* de páginas HTML (base essencial para o desenvolvimento de rastreadores web); e deverá elucidar o resultado obtido em cada possível tarefa dentro da Mineração de Dados. O último objetivo secundário é extremamente relevante, ao passo que cada tarefa possui uma intenção específica, que implica em técnicas, resultados e formas de visualização diferenciados. As tarefas de agrupamento serão apresentadas mais detalhadamente, uma vez que configuram a tarefa de mineração principal deste trabalho.

1 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS E MINERAÇÃO DE DADOS

1.1 Motivação

Com o desenvolvimento de novas tecnologias, presencia-se um crescimento surpreendentemente acelerado na capacidade de geração e armazenamento de dados. Vide os satélites de observação da NASA que geram cerca de um *terabyte* de dados por dia; o projeto Genoma com outros milhares de *bytes* para cada uma das bilhões de bases genéticas; e os milhões de *bytes* resultantes do censo dos Estados Unidos desde 1990 (BRAMER, 2007). Entretanto, ao passo que a quantidade de dados aumenta, "maior é a lacuna entre a geração destes dados e o entendimento deles" (LOPES, 2003).

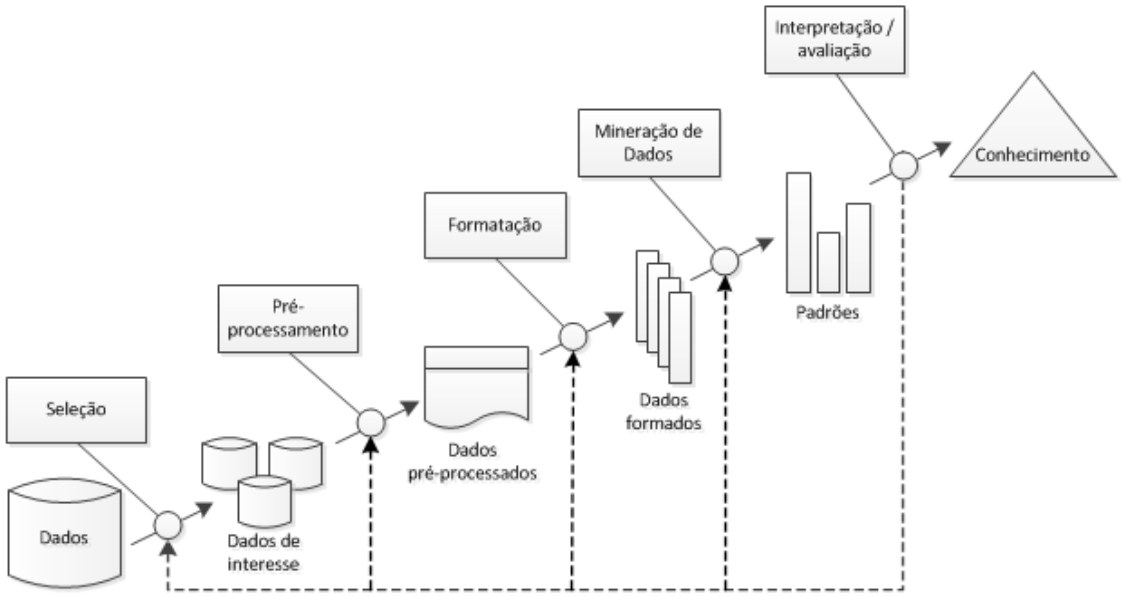
De fato, os métodos tradicionais de análise, na prática, ainda se baseiam muito na percepção e interpretação manual dos dados. Entretanto, essa forma de análise, quando empregada a grandes volumes dados, se torna lenta, cara e altamente subjetiva, visto que não se baseia em uma metodologia única para apreciação daqueles dados, ficando a mercê do indivíduo a informação resultante da análise (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Neste contexto, surge um campo de pesquisas cujo propósito é o estudo e desenvolvimento de metodologias para a extração de informaões de auto nivel (conhecimento) a partir de dados de baixo nível (usualmente estocados em grandes bases dados) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A esse campo, foi dado o nome de "Descoberta de Conhecimento em Bancos de Dados" (em inglês, *Knowledge Discovery in Databases*), também conhecido como KDD. Esta, acabou se tornando uma área de interesse para diversos pesquisadores e profissionais de diversas áreas, incluindo Fayyad (1996): Inteligência Artifical, estatística, reconhecimento de padrões e computação paralela.

1.2 Descoberta de Conhecimento em Base de Dados

A Descoberta de Conhecimento em Bases de Dados é definida por Frawley, Piatetsky-Shapiro e Matheus (1992) como a extração não trivial de informações implícitas, previamente desconhecida e potencialmente útil a partir de um conjunto de dados (F). As metodologias desenvolvidas nesta esta área buscam a extração de informações através do reconhecimento de padrões, associções ou correlações existentes entre subconjuntos de dados. Nesta abordagem, dado uma linguagem L e alguma medida de certeza C, define-se um "padrão" como um estado S em L que descreve uma relação entre os subconjuntos F_S de F com uma certeza c, tal que S é mais simples que totalidade dos dados em F_S

Figura 1 - Processo de Descoberta de Conhecimento em Banco de Dados.



Fonte: Adaptação de Fayyad, Piatetsky-Shapiro e Smyth (1996).

(FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992).

O processo de KDD pode ser entendido, como uma série de atividades (etapas) que são iniciadas a partir da definição de uma meta - aquilo que se deseja extrair de informação - e finaliza com recuperação da informação pertinente (MAIMON; ROKACH, 2005). Estas etapas podem ser vistas através da Figura 1. Nesta imagem, pode-se notar que o processo é iterativo a cada etapa,o que significa que pode ser requerido diversos laços de repetição dentro de uma mesma fase e ou entre etapas, até que o resultado final seja alcançado. Para Maimon e Rokach (2005) o processo possui, ainda, um aspecto "artístico", uma vez que um processo não pode ser definido em uma fórmula única ou realizar uma taxonomia completa para as escolhas corretas em cada passo para cada tipo de aplicação. Desta forma, é necessário um conhecimento profundo do processo e as diferenças necessárias e possíveis em cada etapa.

A seguir, segue uma breve descrição do que se espera a cada etapa do processo ilustrado pela Figura 1 (MAIMON; ROKACH, 2005; CIOS et al., 2007; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996):

- **Seleção:** etapa inicial do processo que consiste no aprendizado do domínio que será realizado a aplicação e definição dos conjuntos que serão usados dentro da base de dados. Estão inclusos nesta fase, conhecimentos *a priori* sobre o conjunto; definição das metas que se deseja alcançar ao final do processo; e seleção de subsonjuntos, variáveis e ou amostras que serão utilizados no decorrer do processo.

- **Pré-processamento:** tem enfoque no tratamento e na preparação dos dados para uso pelos algoritmos. Estão inclusos nesta fase operações básicas para limpeza dos dados, como a remoção de ruídos e *outliers* se for apropriado; coleta de informações necessárias para modelar ou buscar os responsáveis pelos ruídos; e definir estratégias para lidar com valores ausentes ou desconhecidos.

- **Transformação:** objetiva a aplicação, quando necessário, alguma transformação linear, ou não, nos dados, de forma a encontrar aqueles mais relevantes para o problema em estudo. Nesta etapa geralmente são aplicadas técnicas de redução de dimensionalidade e de projeção dos dados.

- **Mineração de Dados (DM):** objetiva a busca por padrões através da aplicação de algoritmos e técnicas computacionais específicas. Esta atividade é considerada a mais importante do processo e pode ser entendida como uma etapa de condensa três grandes atividades:

  - Escolha da tarefa de mineração de dados - Aqui, o pesquisador cruza as metas definidas inicialmente com um método particular de DM, como classificação, regressão e agrupamento.
  - Escolha do algoritmo de mineração de dados - Nesta atividade, o pesquisador seleciona métodos de busca por padrões nos dados e decide qual modelo e parâmetros dos métodos podem ser mais apropriados.
  - Mineração de Dados - Escolha do algoritmo de mineração de dados - Esta atividade gera padrões em uma forma de representação particular para cada algoritmo escolhido, tais como regras de classificação, árvores de decisão e modelos de regressão.

- **Interpretação e Avaliação:** etapa de análise dos resultados da mineração e da geração de conhecimento pela interpretação e avaliação do conteúdo obtido na etapa anterior. O analista deve construir formas de visualização da informação obtida coerentes com os métodos empregados na DM.

A maior parte dos trabalhos encontrados são focados na etapa de Mineração de Dados, entretanto, as outras etapas são igualmente importante para o sucesso da aplicação de um KDD. Todavia, esta fase continua sendo a de maior complexidade, por vezes, sendo usada para denominar o próprio processo de KDD. Em sequência, é apresentado um estudo mais aprofundado nas atividades que compõe esta etapa.

1.3 Mineração de Dados

A Mineração de Dados (em inglês, *Data Mining*) constitui uma das principais fases do processo de KDD, sendo evidenciada pela construção de modelos computacionais para a descoberta automática de novos fatos e relacionamentos entre dados, a partir da aplicação de algoritmos de busca. O termo "minerar"faz alusão ao ato de garimpar por alguma preciosidade; por sua vez, "minerar dados"remete ao ato de garimpar bases de dados por alguma informação de valor significativo para o pesquisador. Para Camilo e Silva (2009) a mineração de dados pode ser definida como "a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados".

Existem diversos métodos de DM para diferentes propósitos e metas. Uma forma de visualizar panoramicamente estes diversos métodos pode ser vista na Figura 2. Nesta imagem, é apresentada a Taxonomia dos métodos de DM feita por Maimon e Rokach (2005), onde apresenta os tipos de métodos, suas interrelações e os grupos a que pertencem.

Os primeiros tipos existentes fazem menção à orientação do algoritmo, podendo ser dos tipos: verificação orientada, onde o sistema verifica a hipótese do usuário; e descoberta orientada, onde o sistema encontra novas regras e padrões autonomamente (MAIMON; ROKACH, 2005).

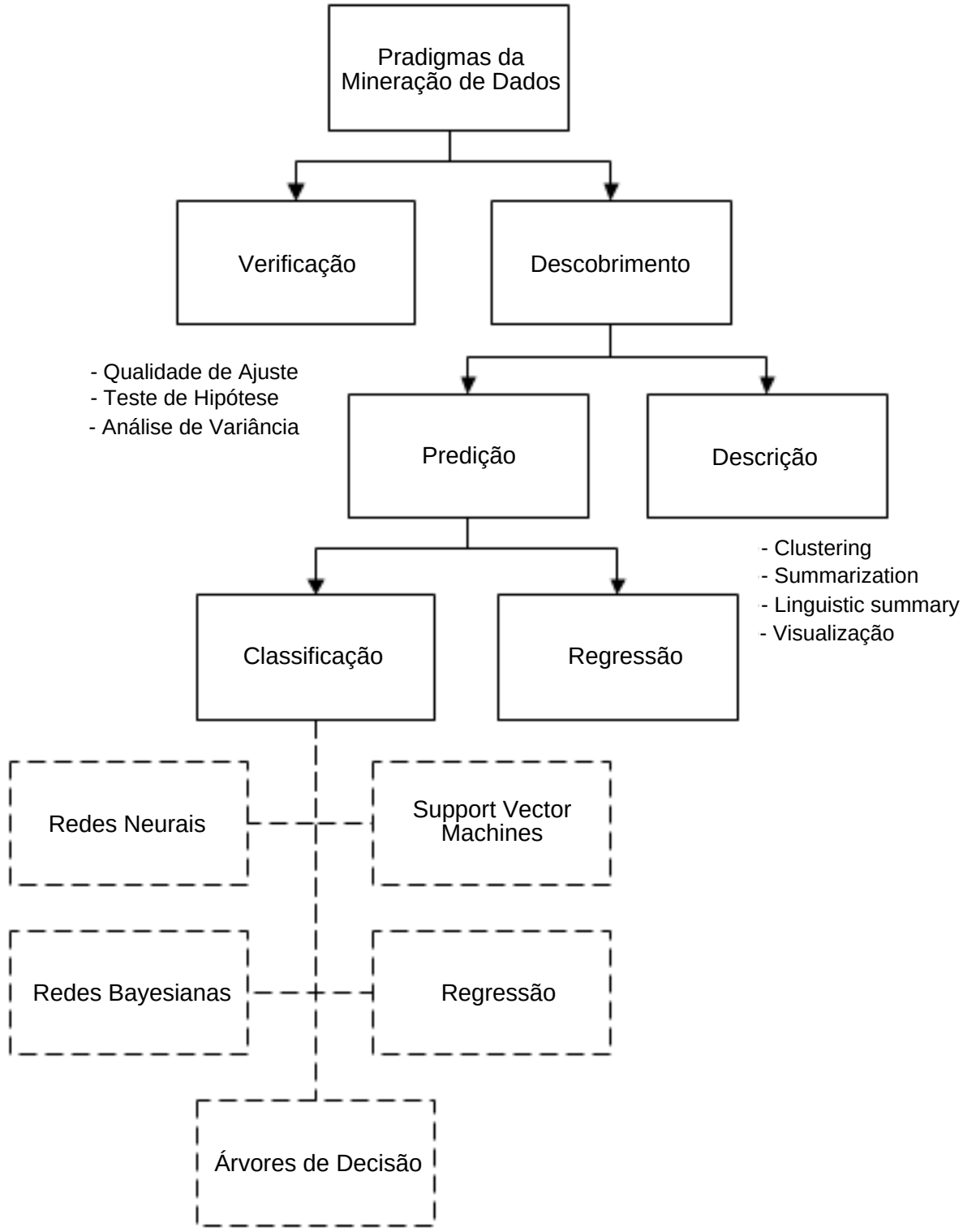
Os ramos descendentes do tipo descobrimento consistem de métodos de predição e descrição. Métodos de descrição são orientados à interpretação, com foco no entendimento de como os dados subjacentes se relacionam com suas partes. Métodos de predição alvejam construir automaticamente um modelo comportamental. Este, é bastante útil para predizer valores de um ou mais variáveis relacionadas a um amostra (MAIMON; ROKACH, 2005). Na Figura 2 são apresentadas grandes áreas que, normalmente, derivaram de outros nichos e foram remodeladas para serem aplicadas na DM.

Conforme Maimon e Rokach (2005) explicita, a maior parte das técnicas orientadas ao tipo descobrimento são baseadas em aprendizagem induzida, onde, o modelo é construído, explicitamente ou implicitamente, através da generalização de um número suficiente de exemplos de treinamento. Estas metodologias visam criar modelos que, a partir dos dados de treino, sejam aplicáveis a dados desconhecidos e retornem predições coerentes. Os outros tipos de metodologias apresentadas na Figura 2 seguem estas premissas e seguem caminhos diferentes para obter os resultados.

Não obstante a forma de representação feita por Maimon e Rokach (2005), Fayyad, Piatetsky-Shapiro e Smyth (1996) preconiza que a maior parte dos métodos empregados em DM podem ser vistos como a composição de diversas técnicas e princípios básicos. Particularmente, algoritmos de mineração de dados consistem na mescla de três componentes Fayyad, Piatetsky-Shapiro e Smyth (1996):

- **O modelo.** Existem dois fatores relevantes na definição do modelo: a função

Figura 2 - Taxonomia da Mineração de Dados.



Fonte: Adaptação de Maimon e Rokach (2005).

do modelo e a forma representacional do modelo. Além disto, o modelo contém parâmetros que estão sendo determinados a partir dos dados.

- **O critério de preferência.** O critério de preferência é, usualmente, alguma função que mensure a qualidade do ajuste do modelo em relação aos dados, possivelmente, sendo constituída de algum limite de nivelamento a fim de evitar excessos de ajustes ou geração de um modelo com muitos graus de liberdade[2].

- **O algoritmo busca.** A especificação de um algoritmo para encontrar modelos e parâmetros específicos, tendo em conta os dados, um modelo (ou família de modelos), e um critério de preferência.

A literarura existente, no entanto, dificilmente faz uma diferenciação clara do modelo, critério de preferência e método de busca usado; estes, frequentemente são agregados a uma descrição única de um algoritmo em particular. Este reducionismo, como menciona Fayyad, Piatetsky-Shapiro e Smyth (1996), acaba tornando turva a contribuição independente de cada componente da método aplicado.

Além da definição citada, é válido dizer que existem algumas variações desta definição, uma vez que a mineração de dados se tornou um ambiente totalmente multidisciplinar. Todavia, todas acabam mantendo o cerne da busca por conhecimento em grandes conjuntos de dados. Entretanto, a forma como são colocadas estas definições e como os pesquisadores expõem a aplicação desta metodologia, pode levar a crer que o processo de extração de conhecimento se dá de forma totalmente automática, fato esse que não é verdade (LAROSE, 2014). Mesmo encontrando diversas ferramentas que auxiliem na execução dos algoritmos de mineração, os resultados ainda precisam de uma análise humana (CAMILO; SILVA, 2009). Conteúdo, a mineração de dados contribuiu imensamente no processo de KDD, permitindo que especialistas concentrem esforços apenas em partes mais significativas dos dados.

1.4 Tarefas de Mineração de Dados

Conforme apresentado na Figura 1, a Taxonomia da DM é composta por diversos ramos e, cada um, possui em suas raízes um conjunto de tarefas. Estas tarefas dizem respeito à aplicação de algoritmos específicos de acordo com o tipo de padrão que se

[2] Os graus de liberdade (DF) são a quantidade de informação que seus dados fornecem que você pode "gastar" para estimar os valores de parâmetros populacionais desconhecidos, e calcular a variabilidade dessas estimativas. Esse valor é determinado pelo número de observações e o número de parâmetros em seu modelo.

deseja garimpar. A seguir, são apresentados algumas das principais tarefas de Mineração de Dados e no Capítulo 3, é explicado mais detalhadamente uma destas tarefas.

1.4.1 Classificação

Uma das tarefas mais comuns, a classificação visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de aprender como classificar um novo registro (aprendizado supervisionado). Este modelo aprendido é então empregado para predizer um valor de entrada de novos exemplos (CAMILO; SILVA, 2009).

Exemplos de sua aplicação nos negócios e pesquisas incluem (LAROSE, 2014):

- Determinar se uma específica transação de crédito é fraudulenta;
- Alocar alunos em faixas particulares, no que diz respeito às necessidades especiais de cada um;
- Diagnosticar se doença em particular está presente no indivíduo; e
- Determinar se um juramento foi escrito pelo real falecido ou se por outra pessoa.

Uma das técnicas mais utilizadas em Classificação é a Árvore de Decisão. Nesta técnica, utiliza-se regras similares às tradiconais *if-then*, que recebem como entrada uma situação descrita por um conjunto de atributos e retorna uma decisão, que é o valor predizido para o valor de entrada. Seus atributos de insumo podem ser discretos ou contínuos (LAROSE, 2014).

As Árvores de Decisão são meios de representar os resultados encontrados na forma de árvore e que lembram um gráfico de estrutura organizacional. Cada folha está associada a uma classe, representando o valor do alvo mais apropriado. De forma alternativa, a folha pode conter um vetor de probabilidades indicando a probabilidade de um atributo alvo ter um certo valor. Instâncias são classificadas através do caminhar do algoritmo que, a partir das raízes da árvore (*Top*), vão "descendo" para suas fohas (*Down*), de acordo com o resultado dos testes ao longo da trajetória (MAIMON; ROKACH, 2005).

Uma das principais vantagens em utilizar as Árvores de Decisão é a sua forma de representação: "uma estrutura hierarquica que traduz uma árvore invertida a qual se desenvolve da raiz para as folhas"(FILHO, 2009). Tal estrutura traduz uma progressão da análise de dados no sentido de desempenhar uma tarefa de previsão, ou classificação.

Um exemplo prático desta técnica é a apresentada em Evans (2002), onde aponta o jogo *Black & White* como o primeiro a utilizar árvores de decisão com sucesso. No jogo,

Figura 3 - Exemplo de Árvore de Decisão Aletatória.

esta técnica foi implementada para representar as informações sobre as experiências que uma criatura do jogo tinha ao longo do evento, por exemplo, as experiências sobre que tipos de objetos foram ingeridos, quanto sua beneficência ou maleficência para a criatura. Neste caso, o indivíduo era capaz de tomar decisões sobre que tipo de objetos seriam mais apropriados, ou não, para se alimentar futuramente. Além disso, toda vez que a criatura realizava alguma ação, a reação do jogador era monitorada, usando a tupla (ação, resposta do jogador) como entrada do mecanismo de indução para construir a aŕvore de decisão que guiaria as ações futuras daquela criatura. Desta forma, ela passa a tender a realizar ações que foram recompensadas pelo jogador, evitando as demais.

1.4.2 Associação

A análise por associação é a tarefa de encontrar relações interessantes em grandes conjuntos de dados. Analogamente, sua intenção é expressar através de uma coleção de regras de associação e conjuntos de itens frequentes, relações inicialmente não pecerptíveis. Os conjuntos de itens frequentes podem ser compreendidos como uma coletânea de itens, ou eventos, que ocorrem frequentemente juntos. Assim, as regras de associação sugerem uma força de relacionamento que existe entre dois itens (LAROSE, 2014).

Matematiamente, as regras de associação são relacionamentos tais que, seja $I = \{I_1, I_2, \ldots, I_m\}$ um conjunto de itens. Seja D o conjunto de bancos de dados transacionais onde cada transação T é um conjunto de itens tal que $T \subseteq I$. Cada transação está associada a um identificador, chamado TID. Seja A um conjunto de itens. Uma transação T é dita conter A se somente se $A \subseteq T$. Uma regra de associação é uma implicação na

forma $A \Rightarrow B$, onde $A \subset I$, $B \subset I$ e $A \cap B = \emptyset$ (HAN; KAMBER, 2006).

Agregando-se a isto, para cada regra de associação é computado um fator de **suporte** e um fator de **confiança**. Estes fatores são meios de se quantificar o sucesso de uma análise por associação. Matematicamente, a regra $A \Rightarrow B$ se prende ao conjunto de transações D com **suporte** s, onde s é o percentual de transações em D que contêm $A \cup B$, ou seja, a união dos conjuntos A e B. Isto é considerado como sendo a probalidade de $P(A \cup B)$. Diz-se que a regra $A \Rightarrow B$ tem **confiança** c no conjunto transacional D, onde c é o percentual de transações em D que contêm A e que também contém B. Isto é considerado como sendo a probabilidade condicional $P(B|A)$ (HAN; KAMBER, 2006). Isto é,

$$\text{suporte } (A \Rightarrow B) = P(A \cup B) \text{ e} \tag{1}$$

$$\text{confiança } (A \Rightarrow B) = P(B|A). \tag{2}$$

Regras que satisfazem tanto o limíte mínimo de suporte (min_sup) e um limite mínimo de confiança ($conf\_min$) são chamados **fortes**. Por convenção, escreve-se valores de confiança e suporte entre 0 e 100%.

Em geral, algoritmos pertencentes a esta tarefa processam e dois grandes passos (HAN; KAMBER, 2006):

1. Encontrar todos os conjuntos de itens frequentes: por definição, cada um dos conjuntos de itens irão ocorrer pelo menos tão frequente quanto a contagem apoio mínima predeterminada, min_sup;

2. Gerar fortes regras de associação a partir dos conjuntos de itens frequentes: por definição, estas regras devem satisfazer o apoio mínimo e a confiança mínima.

Uma das aplicações deste método pode ser visto no trabalho Batista (2006), onde é estudada a associação entre fatores de risco genético e doenças em geral. A pesquisadora aplicou a análise de associação por meio de modelos de regressão logística em que, as análises de dados genéticos são abordadas via dados no nível genotípico e cromossômico.

1.4.3 Agrupamento

O Agrupamento, também conhecido como *Clustering* ou Agregação, é o processo de partição de uma população heterogênea em vários subgrupos, ou clusters, mais hete-

rogêneos. A principal diferença entre sua abordagem e a de Classificação é o desconhecimento prévio sobre o número de classes possíveis, e a possível pertinência dos exemplos usados na modelagem (LAROSE, 2014). (FILHO, 2009) descreve esta tarefa como uma técnica que agrupa um conjunto de itens, indivíduos ou objetos, sendo que os objetos incluídos em um mesmo agrupamento são os mais similares entre si e menos similares em relação aos objetos que estão em outros agrupamentos.

Alguns exemplos de aplicações desta tarefa são:

- Agrupamento de documentos relacionados para pesquisa;
- Agrupamento de genes e proteínas que tenham funcionalidade similar;
- Agrupamento de estoques com flutuações de preço similares; e
- Redução do tamanho de grandes bases de dados.

Os agrupamentos são realizados por meio de uma distância de similaridade (dissimilaridade). Dessa maneira, o usuário que realiza a análise deve possuir conhecimento suficiente sobre o problema, visando distinguir grupos úteis, necessários à realização de consultas. É considerado uma metodologia objetiva para quantificar uma característica estrutural de um conjunto de observações (FILHO, 2009).

Existem dois grandes grupos de métodos de Análise de Agrupamento (FáVERO et al., 2009):

- **Métodos Hierárquicos:** O agrupamento em classes procede por etapas, em geral determinando-se a partir de n subgrupos (de um único indivíduo cada) sucessivas fusões de subgrupoas consideradas mais "semelhantes". Cada fusão reduz, em uma unidade, o número de subgrupos.
- **Métodos não-Hierárquicos:** Fixa-se à partida o número k de classes que se pretende constituir e faz-se uma classificação inicial dos n indivíduos em k classes, ou determinam-se k "sementes"em torno das quais construir as classes. Através de transferências de indivíduos de uma classe para outra, ou de associações dos indivíduos às sementes das classes, procura-se determinar uma "boa"classificação, no sentido de tornar as classes mais internamente homogêneas e externamente heterogêneas.

Um exemplo de sua aplicação pode ser encontrada no trabalho Freitas (2006). Neste trabalho, são utilizadas técnicas de *cluster* hierárquico e *Sorting points into neighborhoods*[3] (SPIN) para estudar como se deu o histórico da evolução das espécies a partir

[3] SPIN significa Ordenação de pontos na vizinhança e é a denominação de uma técnica utilizada para tratamento de dados de múltiplos objetos em aglomerados.

do estudo de uma única proteína comum em diversos seres vivos, a **Lisozima**.

Este tipo de análise será mais aprofundado no Capítulo 3, constituindo-se a tarefa principal da aplicação deste trabalho.

2 MINERAÇÃO DE TEXTOS

2.1 Descoberta de Conhecimento em Dados não Estruturados

O processo de KDD citado no Capítulo 1 funciona de maneira eficiente para dados armazenados de forma estruturada, entretanto, a quantidade de documentos não estruturados que trafegam nas redes é infinatamente maior. Assim, descobrir uma forma de extrair informações úteis deste tipo de dado exije um trabalho diferenciado, e não raramente árduo. A dificuldade de trabalhar estes itens advem do fato de que, informações encontradas em textos são usualmente insatisfatórias como respostas, por serem extensas e, por vezes, prolixas (ARAUJO, 2007).

Para lidar com esta dificuldade, Jones (1997) preconiza a existência de técnicas e ferramentas da área de Recuperação de Informação (RI) que podem auxiliar na obtenção de informações relevantes junto a textos não estruturados. Entretanto, como é exposto por Chen et al. (1994), ferramentas de RI costumam produzir um conjunto massivo de documentos como resposta, acarretando a sobrecarga de informações, que acontece quando o usuário tem muita informação ao seu alcance, mas não dispõem de recursos ou meios para tratá-las, dificultando, ou impossibilitando, o encontro da informação desejada.

Neste contexto, o desenvolvimento de metodologias para o estudo de casos desta natureza foi amadurecendo, evoluindo para o surgimento da área da Descoberta de Conhecimento em Textos (*Knowledge Discovery from Text* - KDT). Araujo (2007) expõem que o termo foi utilizado pela primeira vez por Feldman e Dagan (1995) para designar o processo de encontrar algo interessante em coletâneas de textos (artigos, jornais, e-mails, páginas Web, etc.). Atualmente, esta área é muito conhecida pelo nome de *Text Mining* ou Mineração de Textos.

A Mineração de Textos é um método interdisciplinar que envolve, além da área de RI, aprendizagem de máquina, estatística, linguagem natural e mineração de dados (MACHADO et al., 2010). Cada uma dessas áreas, contribui com técnicas e ferramentas inteligentes e automáticas que visam auxiliar na análise de grandes volumes de dados com a finalidade de "garimpar"conhecimento útil, beneficiando não somente usuários de documentos eletrônicos da Web, mas qualquer domínio que utiliza textos não estruturados (MORAIS; AMBRóSIO, 2007).

Usualmente, a metodologia empregada em Mineração de Textos, enbloba quatro fases (TAN et al., 1999):

- **Coleta de documentos** - etapa de busca dos dados que serão analisados a fim de se retirar o conhecimento desejado.

- **Pré-processamento** - conjunto de ações tomadas sobre os documentos textu-

ais a fim de tonrá-los manipuláveis para a extração do conhecimento. Em geral, "tem como resultado a padronização dos documentos em um formato atributo-valor" (ARAUJO, 2007).

- **Extração de conhecimento** - utilização de algoritmos de aprendizagem com a finalidade de extrair conhecimento na forma de regras de associação, relação, segmentação, classificação de textos, entre outras.

- **Análise de resultados** - etapa final cujo objetivo é a interpretação e análise dos resultados obtidos na fase anterior.

2.2 Coleta de Documentos em páginas HTML

Esta é a primeira etapa do processo de descoberta do conhecimento e tem a finalidade de buscar, formatar e armazenar os documentos obtidos para que possam ser processados nas atividades seguintes. Ve-se, então, que esta atividade está intimamente ligada ao desejo do pesquisador quanto aquilo que se quer estudar, ou se obter, ao final do processo.

Visto que o intuito deste trabalho foi a análise de dados não estruturados disponíveis na internet, fez-se necessário o estudo de formas inteligentes que auxiliassem na obtenção dos dados para que fossem posteriormente analisados. Diante disto, surgiu a ideia de utilizar técnicas de *web crawling* com a finalidade de obter informações textuais à partir de páginas web escritas em HTML[4].

Reis (2013) preconiza que um ambiente Web possui diversas conexões, podendo ser analisado como um grafo direcionado, conexo e esparso, cujos vértices correspondem à páginas e as arestas correspondem aos hiperlinks que as conectam. Assim, cada página X que tem um link com uma página Y pode ser enxergada como duas arestas que se conectam, e sua direção é dada uma vez que nem toda página Y possuirá um link de retorno para X e, caso possua, este link será descartado durante o processo de *web craling*. Este modelo é denominado **grafo da web**.

Ainda pelo autor, é apresentado algumas propriedades relevantes do grafo da web:

- páginas que estão conectadas no grafo, possuem uma relação de recomendação, uma vez que o autor da página-fonte, X, referenciou a página-alvo, Y, usando um link (ou hiperlink);

[4] Abreviação de *HyperText Markup Language*, é uma linguagem de marcação utilizada, principalmente, na construção de páginas web.

- páginas que estão ligadas umas as outras são propensas a possuir conteúdos relacionados ou similares; e
- o texto do hiperlink de uma página-fonte (também conhecido como texto âncora) geralmente descreve a página-alvo a qual ele referencia.

Algoritmos rastreadores web exploram esta estrutura a fim de identificar os caminhos que deverão ser trilhados a fim de se obter a informação que se deseja. Tais algoritmos são conhecidos por nomes como *web crawler*, *web robot*, *web spider* entre outros; todavia, todos se referem a um algoritmo que possui a mesma dinâmica de funcionamento e finalidade (PAES, 2012).

Estes algoritmos iniciam sua busca recuperando um conjunto de páginas informadas inicialmente, analisando, além de seu conteúdo, hiperlinks contidos nessas páginas. A esse conjunto inicial de entrada dá-se o nome de **sementes**. Posteriormente, o rastreador recupera as páginas-alvo ainda não visitadas referentes aos hiperlinks obtidos. Este processo se repete até que um número delimitado de páginas seja recuperada ou até que um condição pré-estabelecida ocorra. Cada repetição deste processo é denominada **iteração de busca** (REIS, 2013).

2.2.1 Estruturação Básica de Páginas HTML

A linguagem HTML é uma linguagem basicamente para a escrita de páginas Web, assim, quando se quer construir um rastreador específico, deve-se estudar a taxonomia do ambiente cujo as informações serão recuperadas. Desta forma, deve-se entender estrutura utilizada pelo designer para criação destas páginas, a fim de, posterioremnte, construir um algoritmo que consiga extrair, automaticamente, informações textuais contidas nestas páginas.

Uma página em HTML possui quatro conceitos fundamentais (UFRJ, 2015):

- **elemento:** uma estrutura (como parágrafo, lista, tabela, etc.) utilizadas em sua composição;
- **etiqueta:** delimitador do início e fim de um determinado elemento;
- **atributo:** característica do elemento que pode ser configurada através da especificiação; e
- **valor:** especificação atribuída a um atributo.

Sua estrutura básica, a partir do qual serão inseridos outros elementos, é composta por três partes: **estrutura principal** - delimitada pelas etiquetas <html> e < /html>;

cabeçalho - delimitado pelas etiquetas <head> e < /head>; e o **corpo** - delimitado pelas etiquetas <body> e < /body>. No corpo é inserido quase em totalidade o conteúdo das páginas, através dele, são definidas propriedades gerais para toda a página, como a cor do plano de fundo ou a cor dos links que serão expostos (UFRJ, 2015).

```
<html>
    <head>
        <title>Pagina Xpto</title>
    </head>
    <body>
        Corpor do Documento. Texto, tabelas, figuras, etc...
    </body>
</html>
```

Além dos elementos básicos utilizados para a estruturação um HTML, outros elementos são largamente usados, como os **cabeçalhos do corpo**. O cabeçalho dorpo da página é usado para introduzir o texto que será elaborado, entretanto, é empregado outros tipos de cabeçalhos no decorer do HTML com a finalidade de organizar e melhorar a apresentação do conteúdo que se quer expor na página. Para sua representação existem as tags "**h**", que possuem seis níveis (<h1> e < /h1>, <h2> e < /h2>, ... e <h6> e < /h6>) e variam a formatação de seu conteúdo de acordo com o tema utilizado (DEVMEDIA, 2014).

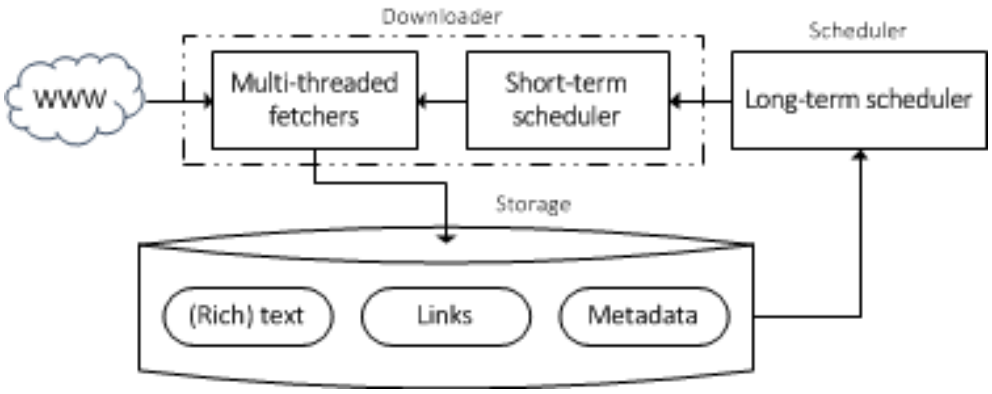
```
<h1> Titulo nivel 1 </h1>
<h2> Titulo nivel 2 </h2>
<h3> Titulo nivel 3 </h3>
<h4> Titulo nivel 4 </h4>
<h5> Titulo nivel 5 </h5>
<h6> Titulo nivel 6 </h6>
```

Naturalmente, há páginas que a quantidade de informação é mais extensa, exigindo uma divisão do conteúdo através de parágrafos, com a finalidade de organizar e apresentar melhor o as informações. Durante a codificação de um texto HTML, utiliza-se as tags <**p**> e < /**p**> para indicar o início e término de um novo parágrafo.

Outro elemento de destaque, principalmente para o foco do trabalho, são os **links**. A HTML possui uma tag chamada âncora representada por <a>, cujo um dos atributos é o "**href**". A exemplo, para o endereço "http://dissertacao.com.br" ser atribuída a palavra "dissertacao", o código HTML seria:

```
<a href=''http://dissertacao.com.br''>Dissertacao</a>
```

Figura 4 - Típica arquitetura alto-nível de um Web Crawler, envolvendo *Scheduler* e um *Downloader*.



Fonte: Obtido em Castillo e Baeza-Yates (2010).

Usualmente os sites possuem uma estrutura padrão, variando apenas seu conteúdo. Devemos, então, aproveitar este fato para que, ao extrairmos as informações através do uso de rastreadores web, estas, sejam devidamente tratadas, garantindo que a informação recuperada contenha exatamente o conteúdo que dever-se-á ser estudado nas próximas etapas da Mineração de Textos.

2.2.2 Arquitetura de Web Crawlers

É notório que algoritmos rastreadores são fundamentais, e por vezes peça principal, para o desenolvimento de motores de busca. Consequentemente, empresas que investem em pesquisas desta natureza acabam mantendo em sigilo o detalhamento da lógica e arquitetura utilizada para a construção de seus programas. Assim, quando um novo *design* de Crawler é publicado, frequentemente são omitidos detalhes importantes a fim de previnir a replicação daquele trabalho. Entretanto, Castillo e Baeza-Yates (2010) preconiza a existência de uma arquitetura base para o desenvolvimento de rastreadores web de alto-nível (Figura 4).

Esta arquitetura "padrão" possui como entrada um conjunto de páginas WWW, que são processadas através de um ou mais módulos denominados "*Downloaders*". Este módulo é responsável pelas operações de coleta de informação na rede e envio de dados para o "*Storage*". O sistema de storage funciona como um armazém de dados e é responsável pelo compartilhamento da informação recebida pelo Downloader para um módulo conhecido como "*Frontier*". O storage pode, ainda, ser compartilhado de forma parcial ou completa, ou seja, seus dados podem ser tramitados totalmente ou parcialmente entre os módulos que estão ligados a ele. O frontier é o módulo responsável pela retroali-

mentação do Downloader, enviando as URLs obtidas através de atividades passadas para um ou mais Downloaders. Durante esta atividade, é comum a utilização de uma regra para priorizar e ordenar estas URLs que deverão ser visitadas, variando de acordo com a intenção do desenvolvedor.

Vê-se ainda na Figura 4 que o storage pode ser subdividido em três partes: **texto** - composto por documentos formatados ou documentos em HTML completos; **metadata** - também conhecido como metainformação, são dados e ou informações acerca de outros dados; e **links**. No caso de web crawlers focados, o texto se torna importante para a priorização e classificação das páginas e, geralmente, os metadatas e links são suficientes para decidir a sequência de páginas que deverão ser visitadas.

A arquitetura apresentada é bastante generalista, Chakrabarti (2002) descreve em seu trabalho detalhadamente a arquitetura de um web crawler, apresentando pontos que podem diferir de um algoritmo para outro, como as técnicas de *parsing* ou busca por textos duplicados.

2.2.3 Recuperando Informações HTML

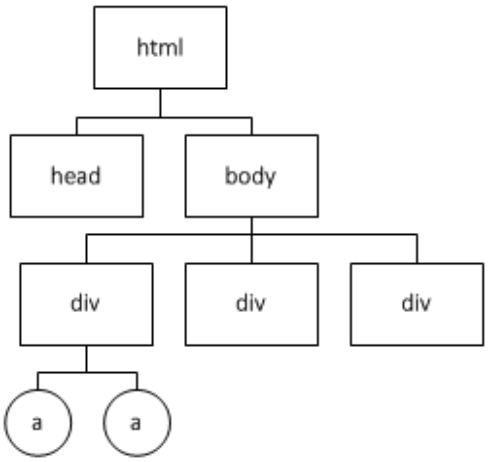
A maioria das páginas na web estão em HTML, o que é ótimo quando se quer obter informações com o auxílio de rastreadores. Entretanto, nem sempre são seguidas as boas práticas de formatação e classificação das páginas, dificultando bastante o trabalho daqueles que querem recuperar conteúdos específicos, ou seja, que estão embutidos em algum trecho daquele código. Desta forma, a maneira mais confiável de obtermos estas informações é utilizando técnicas de ***parsing*** do conteúdo HTML.

Parsing, também conhecido como análise sintática, corresponde ao processo de analisar uma sequência de tokens de entrada para determinar sua estrutura gramatical segundo uma determinada gramática formal (MARCUS, 1980). Na prática, técnicas desta natureza analisam um conjunto de entrada sob a ótica de uma estrutura similar a uma árvore, onde cada nó corresponde a uma "categoria", facilitando o manuseio e manipulação dos dados recebidos.

Em especial, quando fala-se em parsing de HTML, a estrutura da árvore de um parsing é composta por nós que representam as tags utilizadas pelo desenvolvedor durante a sua construção. Assim, ao pretendermos recuperar informações de conteúdos específicos de um página, devemos analisar sua estrutura a fim de utilizarmos o parsing para a extração das informações desejadas, expurgando aquilo que não faz sentido para as análises futuras.

A exemplo, pode-se ver na Figura 5 a árvore de representação para o trecho HTML abaixo:

Figura 5 - Árvore estrutural de uma
página HTML fictícia.




```
<html>
    <head>
        <title>Pagina Xpto</title>
    </head>
    <body>
        <div id=‘‘content’’>
            <div class=‘‘content-title’’>
                Conteudo especial de
                <a class=‘‘link’’ href=‘‘http://www.google.com/’’
                   >exemplo</a>,
                para um texto sem fundamento usado apenas
                na <a class=‘‘link-old’’> dissertacao</a>.
            </div>
            <div class=‘‘content-author’’>
                Escrito por mim.
            </div>
            <div class=’’content-data’’>
                14/11/15
            </div>
        </div>
        <div class=‘‘footer’’>
            Nenhum direito reservado.
        </div>
    </body>
</html>
```

Diante dos fatos abordados, foi pesquisado para o desenvolvimento deste trabalho uma linguagem que possibilitasse tanto a construção de um web crawler quanto a manipulação do conteúdo obtido. Com isso, foi adotada a linguagem **Python**, uma vez que se mostrou extramamente eficiente, com pacotes já elaborados e de fácil utilização, tanto para obtenção dos dados quanto para aplicação de técnicas de parsing e exportação do conteúdo. Em particular, foram utilizadas como alicerce do algoritmo, duas bibliotecas:

- **requests:** desenvolvida para facilitar a integração com os serviços web, mitigando a necessidade de intervenções manuais para a realização de requisição ou *post* de conteúdos, sendo assim, essencial para a captura dos dados HTML;
- **BeautifulSoup:** tem a finalidade de extrair dados a partir de arquivos HTML e XML[5] usando técnicas de parsing, propiciando uma maneira fluida de navegação, busca e manipulação da árvore de representação recebida como entrada.

Para o desenvolvimento deste trabalho, foi construído um algoritmo em Python que tem como entrada um único site e, a partir dele, são levantados os hiperlinks e conteudos das páginas que deverão ser acessados. A partir deste input, o conteúdo em formato HTML é recuperado através do comando a seguir e seu conteúdo armazenado em um objeto denominado *r_start*.

```
r_start = requests.get(``http://blogs.oglobo.globo.com/miriam-
    leitao/")
```

O arquivo recuperado inicialmente, é responsável pelo norteamento dos próximos passos que deverão ser dados pelo crawler. Entretanto, o arquivo obtido desta maneira está em HTML, possuindo um enorme número de informações desnecessárias e que poluem a análise de seu conteúdo. Assim, para retirar as informações necessárias, aplica-se o comando a seguir, a fim de parsear o arquivo obtido e extrair informações de hiperlinks e conteúdos textuais contidos na URL inicial.

```
soup = BeautifulSoup(link.content, 'lxml')
```

Utilizando estas ferramentas como motores para a criação de crawlers, resta apenas estudar a estrutura da página que se quer analisar e exportar as informações desejadas

[5] Do inglês *eXtensible Markup Language*), é uma linguagem de marcação recomendada pela W3C para criação de documentos com dados organizados hierarquicamente, tais como textos, banco de dados ou desenhos vetoriais. A linguagem XML é classificada como extensível porque permite definir os elementos de marcação.

em um formato favorável para as atividades seguintes. No Capítulo 4, será detalhado o desenvolvimento de um web crawler construído para uma aplicação específica, onde será mais facilmente compreendida a interação dos conceitos apresentados nesta seção para o desenvolvimento de rastreadores web.

2.3 Pré-processamento de Textos

A etapa de pré-processamento destina-se à extração de textos escritos em linguagem natural, obviamente não estruturadas, uma representação estruturada, concisa e manipulável através de algoritmos (CORREA; MARCACINI; REZENDE, 2012). Em resumo, inicialmente são realizadas atividades de tratamento e padronização da coletânea de textos, seguidas do estudo de termos mais significativos, finalizando com uma representação da coletânea textual em um formato estruturado, preservando características de interesse para o trabalho. Esta atividade irá variar de acordo com o que se deseja obter de informação a partir dos dados que serão analisados.

O pré-processamento, por muitas vezes, é o processo mais oneroso da metologia de Mineração de Textos, uma vez que não existe uma única técnica a ser aplicada para a obtençãode uma representação satisfatória em todos os meios e pesquisas, exingindo-se a realização de vários experimentos empíricos para se chegar a um representação adequada citebib:paes2012.

2.3.1 Padronização de Termos

Os documentos obtidos na etapa anterior, não raramente resulta em um coletânea de arquivos com formatos diferenciados, uma vez que existem diversos aplicativos destinados a geração e publicação de textos eletrônicos. Por isso, convém realizar a conversção dos textos para um formato plano sem formatação, garantindo que as análises que serão realizadas não gerem resultados espúrios em decorrência da falta de padronização dos dados.

Além disto, Correa, Marcacini e Rezende (2012) expõe que uma das maiores dificuldades enfrentadas pelo processo de mineração de textos é a elevada dimensionalidade dos dados. Uma pequena coletânea textual pode facilmente conter milhares de termos, muitos deles redundantes ou desnecessários, tornando as etapas posteriores lentas e comprometendo a qualidade dos resultados.

A **seleção de termos** é aplicada para driblar este problema, tendo como propósito a obtenção de um subconjunto conciso e representativo de termos da coletênea textual recebida. Para isso, Manning et al. (2008) cita algumas técnicas que auxiliam nesta

tarefa, como a eliminação de termos comuns (*stopwords*) e a aplicação de tecnicas de Normalização, como *stemming* e lematização.

2.3.1.1 Remoção de *Stopwords*

Stopwords são aquelas palavras empregadas com elevada frequência em documentos textuais e que, por serem muito comuns, acabam não contribuindo significativamente para a seleção e determinação do conteúdo do documento (MANNING et al., 2008). Wives (1999) apresenta estes termos como **palavras negativas** e preconiza que mantê-las, prejudica a qualidade da análise e torna o processo mais moroso, visto a frequência com que estes termos aparecem nos documentos que serão processados.

Geralmente, a estratégia de eliminação destes termos é através da criação de uma lista de *stopwords*, conhecida como *stoplist*. Em seguida, percorre-se a coletânea de entrada elimando os termos que compõem a *stoplist*. Uma *stoplist* usualmente é constituída por **artigos**, **preposições**, **pontuação**, **conjunções** e **pronomes** de uma língua. A identificação e remoção desta classe de palavras reduz de forma considerável o tamanho final do documento, tendo como consequência benéfica o aumento de desempenho do sistema que se está criando para a análise da coletânea.

A *stoplist* pode ser definida manualmente por um especialista no domínio do assunto ou de forma automática, através da frequência de aparição dos termos no documento. Um percentual T das palavras de maior aparição define a lista de remoções. Para este trabalho foi usada como base a *stop list* presente na biblioteca **tm** disponível para o software **R**. Sua adoção como recurso para a aplicação se mostrou pertinente, otimizando o tempo e a qualidade da fase de eliminação das *stopwords*.

2.3.1.2 Normalização

Normalização é o processo que visa reduzir a quantidade de termos diferentes através de um agrupamento de palavras que compartilham de um mesmo padrão. Existem diversas abordagens de agrupamento, como a identificação morfológica do termo ou o reconhecimento de sinônimos e conceitos similares (CARRILHO, 2007).

A aplicação de técnicas de Normalização introduz uma melhora significativa nos sistemas de Mineração de Textos. Esta melhora varia de acordo com o escopo, o tamanho da massa textual e o que se pretende obter como saída do sistema. De acordo com a forma de agrupamento das realizações das palvras, os processos de normalização podem ser de vários tipos. Os principais são: ***Stemming*** e **Lematização**.

Por razões gramaticais, uma mesma palavra pode ser empregada de maneiras diver-

sas, como organizar e organização. Além disso, existem famílias de palavras com escritas e sentidos similares, como construto, construção e construtor. Em muitas situações, é conveniente utilizar esta característica para buscar, a partir de uma palavra, documentos com possíveis similaridades.

O objetivo do *stemming* e da lematização é sintetizar palavras flexionadas, e ou similares, a uma forma base, reduzindo o conjunto de palavras distintas contidas na coletânea inicial (MANNING et al., 2008). Assim, este mapeamento dos textos resultará em palavras do tipo:

- Os garotos dos carros amarelos ⇒ O garot do carr amarel.

Manning et al. (2008) define *stemming* como um "processo heurístico bruto", geralmente carcaterizado pelo corte das extremidades de palavras na esperança de atingir seu objetivo corretamente a maior parte das vezes. Para isso, não raro são removidos os morfemas derivacionais[6] ligados aquelas palavras. Em outras palavras, esta técnica concentra-se na redução de cada termo do documento, até que seja obtida sua respectiva raiz. Seu benefício se dá pela eliminação de sufixos que indicam variação na forma da palavra, como plural e tempos verbais. A seguir, seguem os principais métodos apresentados por Carrilho (2007):

- **Método de *Stemmer* S:** Método simples que foca apenas em algumas poucas terminações de palavras do inglês. Os principais sufixos a serem removidos são: *ies*, *es* e *s*. Embora este método não introduza muito impacto nos documentos, é bastante utilizado por seu caráter conservador e que raramente surpreende negativamente o usuário. Seu uso pode ser aliado a um tradutor de conteúdo para aplicá-lo em documentos escritos em outras linguas.

- **Método de *Porter*:** Método que se concentra na identificação das diversas formas e inflexões referentes à um mesmo termo e sua substituição por um radical comum. A exemplo, as palavras correr, corrida e corrido compartilham de um mesmo radical e, aplicando-se este método, todas as palavras seriam reduzidas a um radical comum "corr".

- **Método de *Lovins*:** Método de passo único. É Sensível ao contexto e abrange uma gama maior de sufixos. Baseia-se em uma lista de regras, chamada de regras de

[6] Afixos participam em processo de formação de palavras e determinam a classe gramatical da palavra em que ocorrem. A exemplo, o sufixo **-idade** origina apenas nomes e agrega-se apenas a bases adjetivais: formalidade, vivacidade, materialidade, entre outros. Ademais, os morfemas derivacionais são responsáveis pela determinação do valor das categorias morfológicas, morfossintáticas e morfossemânticas relevantes. A exemplo, os sufixos empregados a palavra **formal**, como formalizado ou formalizar.

Lovins e que, num passo único, faz a remoção de, no máximo, um único sufixo por palavra. Apesar de não incluir vários sufixos em sua abordagem, é o mais agressivo dos algoritmos apresentados.

O processo de *Lematização* pode ser entendida como um processo que objetiva remover de forma otimizada as terminações de inflexão das palavras, retornando-as a uma forma base conhecida como *lema* (MANNING et al., 2008). Este processo, diferente do *stemming*, baseia-se na utilização de um vocabulário pré-determinado e na análise morfológica das palavras, reduzindo de maneira mais eficiente os termos contidos em uma coletânea. Devido sua aplicação exigir um trabalho de análise mais apurado e robusto, por vezes, acaba sendo inviabilizado dependendo da variedade e massa textual contidas na coletânea estudada.

Outros métodos também costumam ser usados com a finalidade de tornar os documentos mais polidos, tais como:

- Remoção de acentos quando se estuda línguas com este recurso gramatical;
- Remoção de números e termos relacionados, como $ ou %;
- Transformação dos termos para maiúsculo ou minúsculo, garantindo que termos iguais não sejam interpretadas como distintos devido a isto;
- Remoção de espaços extras no meio do documento;
- Remoção de pontuações quando não se quer analisar semanticamente os termos que compõem a coletânea; e
- Análise de palavras compostas para que não sejam tratadas como termos independentes.

2.3.2 Seleção de Termos

O método de Luhn (LUHN, 1958) é uma técnica tradicional para a seleção de termos utilizando a frequência dos termos. Conforme é exposto por Rezende, Marcacini e Moura (2011), este método foi baseado na Lei de Zipf, também conhecida como Princípio do Menor Esforço. Para dados textuais, ao realizar a contabilização da frequência dos termos e ordená-los, o histograma resultante em ordem decrescente, forma-se a chamada Curva de Zipf, na qual o "k-ésimo termo mais comum ocorre com frequência inversamente proporcional a k. Os termos de alta frequência são julgados não relevantes por geralmente aparecerem na grande maioria dos textos, não trazendo, em geral, informações úteis"

(REZENDE; MARCACINI; MOURA, 2011). Em contrapartida, os termos de com menor frequência são considerados muito raros e não possuem caráter discriminatório. Em decorrência disto, são definidos pontos de corte superior e inferior da Curva de Zipf, de maneira que "termos com alta e baixa frequência são descartados, considerando os termos mais significativos os de frequência intermediária" (REZENDE; MARCACINI; MOURA, 2011).

Notoriamente, este método é facilmente escalável para bases de dados muitos grandes, visto sua baixa exigência de processamento. Entretanto, como Rezende, Marcacini e Moura (2011) menciona, este método é extramamente subjetivo, visto que os pontos de corte irão ficar a critério do pesquisador.

2.3.3 Matriz Termo-Documento

Uma vez selecionados os termos mais frequentes e já tendo sido realizadas as técnicas para redução e limpeza dos dados, deve-se buscar a estrturação dos documentos, de maneira a torná-los processáveis pelos algoritmos que virão (REZENDE; MARCACINI; MOURA, 2011). O modelo mais usado para representação dos dados é o modelo de espaço vetorial. O modelo espaço vetorial é uma representação matemática de **termos** e **documentos** de uma **coletânea** textual (BERRY; CASTELLANOS, 2007). Nesta abordagem, cada componente do vetor-documento expressa, numericamente, a importância semântica de um termo presente. A coletânea é modelada por meio de uma matriz chamada **termo-documento**. Esta forma de visualização é usualmente utilizada para os próximos passos da Mineração de Textos e pode ser um dos resultados esperados do pré-processamento.

Um coletânea composta de n documentos indexados por m termos pode ser representada por uma matriz termo-documento A de ordem $m \times n$. Os vetores-documento estão dispostos como colunas na matriz A e cada elemento $a_{i,j}$ representa a frequência ponderada que o termo i ocorre no documento j (BRAGA, 2011). Inexiste uma interpretação específica para cada vetor representado nas coluna da matriz, uma vez que a combinação de quaisquer dois documentos não produz necessariamente um documento viável. Todavia, usar esta estrutura possibilita explorar poderosas relações geomátricas e algébricas entre termos e documentos (vetores) para avaliar semelhanças e diferenças semânticas de conteúdo. Como exemplo, a Tabela 1 mostra a preparação (análise + remoção de *stopwords* + processo de *stemming*) de uma coletânea de documentos.

Com a coletânea preparada, os documentos poderão ser estruturados na matriz exposta na Tabela 2.

Esta representação, a matriz é constituída pela frequência simples de ocorrências dos termos em cada documento (coluna). Entretanto, Berry e Castellanos (2007) preco-

Tabela 1 - Preparação de uma pequena coletânea textual.

Documento		Conteúdo Original		Termos
D_1	$\rightarrow$	Letras brasileiras	$\rightarrow$	Letr brasil
D_2	$\rightarrow$	Corredores do brasil	$\rightarrow$	Corredor brasil
D_3	$\rightarrow$	Controle estatístico e de letra	$\rightarrow$	Control estatístic letr

Tabela 2 - Exemplo de Matriz Termo Documento.

	D_1	D_2	D_3
Letr	1	0	1
brasil	1	1	0
Corredor	0	1	0
Control	0	0	1
estatístic	0	0	1

niza em seu trabalho que ponderar os termos otimiza potencilamente o desempenho na recuperação da informação. Assim, o autor sugere a ponderação dos elementos da matriz no formato mostrado na equação 3, para parametrizar aspectos elevantes na recuperação.

$$a_{ij} = l_{ij} g_i d_j \tag{3}$$

O fator l_{ij} representa o peso local para o termo i presente no documento j e regula a importância de cada termo internamente ao documento, salientando sua essencia semântica. O parâmetro g_i reflete a ponderação global do termo i na coleção considerando conteúdos individuais no ambiente mais amplo da coleção e atuando como moderador da heterogeneidade da base. Finalmente, d_j especifica a normalização aplicada nos documentos e estabelece um patamar homogêneo na avaliação dos documentos. Modelos reconhecidos para equacionamento dos elementos podem ser encontrados em Berry e Castellanos (2007), sendo a_{ij} o produto final.

Com esse entendimento, neste exemplo básico será usada a frequência de ocorrência simples no p eso local sem considerar ponderação global e normalização, assumindo portanto valor unitário. Neste contexto, o valor de a_{ij} é dado pelo número de ocorrências do termo i no documento j. Intuitivamente esta configuração simples atende bem a abordagens iniciais em bases de dados desconhecidas e serve de parâmetro de comparação na avaliação de outros modelos mais sofisticados.

3 ANÁLISE DE AGRUPAMENTO (*CLUSTERING*)

3.1 Introdução a Análise de Conglomerados

A técnica de análise de conglomerados (*clustering*), também conhecida como análise de agrupamentos, é um método estatístico de interdependência que permite agrupar casos ou variáveis em grupos homogêneos em função do grau de similaridade entre os indivíduos, a partir de critérios predeterminados. Sua ideia princial é agrupar objetos com base em suas próprias características, buscando, desta forma, a estrutura "natural" desses termos (FáVERO et al., 2009).

Esta metodologia pode ser aplicada em diversas áreas do conhecimento cujo objetivo seja segmentar as observações em grupos homogêneos internamente e heterogêneos entre si. Quando aplicada na Mineração de Textos, seu uso é bastante extenso, podendo auxiliar na indexação de documentos por conteúdo, evidenciar automaticamente palavras-chave ou auxiliar na elaboração de resumos da coletânea textual.

A análise de conglomerados é uma importante técnica exploratória, uma vez que, ao estudar a estrutura natural de grupos, possibilita avaliar a dimensionalidade dos dados, identificar *outliers*[7] e levantar hipóteses relacionadas à estrutura (associações) dos objetos (JOHNSON; WICHERN et al., 2007).

As variáveis estatísticas de agrupamento podem ser definidas como o conjunto de atributos ou características das observações que servirão de base para a determinação da similaridade entre elas. É válido destacar ainda que, esta metodologia difere de outras técnicas multivariadas quanto a seleção da variável estatística. Enquanto outras técnicas estimam empiricamente a variável estatística, nesta, a variável é representada por um conjunto de variáveis selecionadas diretamente pelo pesquisador (FáVERO et al., 2009).

Sharma (1996) apresenta a técnica de *clustering* geometricamente, o que contribui bastante para os conceitos que se seguirão. O autor argumenta que cada observação de um conjunto de dados pode ser observada como um ponto em um espaço p-dimensional, onde p corresponde ao número de variáveis ou características usadas para descrever o sujeito. Neste contexto, pode-se representar a Tabela 3 geometricamente sob a forma da Figura 6, onde cada variável corresponde a um ponto em um espaço bidimensional. Suponha agora que se quer aglutinar as observações em três grupos homogêneos. Visualmente a imagem sugere que $S1$ e $S2$ formem um grupo, $S3$ e$S4$ outro grupo, e $S5$ e $S6$ um terceiro grupo.

[7] Em estatística, *outlier*, é o nome dado a valores aberrantes ou atípicos em um conjunto de dado, podendo ser interpretado como uma observação que apresenta grande afastamento dos demais dados da série.

Tabela 3 - Dados hipotéticos.

Id	Salário (R\$1.000)	Educação (anos)
S_1	5	5
S_2	6	6
S_3	15	14
S_4	16	15
S_5	25	20
S_6	30	19

Entretanto, utilizar estes procedimentos gráficos para identificar conglomerados se torna impraticável quando se possui muitas observações ou quando existem mais de três variáveis ou características. O que é necessário, nesses casos, é a aplicação de técnicas analísticas adequadas para identificar grupos ou conglomerados de pontos em quaisquer espaço p-dimensional.

Neste trabalho, serão apresentados os conceitos para aplicação desta metologia segundo as etapas sugeridas pelo autor Fávero et al. (2009):

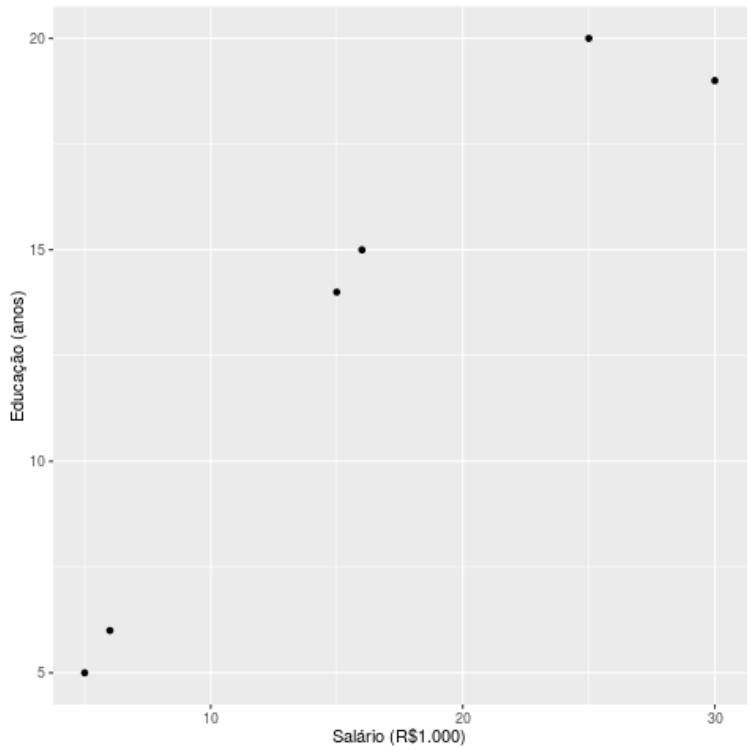
- análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificaçãode *outliers* e padronização);
- seleção da medida de similaridade entre cada par de observações;
- seleção do algoritmo de agrupamento: método hierárquico ou não-hierárquico;
- definição da quantidade de conglomerados formados; e
- interpretação e validação dos conglomerados.

Estas etapas, possuem grande semelhança com as etapas da Mineração de Dados, semelhança esta que contribui para sua aplicação nesta temática. A primeira etapa apresentada pode ser assemelhada a fase de tratamento da coletânea textual, onde os documentos são coletados e padronizados (pré-processamento); a última etapa corresponde ao julgamento dado após o processamento da Mineração dos Dados; e as atividades meio, podem ser equiparadas a fase de extração da informação, ou pós-processamento.

3.2 Medidas de Similaridade e Dissimilaridade

O conceito de similaridade em *clustering* é o cerne desta metodologia, uma vez que a identificação de agrupamentos de sujeitos ou variáveis só é possível com a adoção

Figura 6 - Plot de dados hipotéticos.



de alguma medida de semelhança que permita a comparação objetiva entre os sujeitos. Quando as observações são agrupadas, a medida de proximidade é usualmente indicada por alguma medida de distância. Não obstante, o agrupamento das variáveis pode, ainda, ser feito através de medidas de correlação ou de associação (JOHNSON; WICHERN et al., 2007). Fávero et al. (2009) preconiza que a escolha das medidas de similiradade implica no conhecimento da natureza das variáveis (discreta, contínua, binária) e da escala de medida (nominal, ordinal, intervalar ou razão).

Geralmente, estas medidas são dispotas em uma matriz $\mathcal{X}, p$, onde n representa os objetos e p as variáveis. A similaridade entre os objetos é descrita por uma matriz $D_{n,n}$ (HäRDLE; SIMAR, 2003) descrita na (4).

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}. \tag{4}$$

A matriz $\mathcal{D}$ contém as medidas de similaridade ou dissimilaridade entre os n objetos. Se os valores d_{ij} são distâncias, então D é donominada **Matriz de Dissimilaridade**; quanto maior a distância entre os objetos, menor a similaridade entre eles. Se os valores

d_{ij} são medidas de proximidade, então a recíproca é verdadeira, isto é, quanto maior o valor de proximidade, mais similares são os objetos (HäRDLE; SIMAR, 2003).

3.2.1 Medidas de Distância

Como supracitado, as medidas de distâncias são consideradas medidas de dissimilaridade, uma vez que quanto maiores seus valores, menor a semelhança entre os objetos, e vice-versa. Além disto, para que uma função seja de *distância* é necessário e suficiente que as seguintes condições sejam satisfeitas, para quaisquer objetos i, j, k (AMO, 2003):

1. $d(i,j) \geq 0$;
2. $d(i,i) = 0$;
3. $d(i,j) = d(j,i)$ (simetria); e
4. $d(i,j) \leq d(i,k) + d(k,j)$ (desigualdade triangular).

A seguir são apresentadas algumas das principais medidas de distâncias aplicadas em análise de conglomerados são (JOHNSON; WICHERN et al., 2007; FáVERO et al., 2009).

Distância Euclidiana

Distância Euclidiana é a distância linear direta entre dois pontos em um espaço p-dimensional. Sua representação é dada pelas equações a seguir:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{5}$$

ou,

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ik} - x_{jk})^2} \tag{6}$$

Em que x_{ik} é o valor da variável k refrente à observação i e x_{jk} representa a variável k para a observação j. Nesta abordagem, quanto menor a distância, maior a similaridade entre as observações.

Distância de Manhatan

Esta medida representa a soma das diferenças absolutas entre os valores das n variáveis para os dois casos, possuindo sua equação definida por:

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}| \tag{7}$$

Distância de Mikowski

Tanto a distância euclidiana quanto a distância de Manhatan podem ser descritas como um caso particular de uma distância mais geral, denominada distância de Minkowski. Sua forma de medida é dada pela seguinte expressão:

$$d_{ij} = \left(\sum_{k=1}^{n} (|x_{ik} - x_{jk}|)^p \right)^{\frac{1}{p}}, \tag{8}$$

em que d_{ij} é a distância de Mikowski entre as observações i e j, n é o número e variáveis, e $p = 1, 2, \ldots, \infty$.

Distância de Mahalanobis

Representa a distância estatística entre dois indivíduos i e j, considerando a matriz de convariância para o cálculo das distâncias. Sua equação é dada por:

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}, \tag{9}$$

em que S é a estimativa amostral da matriz de variância-covariância Σ dentro dos conglomerados.

Métrica de Canberra

Distância aplicada para variáveis não negativas. É expressada pela equação:

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{x_{ik} - x_{jk}}. \tag{10}$$

3.2.2 Medidas de Correlação

De acordo com Fávero et al. (2009), "as medidas correlacionais representam similaridade pela correspondência de padrões ao longo das características (X variáveis)". O autor aponta que, nas ciências sociais, a correlação e Pearson, dentre as medidas correlacionais é a mais utilzada. Este coeficiente pode ser medido através da fórmula:

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{1k} - \bar{x}_i)(x_{1j} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{1k} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{1j} - \bar{x}_j)^2}}, \tag{11}$$

Tal que, x_{ik} e x_{jk} são os valores da variável k para as observação i e j, respectivamente; $\bar{x}_i$ e $\bar{x}_j$ rperesentam as médias de todas as variáveis para os indivíduos i e j, repsectivamente; e n é o número de variáveis.

O coeficiente de correlação de Pearson mede o grau da correlação linear entre duas variáveis quantitativas, sendo um índice adimensional com valores variando entre -1 e 1 inclusive, que reflete a intensidade de uma relação linear entre dois conjuntos de dados. Nesta abordagem, 1 significa uma correlação positiva perfeita entre os dados; -1 significa uma correlação negativa perfeita entre as duas variáveis, ou seja, se uma aumenta, a outra diminui; 0 significa que as variáveis não possuem dependência linear.

É válido destacar que, as medidas mais utilizadas em análises de conglomerados são as de distância, uma vez que as medidas correlacionais não focam na magnitude dos objetos, mas apenas a correlação entre seus perfis. Além disso, apesar desta medida poder apontar ausência de correlação entre os dados, pode haver uma correlação **não linear** entre eles, sendo mais uma razão para inviabializar seu uso como medida exclusiva para análise de conglomerados.

Tabela 4 - Tabela de Contingência Padrão.

Indivíduo i \ *Indivíduo j*	1	0	Total
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	$p=a+b+c+d$

3.2.3 Medidas de Associação

As medidas de associação são utilizadas para representar a similiradidade quando tratamos de variáveis nominais, baseando-se em tabelas de contingência (FáVERO et al., 2009). Para a construção das tableas de contingência, usa-se o recurso de transformar os atributos em variáveis binárias, assumindo o valor 1 caso possuam aquela característica e 0 caso contrário.

Assim, seja i e j dois indivíduos caracterizados por p variáveis nominais dicotômicas, em que 1 significa a presença da característica, pode-se cofeccionar a seguinte tabela de contingência:

Onde a representa o número de características presentes em ambos os indivíduos, b representa o número de características presentes no indivíduo i e ausente no j, c representa o número de características ausente no i e presente no j e d representa a ausência simultânea de carcterísticas em i e j.

A medida de associação mais utilizada, segundo Sharma (1996), Fávero et al. (2009), são os coeficientes de emparelhamento simples definidos como:

$$S_{ij} = \frac{a+d}{a+b+c+d}, \tag{12}$$

ou

$$d_{ij} = \frac{b+c}{a+b+c} \tag{13}$$

Tal que S_{ij} (medidade de semelhança) é a relação entre o número de características presentes e ausentes simultaneamente para os dois indivíduos e o número total de características; e d_{ij} (medida de distância) representa o coeficiente entre o número de características presentes em um indivíduo e ausentes no outro e o número total de características.

Outros coeficientes de similaridade são apresentados por Johnson, Wichern et al. (2007) e estão dispostas na Tabela 5 em termos das frequências da Tabela 4.

Tabela 5 - Coeficientes de Similiradade.

Coeficente	Análise Racional
$\frac{a+d}{p}$	Pesos iguais para a presença e ausência das características em ambos indivíduos.
$\frac{2(a+d)}{2(a+d)+b+c}$	Peso duplo para a presença e ausência das características em ambos indivíduos.
$\frac{a+d}{a+d+2(b+c)}$	Peso duplo para as presenças particulares de cada indivíduo.
$\frac{a}{p}$	Sem o par $(0,0)$ no numerador.
$\frac{a}{a+b+c}$	Sem o par $(0,0)$ no numerador e no denominador.
$\frac{2a}{2a+b+c}$	Sem o par $(0,0)$ no numerador e no denominador. Peso duplo para o par $(1,1)$.
$\frac{a}{a+2(b+c)}$	Sem o par $(0,0)$ no numerador e denominador. Peso duplo para os pares $(0,1)$ e $(1,0)$.

3.3 Métodos de *Clustering*

Os métodos de *clustering* podem ser divididos essencialmente em dois grupos: **hierárquicos** e **não-hierárquicos**. Além disso, os métodos hierárquicos podem, ainda, ser **aglomerativos** ou **divisivos**.

O primeiro tipo de agrupamento hierárquico começa a partir do melhor particionamento possível, onde cada objeto pertence a um conglomerado exclusivo. Daí, os objetos começam a se reagrupar de acordo com suas similaridades, fundido os grupos a que pertencem. Eventualmente, de acordo com o decréscimo da similaridade, todos os subgrupos são fundidos a um único conglomerado (HäRDLE; SIMAR, 2003; JOHNSON; WICHERN et al., 2007).

Os métodos hieráquicos divisivos funcionam de maneira oposta aos aglomerativos. Ou seja, o processo é iniciado a partir da partição mais grosseira possível, com todos os objetos pertencentes a um mesmo grupo. O algoritmo se dá a partir da divisão sucessiva deste bloco em grupo menores, utilizando como preceito a dissimilaridade dos objetos contidos em cada grupo (HäRDLE; SIMAR, 2003; JOHNSON; WICHERN et al., 2007).

Os métodos não-hierárquicos são constituídos de algoritmos que visam dividir os dados em k partições, onde cada partição representa um conglomerado. Todavia, em oposição aos métodos hierárquicos, o número de grupos k deve ser conhecido *a priori* (SHARMA, 1996).

3.3.1 Métodos Hierárquicos

Este trabalho irá se concentrar nos métodos hierárquicos aglomerativos, visto sua grande aplicação prática e extensa bibliografia. Para maior detalhamento dos procedimentos hierárquicos divisivos Johnson, Wichern et al. (2007) faz uma excelente apresentação do tema, indicando, inclusive, alguns trabalhos que abordam este assunto.

Em síntese, um algoritmo aglomerativo consiste nos seguintes passos (HäRDLE; SIMAR, 2003):

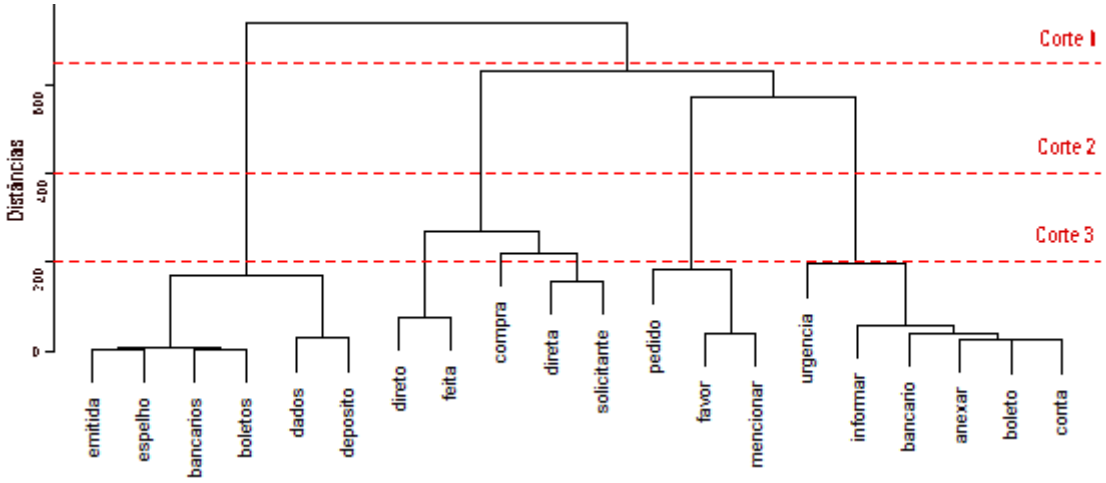
1. Dados N elementos de entrada, definir o melhor particionamento possível, onde cada elemento dá origem a um grupo.
2. Computar as distâncias da matriz $\mathcal{D}_{\mathcal{N},\mathcal{N}}$.
3. Encontrar os dois *clusters* com a menor distância.
4. Agregar os dois *clusters* em um único grupo.
5. Computar as distâncias entre os novos grupos e obter uma matriz de distâncias $\mathcal{D}$ reduzida.
6. Verificar se todos *clusters* estão aglomerados em $\mathcal{X}$, caso não estejam, retornar ao item 3.

Usualmente, o resultado de uma aplicação desta natureza é vizualizado através de um diagrama bidimensional conhecido como **dendograma**. O dendograma ilustra a convergência dos grupos a cada nível percorrido pelo algoritmo, onde cada ramos representa um elemento, enquanto a raiz representa o agrupamento de todos os elementos (DONI, 2004).

Na Figura 7 é apresentado um exemplo de diagrama resultante da aplicação de um algoritmo hierárquico em uma base de dados textuais. Nele, pode-se ver a tendência de alguma palavras pertencerem a um mesmo grupo, ou seja, palavras com maior similaridade. Entretanto, o último ramo do dendograma sempre será composto por objetos únicos representando todos como um grande grupo; devido a isto, cabe ao pesquisador definir uma distância mínima aceitável para que objetos que estejam "até" uma distância q, pertençam a um mesmo grupo. No exemplo apresentado, são apontados alguns valores para q, sendo representados pelas linhas de corte 1, 2 e 3. No corte 1, os dados seriam agrupados em apenas dois conglomerados; no corte 2, os dados seriam agrupados em quatro conglomerados; e no corte 3, os dados seriam agrupados em seis conglomerados.

Existem, no entanto, diversos métodos para computar as distâncias entre estes objetos, sendo a principal diferença entre estes métodos, a função utilizada para o cálculo destas distâncias. Assim, a configuração do dendograma pode, e irá, variar de acordo com

Figura 7 - Exemplo de dendograma resultante de uma aplicação em uma coletânea textual.



o método empregado. Dadas duas classes, G e H, Fávero et al. (2009), Härdle e Simar (2003) destacam alguns métodos habituais de dissemelhança:

- **Menor Distância** (Em inglês, *Single Linkage* ou *Nearest Neighbor*): Consiste em considerar que a distância entre dois subgrupos é a menor distância entre um elemento de um grupo e um elemento do outro subgrupo:

$$D_{GH} = \min_{k \in G,\, l \in H} d_{kl}. \tag{14}$$

- **Maior Distância** (Em inglês, *Complete Linkage* ou *Furthest Neighbor*): Consiste em considerar que a distância entre dois subgrupos é a maior distância entre um elemento de um subgrupo e um elemento do outro subgrupo:

$$D_{GH} = \max_{k \in G,\, l \in H} d_{kl}. \tag{15}$$

- **Ligação Média** (Em inglês, *Average Linkage*): Consiste em considerar que a distância entre duas classes é a média de todas as distâncias entre pares de elementos (um de cada classe):

$$D_{GH} = \frac{1}{n_G n_H} \sum_{k=1}^{n_G} \sum_{l=1}^{n_H} d_{kl}. \tag{16}$$

- **Inércia Mínima (Método de Ward)**: Considera-se a inércia de uma classe G, isto é, a soma de quadrados das diferenças entre cada indivíduo e o "indivíduo médio" dessa classe (dado pelo centro de gravidade da nuvem de n pontos em $\mathbb{R}^p$):

$$I_G = \sum_{l=1}^{p}[\sum_{k\in G}(x_{kl} - \bar{x}_{G_l})^2]. \tag{17}$$

  Tal que, $\bar{x}_{G_l}$ é a média dos valores da variável l para os indivíduos da classe G. Tome-se agora a distância entre as classes G e H como sendo o *aumento na soma total das inércias provocado pela fusão dos grupo G e H*. Neste contexto, seja I_G a inércia da clasee G, I_H a inércia da classe H e $I_{G\bigcup H}$ a inércia da classe resultante de fundir as classes G e H, então (uma vez que a fusão de G e H nãe afeta as inércias das restantes classes):

$$D_{GH} = I_{G\bigcup H} - (I_G + I_H). \tag{18}$$

- **Centróides**: Consiste em utilizar as distâncias entre duas classes como sendo a distância entre os centros de gravidade[8], ou outros pontos considerados representativos (centróides), das classes:

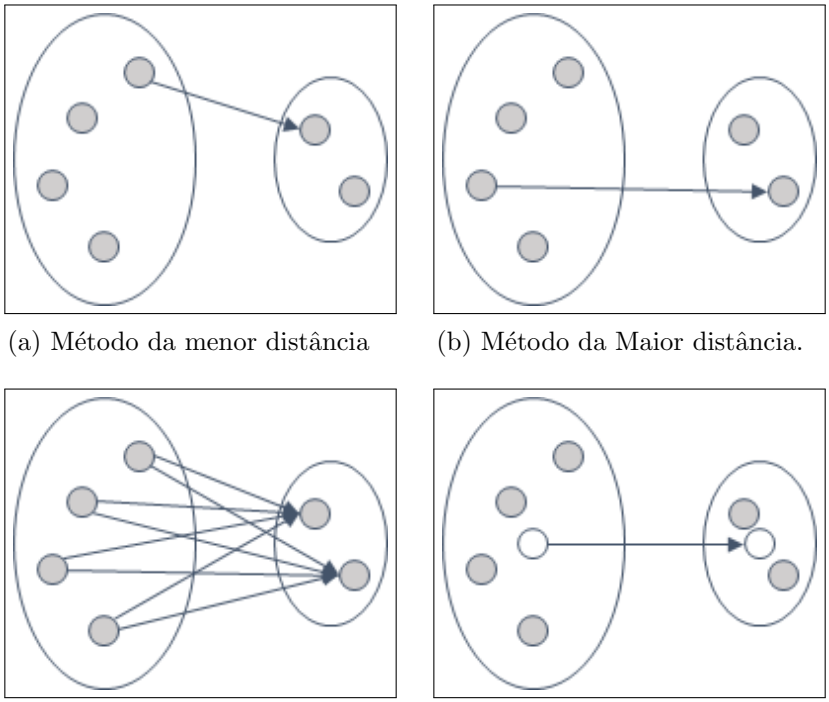
$$D_{GH} = ||\bar{X}_G - \bar{X}_H||. \tag{19}$$

Härdle e Simar (2003) preconiza que as definições de distância supracitadas podem, ainda, serem vistas como casos particulares de uma função distância geral. Nesta abordagem, se dois objetos, ou grupos, P e Q estão unidos, o cálculo da distância (idenpendente do método) entre um novo grupo $P+Q$ e um grupo R pode ser medido através da função:

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|, \tag{20}$$

[8] "O centro de gravidade de um corpor rígido é o ponto tal que, se imaginarmos o corpo suspenso por este ponto e com liberdade para girar em todos os sentidos ao redor deste ponto, o corpo assim sustentado permanecerá em repouso e preservará sua posição original, qualquer que seja a orientação do corpo em relação à Terra" (ASSIS; RAVANELLI, 2008)

Figura 8 - Ilustração das interações realizadas em métodos de *clustering* hierárquico.



(a) Método da menor distância (b) Método da Maior distância.

(c) Método da Ligação Média. (d) Método do Centróide.

onde os δs representam os fatores ponderadores que conduzem aos diferentes métodos aglomerativos já mencionados, como pode ser visto na Tabela 6. Neste ponto, $n_p = \sum_{i=1}^{n} I(x_i \in P)$ é o número de objetos no grupo P. Os valores n_q e n_R são definidos de forma análoga.

Vale ressaltar que a escolha do método de agregação condicionará a classificação resultante. Assim, a opção por um ou outro método deve, pois, ser justificada com base na natureza dos dados e no objetivo da análise.

Um exemplo destas variações pode ser vista em Albuquerque, Ferreira et al. (2006). Neste trabalho, uma equipe de pesquisadores utilizou um conjunto de dados ambientais sobre a Mata de Sivicultura da Universidade de Viçosa (Viçosa, Minas Gerais) para propor uma sistemática de estudo e interpretação da estabilidade dos métodos de *clustering*. O trabalho utilizou a distância de Mahalanobis para a construção das matrizes de distância, utilizando como variáveis, as espécies do local e suas densidades por parcelas de $20 \times 50\ m$. Aplicando-se apenas métodos hierárquicos aglomerativos, tanto para os dados

Tabela 6 - Fatores ponderadores para computação da distância entre grupos.

Método	δ_1	δ_2	δ_3	δ_4
Menor Distância	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Maior Distância	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ligação Média	$\frac{n_p}{n_p+n_Q}$	$\frac{n_Q}{n_p+n_Q}$	0	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_G+n_H+n_K}$	0
Centróide	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0

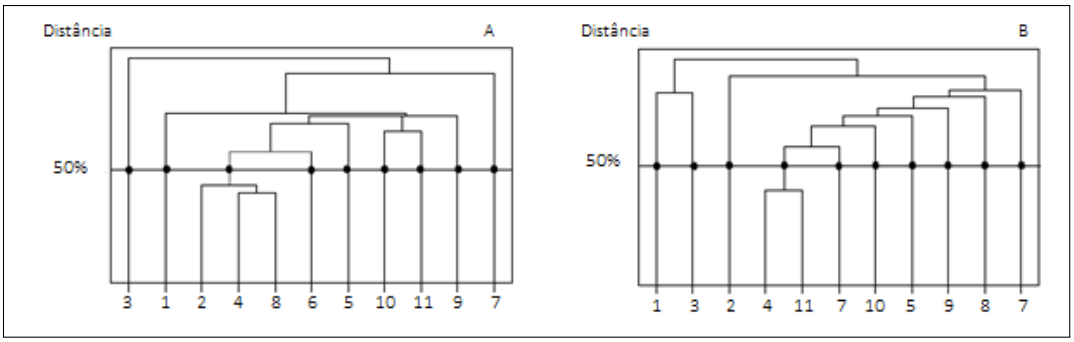
originais quanto para uma amostra proveniente de "*bootstrap*"[9], a Figura 9 representa os dendogramas resultantes de alguns dos métodos aplicados na pesquisa e exemplifica como a conclusão da análise pode variar de acordo com o algoritmo utilizado.

É válido ressaltar que qualquer método produzirá sempre uma classificação, um agrupamento (em qualquer número de classes, dependendo da altura que se opte por cortar o dendograma). Assim, o *clustering* produz "classificações" mesmo onde elas possam não se justificar e, por esse motivo, é importante verificar a robustez dessas classificações. Cadima (2010) expõe que uma boa classificação deverá corresponder a um corte significativo desse dendograma, sendo realizado em uma zona onde as seprações resultantes entre as classes correspondam a grandes distâncias (o que se traduz em barras de junção de classes mais compridas). Este fato irá refletir em maior heterogeneidade entre classes que, como já mencionado, é um dos objetivos dos métodos hierárquicos. Outro objetivo, o da homogeneidade interna das classes "será tanto melhor conseguido quanto mais próximo das folhas (indivíduos) do diagrama se fizer o corte" Cadima (2010).

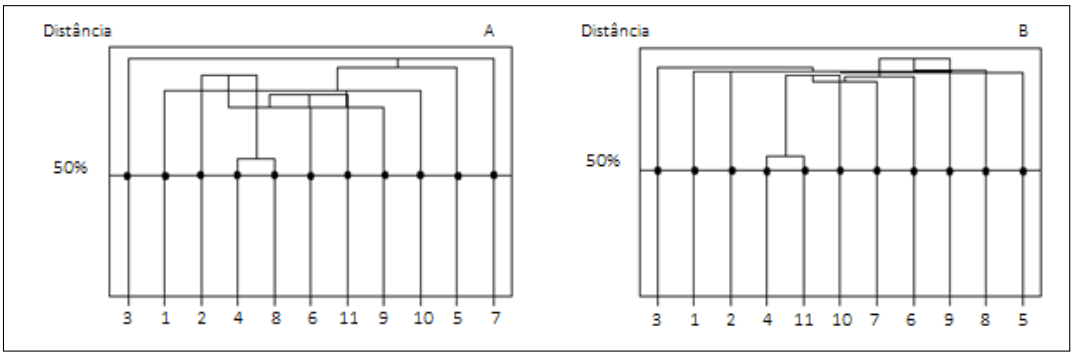
No exemplo ilustrado pela Figura 9, o corte foi escolhido como 50% da amplitude da distância (independente do método) e, com esta definição, note que houve uma variação significativa no número de grupos formados de acordo com o método empregado. Enquanto as Figuras 9a e 9a sugeriram a classificação dos dados em nove grupos, a Figura 9b já aponta para um total de onze grupos. Este fato, corrobora a afirmação anterior e

[9] Este técnica tem como ideia básica reamostrar um conjunto de dados inicial, diretamente ou via um modelo ajustado, a fim de criar réplicas dos dados, a partir das quais pode-se avaliar a variabilidade de quantidades de interesse, sem usar cálculos analíticos. Esta técnica é bastante utilizada quando se possui uma amostra de tamanho reduzido ou não satisfatório.

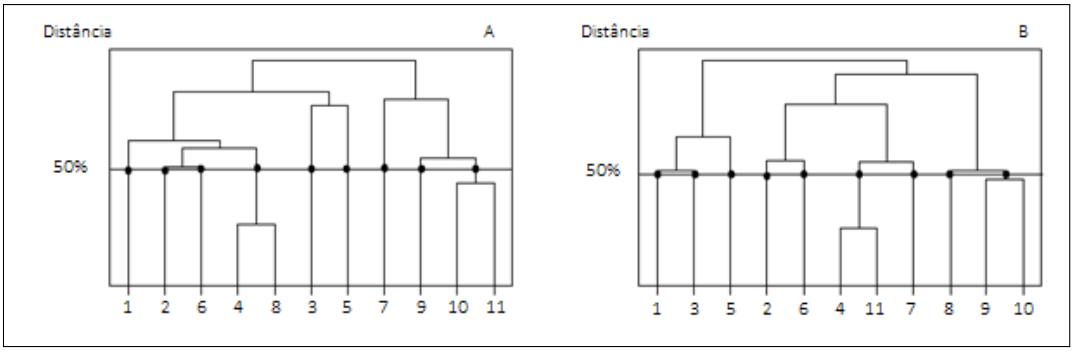
Figura 9 - Dendogramas representando as sequências de fusões das parcelas, com base nas distância de Mahalanobis, a partir dos dados originais (A) e “bootstrap” (B).



(a) Método da menor distância.



(b) Método do centróide.



(c) Método de Ward.

Fonte: Estabilidade em Análise de Cluster: Estudo de Caso em Ciências Florestais (ALBUQUERQUE; FERREIRA et al., 2006)

reforça a importância do emprego de uma metodologia para o corte do dendograma, de forma que o resultado obtido seja o mais eficiente possível.

Em consequência dos cálculos diferenciados para a obtenção das distâncias, cada método possui características próprias que podem favorecer, ou desfavorecer, seu uso em uma dada pesquisa. Dentre estas características, destacam-se (CADIMA, 2010; SHARMA, 1996; FáVERO et al., 2009):

- O método da Menor Distância tende a produzir classes mais alongadas, com indivíduos que podem estar muito distantes entre si, mas pertencendo a uma mesma classe (evento conhecido como "encadeamento"). Devido a isto, este método possui pouca tolerância a ruídos, visto que, basta que exista um elemento de uma classe "próximo" a um elemento de outra classe para que estas sejam atraídas.

- O método da Maior Distância tende a formar grupos compactos e os ruídos demoram a serem incorporados aos grupos;

- O método da Ligação Média possui menor sensibilidade à ruídos que os métodos de Menor e Maior Distância e, tende a formar grupos com número de elementos parecidos;

- O método do Centróide apresenta forte robustez à presença de ruídos, entretanto, eventualmente pode gerar dendogramas confusos caso sejam geradas distâncias entre centróides menores do que as distâncias entre grupos já formados. Este fenômeno recebe o nome de reversão e, por isso, pode inviabilizar a utilização deste método para determinados conjuntos de dados. Um exemplo de dendograma com a presença de reversão pode ser visto na Figura (referenciar acima), onde uma das linhas é gerada em um nível superior a um grupo que já havia sido desenhado.

- O método de Ward tende a produzir grupos com um número aproximadamente similar de indivíduos e é sensível à presença de *outliers*.

3.3.2 Métodos não-Hierárquicos

Os procedimentos não hierárquicos são utilizados para agrupar indivíduos (e não variáveis) cujo número inicial de *clusters* é definido *a priori*. Este número pode ser definido antecipadamente ou determinado como parte do procedimento de *clustering* (JOHNSON; WICHERN et al., 2007). Não obstante, as técnicas não hierárquicas são métodos que objetivam "encontrar diretamente uma partição de n elementos em K grupos (*clusters*)" (FáVERO et al., 2009), de modo que as partições otimizem a homogeneidade dentro de cada grupo e a heterogeneidade entres os grupos formados (BELBIN, 1987).

Os métodos não hierárquicos não exigem a construção de matrizes de distância (ou dissimilaridade), não sendo necessário o cálculo e armazenamento de matrizes a cada etapa do processo. Como resultado, procedimentos não-hierárquicos tendem a possuir maior eficiência computacional. Este fator faz com que algoritmos desta natureza tenham alta aplicabilidade quando se analisa grandes conjuntos de dados (JOHNSON; WICHERN et al., 2007).

A escolha do valor de K ou dos elementos que deverão compor estes K grupos iniciais, apesar de crucial, não possui uma definição única. Sharma (1996) propõe a utilização de observações aleatórias como centro dos *clusters*; (FáVERO et al., 2009) sugere a aplicação de um *clustering* hierárquico para, posteriormente, definir um número embasado nos resultados obtidos; Belbin (1987) sugere a extração de sementes em intervalos regulares, de modo a reduzir qualquer viés de amostragem original; e Ball e Hall (1967) utilizaram a determinação de um grande centróide, definindo um valor limiar, tal que, os elementos que estavam a uma distância maior do que este limite eram selecionados para criar os grupos. Vide então, que a definição de K pode, e irá, variar de acordo com o estudo e a bibliografia seguida pelo pesquisador.

Em linhas gerais, as técnicas empregadas em métodos não-hierárquicos seguem a seguinte lógica (HUANG, 1997):

1. Selecionar k centróides ou sementes para os *clusters* iniciais, onde k é o número de grupos desejado;

2. Atribuir cada observação ao *cluster* mais próximo;

3. Reatribuir ou realocar cada observação cada observação a um dos k *clusters* de acordo com uma regra de parada pré determinada;

4. Parar se não houverem mais realocações de dados ou se as reatribuições satisfazem um conjunto de critérios determinados pela regra de parada empregada. Caso não atenda este item, retornar ao passo 2.

Em geral, Sharma (1996) relata ainda que outro ponto de diferenciação entre algoritmos desta natureza (além da regra definida para definir os grupos iniciais) está nas regras de realocação das observações. Dentre estes métodos, o mais usual é o *k-means*. Este método será descrito detalhadamente a seguir e, posteriormente, será apresentado o método *k-medoid*. Ambos possuem uma lógica muito similar, variando entre si apenas no critério adotado para o cálculo do valor de referência a ser atribuído durante o *clustering*.

3.3.2.1 Método *K-means*

MacQueen et al. (1967 apud JOHNSON; WICHERN et al., 2007) sugere o termo "*k-means* (em português, k-médias) para descrever um algoritmo que aloca cada item a um *cluster* através do centróide (média) mais próximo". Na sua versão mais simplificada, o processo é composto por quatro etapas (JOHNSON; WICHERN et al., 2007; DONI, 2004):

1. Particionar o os items dentro de k grupos iniciais;
2. Prosseguir com a lista de itens, alocando-os ao grupo cujo centróide (média) está mais próximo (A distância computada, geralmente, utiliza a distância Euclidiana);
3. Recacular os centróides para ambos os grupos que sofreram alteração quanto ao volume de itens; e
4. Retornar ao passo 2 até que não haja mais realocações, ou seja, seja alcançada um particionamento ótimo para o grupo inicial escolhido.

Johnson, Wichern et al. (2007) pontua ainda que, ao invés de iniciar o algoritmo através da alocação de todos os itens nos K grupos no passo 1, pode-se especificar K centróides iniciais e então prosseguir para o passo 2. Contudo, esta seleção arbitrária inicial pode ser melhorada. Para isto, calcula-se o grau de homogeneidade interna dos grupos através da *Soma de Quadrados Residual* (SQRes), que é a medida usada para avaliar o quão boa é uma partição (PRASS, 2004). A SQRes é medida através da função:

$$SQRes(j) = \sum_{i=1}^{n_j} d^2(o_{ij}, \bar{o}_j), \tag{21}$$

tal que o_{ij} representa o valor do i-ésimo registro; $\bar{o}_j$ o centróide do grupo j; e n_j o número de registros no grupo j.

Após a obtenção do SQRes, move-se o primeiro objeto para os demais grupos com o intuito de verificar se ocorrem diminuição em seu valor. Caso haja, este objeto é movido para o grupo que produzir maior ganho, a SQRes é recalculada e passa-se ao objeto seguinte. Eventualmente, as mudanças são cessadas e o processo interrompido (PRASS, 2004).

A Tabela 7 exemplifica a execução do algoritmo utilizando o conjunto $\{2, 5, 7, 4, 8, 9, 1\}$ sob a definição de $K = 2$. Observa-se que como sementes foram escolhidos os dois primeiros elementos do conjunto e, como critério para definir o valor do centróide foi utilizado o valor médio do *cluster*. C_1 e C_2 representam os centróides obtidos em cada etapa do processo.

Algumas características desse método são:

Tabela 7 - Exemplo de execução do algoritmo *k-means*.

Iteração	Conjunto não agrupado	Grupos	Centróides
0	$(7, 4, 8, 9, 1)$	(2) (5)	$C_1 = 2.0$ $C_2 = 5.0$
1	$(4, 8, 9, 1)$	(2) $(5, 7)$	$C_1 = 2.0$ $C_2 = 6.5$
2	$(8, 9, 1)$	$(2, 4)$ $(5, 7)$	$C_1 = 3.0$ $C_2 = 6.5$
3	$(9, 1)$	$(2, 4)$ $(5, 7, 8)$	$C_1 = 3.0$ $C_2 = 6.3$
4	(1)	$(2, 4)$ $(5, 7, 8, 9)$	$C_1 = 3.0$ $C_2 = 7.3$
5	$\emptyset$	$(2, 4, 1)$ $(5, 7, 8, 9)$	$C_1 = 2.3$ $C_2 = 7.3$

- Sensibilidade a ruídos, uma vez que um objeto com valores extremamente altos podem distorcer substancialmente a distribuição dos dados (HAN; KAMBER, 2006); e
- Aplicado apenas a dados binários e numéricos (PRASS, 2004).

3.3.2.2 Método *k-medoid*

Como já mencionado anteriormente, algoritmos *k-medoid* são semelhantes ao *k-means*, variando apenas no que tange o valor de referência utilizado para o agrupamento dos itens. Ao invés de utilizar o valor médio dos objetos para o cálculo dos centróides, este método sugere a escolha de objetos reais para reprensentar os grupos, usando um objeto representativo por grupo (HAN; KAMBER, 2006).

A dinâmica do processo se dá através troca, iterativa, dos objetos elencados para representar os grupos por um não elencado, de modo que a qualidade dos agrupamentos resultantes sejam continuamente melhorada até um ponto em que se estabiliza. Esta lógica de funcionamento se baseia no princípio de minimizar a soma das dissimilaridades entre cada objeto e seu ponto de referência correspondente (objeto selecionado para representar um dado *cluster*). Para isso, é utilizado um *critério de erro absoluto* definido por (HAN; KAMBER, 2006):

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|, \tag{22}$$

onde E é a soma do erro absoluto para todos os objetos no conjunto de dados; p é o ponto no espaço que representa um dado objeto no grupo C_j; e o_j é o objeto representativo de C_j.

Alguma características desse método são:

- Os resultados serão os mesmos independente da ordem inicial;
- Devido ao cálculo recorrente de E, o processamento é mais custoso, tornando-o inviável quando se quer analisar grandes conjuntos de dados; e
- Apresenta maior robustez a presença de ruídos do que o método *k-means*, já que é menos influenciado pelos ruídos do que a média.

4 APLICAÇÃO PRÁTICA - O QUE FALAM SOBRE ECONOMIA?

4.1 Descrição do Cenário

Diante da situação atual do país, existem alguns temas que andam "esquentando" os meios de comunicação. Desde os jornais impressos até as grandes redes sociais, como Tweeter e Facebook, temas relacionados a economia e política do país estão a todo vapor. Assim, pegar uma carona nestas temáticas e desenvolver estudos e aplicações que envolvam estes assuntos é algo bastante interessante.

Neste contexto, este trabalho focou-se no estudo de apenas um destes meios de comunicação, de forma que, a partir dele, fossem retirados *insights* sobre a situação da economia do país e os principais assuntos abordados até o início de 2016. Para isso, foi levada em consideração a frequêcia de publicação de conteúdo, influência de seus escritores sob a população, frequência de acessos e a existência de canais de discussão acerca da matéria publicada.

Embasado nestas métricas, foi escolhido o blog da jornalista Míriam Leitão, alocado na seção de economia do jornal O Globo (edição digital), como principal motor da pesquisa. Seu blog tem análises exclusivas sobre a economia nacional e estrangeira feitas pela própria jornalista e pelos autores Álvaro Gribel e Marcelo Loureiro. Além disto, são postados produtos divulgados em vários veículos do Grupo Globo pela jornalista, os comentários na TV e Rádio, e a coluna no GLOBO.

Após a escolha do portal, analisou-se a estrutura das páginas que compunham sua coluna a fim de construir um web crawler específico para a busca das matérias publicadas no blog. Posteriormente, os dados foram analisados de forma descritiva, buscando empregar meios de visualização dos dados pertinentes aos tipos de dados que se está analisando (textos) e finalmente, empregar tecnicas de *clustering* para obter informações úteis e interessantes em cima da coletânea obtida.

Alguns resultados esperados são: segregação de matérias por autor de forma automática; segregação de matérias por conteúdo; e segregação de assuntos por período de publicação.

Todas os algoritmos para recuperaçãos de conteúdos e estruturação dos dados foram implementadas em Python, enquanto que as análises posteriores foram realizadas através do software R. A opção por estas linguagens se deu por sua robustez em realizar tais tarefas e por serem sistemas *free* e *open-source*.

4.2 Coleta e Estruturação da Coletânea

Como já mencionado neste trabalho, a obtenção de conteúdos via *web crawler* não possui uma maneira única, visto que cada algoritmo poderá variar de acordo com o dado que se deseja obter e o caminho que se irá percorrer para obte-lo. Sendo assim, a maneira escolhida para recuperar informações referentes às postagens do blog em apreço, foi através do estudo de duas estruturas html básicas que compõem o portal e são ligadas através de hiperlinks.

A primeira estrutura (E1) possui uma listagem de no máximo quinze postagens,onde cada item possui uma breve introdução da matéria, seu título e autor. Esta listagem está ordenada segundo a ordem de publicação de cada matéria, apontando (através de hiperlinks) para a segunda estrutura de páginas. Além disto, um fator importante destas páginas é que, apesar de apresentar apenas quinze postagens, ela também possuem hiperlinks para outras páginas semelhantes que contêm postagens anteriores. Sendo assim, analisar as conexões existentes nesta estrutura foi o que propciou que fossem obtidos todos os hiperlinks necessários para a recuperação das matérias, visto que a segunda estrutura (E2) de páginas corresponde a apresentação de cada matéria na íntegra.

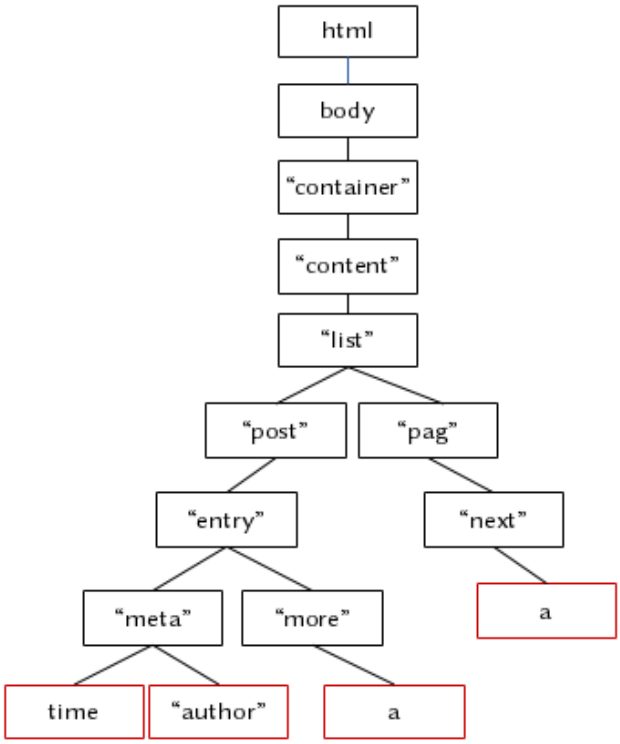
Além de apresentar o conteúdo expandido, a segunda estrutura analisada possuia uma informação bastante relevante, que eram os comentários dos leitores acerca daquela publicação. Entretanto, como o enfoque do trabalho foi o estudo acerca dos temas e assuntos abordados pelos autores, não foi implementado rastreadores para a obtenção deste tipo de conteúdo.

Sabendo-se que páginas HTML apresentam estruturas similares a árvores, analisou-se os "caminhos" para se buscar as informações sobre autor, data, título e conteúdo expandido de cada publicação. Esta análise contribuiu para que fossem empregadas técnicas de *parsing* a fim de extrair exatamente os dados que eram desejados.

A Figura 10 apresenta as principais tags identificadas e utilizadas na implementação do algoritmo para obtenção das informações das páginas E1. Nesta, foram suprimidos os outros 15 (quinze) nós do tipo "post" que compunham esta estrutura. Cada um desses nós corresponde a uma matéria listada na página, onde: "time" era responsável pela alocação da data e hora da publicação; "author", pelo nome do autor da matéria; "more > a", contem o hiperlink para a página da matéria na íntegra; e "next > a", possui o hiperlink para a próxima página do tipo lista contendo matérias anteriores.

As páginas E2, não exigiram grandes análises, visto que, os dados referentes ao título, data e autor, já poderiam ser obtidos na primeira página, restando apenas a recuperação do conteúdo referente ao texto da matéria. Esta informação estava alocada em uma trilha do tipo "div > entry > p", onde cada "p" representa um parágrafo do texto. Os dois primeiros continham informações sobre autor e data de publicação, respectivamente, e os dois últimos, informações sobre compartilhamentos e comentários. Contudo,

Figura 10 - Trilhas de tags principais utilizadas para extraçao de dados (E1).



como estas informações ou já foram obtidas ou não serão utilizadas, decidiu-se expurgá-las no momento da obtenção dos dados, recuperando apenas os textos escritos em cada publicação.

```
def BuscaPosts(linkStart, saida, total, Start = True):

    soup = BeautifulSoup(linkStart.content)

    for i in range(len(soup.find_all('article'))):
        link = soup('article')[i]('h2')[0]('a')[0].get('href')
        autor = soup('article')[i]('span')[0].string
        data = soup('article')[i]('time')[0].string[:10]
        titulo = soup('article')[i]('h2')[0].string
        saida.append((link, autor, data, titulo))

    if len(saida) < total:
        linkNext = soup('div', attrs={'class':'pag'})[0]('p',
           attrs={'class':'next'})[0]('a')[0].get('href')
        linkNext = requests.get(linkNext)
        BuscaPosts(linkNext, saida, total, Start = False)

return saida
```

Conhecendo estas estruturas, foi implementada a função *BuscaPosts*, reponsável pela busca e armazenamento do conteúdo pertinente às páginas E1. Esta função, recebe como entrada uma URL [10] inicial - *linkstart*; uma lista para alocação do conteúdo recuperado - *saida*; o total de matérias que se quer buscar - *total*; e uma informação, do tipo lógica, se a URL utilizada é a inicial - *Start*.

Esta função é o motor principal do algoritmo, uma vez que, após sua implementação, o restante do código teve a finalidade apenas de obter os textos expandidos alocados nos hiperlinks apontados pelas péginas E1 e exportá-los de forma estruturada em formato *.csv* para que fosse posteriormente analisado. O código completo pode ser visto em Anexo B, onde pode ser visto mais detalhadamente a lógica empregada pelo algoritmo.

[10] URL é o endereço de um recurso disponível em uma rede, seja na internet ou intranet, e significa em inglês *Uniform Resource Locator*.

4.3 Pré-processamento da Coletânea

A coletânea obtida possuia 510 matérias publicadas no período de 25/08/2015 até 12/12/2015, apresentando 15170 termos diferenciados. Destas matérias, 41 (8.04%) foram escritas pelo autor Álvaro Gribel, 218 (42.75%) por Marcelo Loureiro e 251 (49.22%) pela jornalista Míriam Leitão. Os autores utilizaram 2489, 3693 e 12998 termos diferenciados, respectivamente.

Após esta análise, foram aplicadas técnicas de:

1. Transformação das letras dos termos para minúsculo, visto que o sentido do termo não se altera;
2. Remoção da pontuação;
3. Remoção de *stopwords* segundo Anexo A;
4. Remoção de números; e
5. Remoção de espaços em branco ocasionados por erros de digitação.

Estas manipulações foram realizadas sequencialmente, entretanto, após os primeiros resultados, foi necessário processar alguns ajustes manualmente a fim de que fosse obtido uma coletânea mais expressiva dos termos empregados pelos autores. Dentre estes ajustes, o mais moroso foi a análise das *stopwords*.

Apesar de ser disponibilizada uma *stoplist* preliminar para a língua portuguesa, nem todos os termos considerados *stopwords* para a realização deste trabalho estavam ali inseridos. Termos como "ser, sobre e $", que não contribuem para as análises seguintes, tiveram que ser analisados caso a caso e processados separadamente.

O resultado do processamento da coletânea foi uma redução de 37.31% dos termos iniciais, resultando em 9510 termos. Em relação aos autores, os termos empregados por Álvaro Gribel, Marcelo Loureiro e Miriam Leitão, foram reduzidos para 1176 (47.44%), 2311 (62.58%) e 8592 (66.10%) termos, respectivamente.

4.4 Análise e Interpretação de Conteúdo

Com foco na análise da coletânea de matérias obtidas, não fazendo distinção incial quanto o autor que a escreveu, as análises posteriores se ateram ao estudo do conjunto completo de dados, não fazendo distinção entre seus autores.

Dentre os 9510 termos resultantes do pré-processamento, podemos ver na Tabela 8 os 25 termos mais frequentes. Intuitivamente, vemos alguns termos que condizem com

Tabela 8 - Tabela de Frequência dos Top 25 Termos.

Termo	Frequência
Governo	800
ano	546
inflação	326
país	317
presidente	294
economia	285
brasil	276
banco	221
alta	214
agora	210
queda	206
pib	204
anos	203
fiscal	201
dilma	185
mercado	175
déficit	172
contas	164
crise	162
dólar	159
dois	156
meta	155
congresso	153
maior	150
ministro	147
meses	147

Figura 11 - *Wordcloud* dos Top 100 Termos.

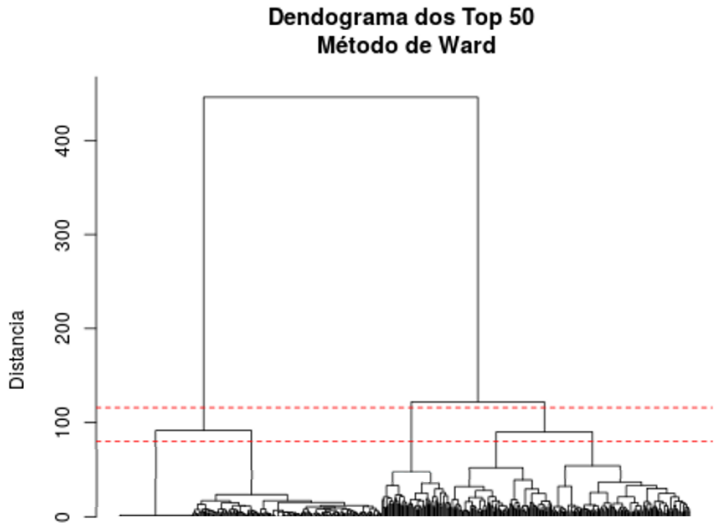
a situação pela qual o Brasil vem passando, corroborando o uso desta técnica como fonte de *insights* sobre assuntos, períodos ou personalidades com maior evidência.

Entretanto, quando trabalhamos dados desta natureza, esta forma de visualização não é muito eficiente. A fim de lidar com este problema, uma solução é o emprego de diagramas do tipo *wordcloud*. Estes diagramas apresentam os termos em um espaço bidimensional, variando o tamanho e a cor de cada termo de acordo com a frequência com que surgem na coletânea. Sua forma de visualização é extramente intuitiva e limpa, otimizando o entendimento dos dados investigados e propiciando uma análise com um número maior de termos. Para esta coletânea, foi construído um diagrama utilizando os 100 termos mais frequentes e o resultado pode ser visto na Figura 11.

Após compreender panoramicamente a coletânea de matérias, foram aplicadas técnicas de *clustering* para analisar possíveis padrões implícitos nestes dados. Visto a magnitude dos dados que se estava estudando, optou-se pelo emprego do método não-hierárquico *k-means*, entretanto, este método apresenta um complicador, que decisão prévia acerca do número de agrupamentos que obterá ao final da tarefa. Diante disto, foi adotado como parte da metodologia, o emprego *a priori* de um método hierárquico, tal que, a partir dos resultados retornados desta tarefa, se tivesse um *insight* sobre alguns números razoáveis de k.

Métodos hierárquicos, no entanto, não são eficientes para grandes números de dados, razão qual não o empregamos como único método da tarefa de *clustering*. Assim, optou-se pela utilização dos 50 termos mais frequentes, visto que o único fundamento desta tarefa inicial é dar subsídios para o processamento via *k-means*. Esta tarefa utilizou

Figura 12 - Dendograma dos Top 50 Termos, via método de Ward.



o método de Ward e medida de distância euclidiana simples, seu resultado pode ser visto através do dendograma apresentado na Figura 12.

Seja $A = 445.97$ a amplitude das distâncias do dendograma apresentado na Figura 12, foram definidas como linhas de corte $d_1 = A \times 0.20 = 89.19$ e $d_2 = A \times 0.25 = 111.50$. Estes valores foram definidos após uma análise visual do dendograma e sugerem $k_1 = 5$ e $k_2 = 3$, como números razoáveis de agrupamentos. Além destes grupos, decidiu-se pelo estudo para $k_3 = 6$, apenas para verificar se há melhora significativa em relação a k_1.

Após a definição dos agrupamentos que seriam analisados, empregou-se a tarefa de mineração de agrupamento não-hierárquico, *k-measn*. O retorno desta tarefa é a alocação de cada matéria em um determinado *cluster*, variando de acordo com o número de grupos definidos inicialmente. Por isto, foram definidos alguns critérios mínimos para serem avaliados em cada aplicação:

- Relação entre grupos resultantes e período de publicação;
- Relação entre autores e grupos resultantes; e
- Similaridade de assuntos principais de cada matéria em cada grupo.

4.4.1 *k-means* com $k_3 = 6$

Ao utilizar $k = 8$, observou-se que 2 grupos, (a) e (b) de matérias foram publicadas apenas em novembro e dezembro, todas pela jornalista Míriam Leitão. Em relação a seu conteúdo, um grupo (a) continha única e exclusivamente matéria sobre o banqueiro André

Esteves e seu vínculo com o banco BTG Pactual, em especial, sua trajetória profissional como banqueiro, envolvimentos com a política, acusações sobre corrupção e, findando, com as consequências de seu desligamento junto ao banco. Um ponto interessante é que estas notícias (7) começaram no dia 25/11/15 e cessaram em 06/12/15, mostrando o quão pontual, porém significativo, foi este tema.

O *cluster* (b) agregou unicamente informações o rompimento da barragem de Mariana - MG e suas consequências ambientais e economicas. Matérias sobre o envolvimento das empresas Samarco e Vale no ocorrido, impactos diretos na cidade de Colatina - ES e custo repentino com o FGTS da população afetada foram os assuntos principais destas matérias. Suas publicações (11) foram realizadas apenas dentro do mês de novembro, aparentando uma possível aquietação em relação ao tema.

O grupo (c) foi composto por 57 matérias, quase exclusivas da autora Míriam Leitão (54), publicadas ao longo de todos os meses analisados. Os assuntos abordados neste grupo são relacionados a temas fiscais, envolvendo desde decisões em votação - como a volta da CPMF[11] e as famosas "Pedaladas Fiscais", até o encontro da atual Presidente com líderes de outros países para discutir assuntos relacionados a este tema. Notoriamente, este tema é bastante abrangente, possuindo diversas matérias, todavia, os resultados obtidos foram bastante precisos quanto ao tipo de Tema abordado em cada matéria.

O grupo (d) agregou 352 publicações, sendo Marcelo Loureiro seu principal autor, com 217 publicações. Outro fato relevante é que 27 matérias foram escritas por Álvaro Gribel, algo interessante, visto que o autor possui apenas 41 matérias publicadas dentro da coletânea recuperada. O grande tema abordado neste grupo são ações na bolsa de valores e, portanto, acaba envolvendo alguns assuntos mencionados em outros grupos, como a queda das ações do banco BTG Pactual devido a tragédia mencionada em (a). Alguns nomes chaves deste grupo são: BTG, inflação, bolsa, IPCA e Levy.

O grupo (e) consolidou 47 matérias, sendo em sua maioria escritas pela jornalista Míriam Leitão (35) e o restante pelo autor Álvaro Gribel. O temas abordados neste grupo foram sobre a temática "inflação", englobando publicações sobre a recessão do país, as variações crescentes da inflação, a perda do poder de investimento do Brasil e os impactos da (e na) alta do dólar. Nomes como Banco Central (BC), Levy e inflação foram mais significativos dentre as postagens.

O último *cluster* (f) foi composto por 37 matérias, todas escritas pela jornalista Míriam Leitão. O conteúdo das matérias que constituem este agrupamento é mais abrangente do que as anteriores, tratando de assuntos que, em sua maioria, tratam de temas

[11] Contribuição Provisória sobre Movimentação Financeira (CPMF) é um tributo brasileiro aplicado em esfera nacional entre os anos de 1997 e 2007.

mais polêmicos, envolvendo a insatisfação com medidas políticas adotadas e discussões sobre a integridade e competência de alguns atores. Insatisfação e pedido de impeachment da atual Presidente, publicações sobre a integridade do Presidente da Câmara Eduardo Cunha e discussões sobre a competência do ministro da Fazenda, Levy e sua possível substituição, são temas principais de algumas matérias que estão aqui inseridas.

É válido dizermos que este número de agrupamentos gerou um resultado bastante satisfatório. A aglutinação das matérias foi realizada através de assuntos concisos e que refletia a situação do país na época; possuíam temas muito similares em cada *cluster*, havendo poucas matérias que distoavam em cada grupo; e possibilitou uma interpretação consistente quanto aos principais temas

4.4.2 *k-means* com $k_2 = 5$

Para um total de 5 grupos, abtivemos um resultado similar ao visto em k_3, porém, alguns grupos ficaram mais mal estrturado, sendo compostos por assuntos tão diversos que era difícil alinhar em temas muito específicos. Entretanto, foram vistos alguns resultados interessantes, com alguns agrupamentos que obtiveram como tema principal outros assuntos que foram tratados implicitamente na tarefa anterior.

O grupo (a) agregou a maior parte das matérias, com 82 publicações da jornalista Míriam Leitão, 216 do autor Marcelo Loureiro e 16 de Álavro Gribel, somatizando 314 matérias. Este primeiro grupo foi composto por matérias cuja temática principal foram "inflação" e ou "ações da bolsa de valores". Entretanto, possivelmente devido ao número reduzido de agrupamentos, algumas matérias atreladas ao banco BTG e a tragédia de Mariana - MG acabaram sendo incluídos aqui, visto que o banco BTG é, também, bastante comentado em matérias relacionados a inflação e deflação do país. Grandes atores aqui são: BTG, IPCA, inflação e bolsa.

O *cluster* (b) é composto por 6 matérias publicadas dentro do mês de novembro, todas por Míriam Leitão. As matérias alocadas aqui são única e exclusivamente sobre temas relacionados a tragédia de Mariana - MG, mencionando frequentemente nomes como Vale e Samarco.

O grupo (c) aglutinou 79 matérias, havendo apenas 2 matérias não escritas por Míriam Leitão. Dentre os assuntos abordados, notamos que todas possuiam um cunho de discussão acerca da indoneidade, integridade e ou competência de alguma figura pública. Debates sobre o impeachment da atual presidente DIlma, incompetência do presidente da câmara Eduardo Cunha, a prisão e ou envolvimento de personalidades públicas, como Delcídio e André Esteves, estão dentre os principais assuntos abordados em alguma matérias do grupo.

O grupo (d) apresentou um resultado bastante interessante, visto que, das 57

matérias agrupadas, 48 participam do grupo (c) para k_3. Desta forma, podemos intuir sobre as principais características deste grupo analisando seu "par" para k_3. Esta forma de interpretar os dados também foi empregada para o grupo (e) que, das 44 matérias aglutinadas, 43 estavam alocadas no grupo (e) utilizando k_3.

4.4.3 *k-means* com $k_1 = 3$

O grupo (a) resultante da aplicação de k_1 é composto por 47 matérias, estando todas contidas no grupo (e) de k_3, valendo portanto, a mesma interpretação supracitada.

O grupo (b), apresentou um total de 95 publicações, estando estas divididas quase igualmente entre os grupos (c) e (d) encontrados para k_2 - 45 e 53 matérias, respectivamente. Diante disto, observou-se matérias mais genéricas que enblobam assuntos como as diversas dicussões políticas ocorridas na câmara dos deputados, debates acerca de integridade de algumas figuras políticas e temas fiscais, como o retorno da CPMF.

O grupo (c) foi construído através da aglutinação de 355 matérias, estando 324 delas atreladas ao grupo (a) de k_1. Consequentemente, podemos intuir os mesmo comentários já citados para as matérias que estão contidas neste *cluster*.

4.5 Conclusão

A proposta inicial desta aplicação era analisar o cenário economico do país via análise de matérias publicadas em um portal de economia específico. Para isso, foi utilizada a metodologia de mineração de textos, desenvolvendo um rastreador web específico para recuperação e estruturação do banco em um formato facilmente manipulável. Esta etapa inicial foi bastante onerosa, visto que exigiu uma análise criteriosa das estruturas de páginas HTML construídas pelos desenvolvedores do portal, porém, tornou viável a análise de um número muito maior de matérias, além de possibilitar a replicação do cenário.

A utilização de dois softwares distintos, aparentemente fugindo as boas práticas de programação, fez com que fosse utilizado o que há de melhor em cada linguagem, aliando pacotes robustos de *crawling* para Python e bibliotecas bastante eficientes para mineração de textos e análise de bancos de dados para o R.

Podemos ressaltar, tambem, que adotar o uso de técnicas de agrupamentos hierárquicos e não-hierárquicos, teve um grande efeito na análise. Caso tivessemos escolhido o número de grupos ao acaso para a aplicação *k-means*, possilvelmente chegaríamos aos números de k adotados, entretanto, se quisermos replicar esta análise para períodos distintos ou para um autor específico, teríamos uma dificuldade maior. Observamos que, para a coletânea

obtida, o número $k_3 = 6$ retornou resultados muito interessantes, e suficientes, para a compreensão dos assuntos abordados pelos colunistas. Ressaltamos ainda que, a escolha pelo blog da jornalista Míriam Leitão foi parte fundamental para os resultados, visto que conseguimos evidenciar diversos temas que impactaram a economia, analisando inclusive, o período que começaram a ser mais discutidos.

É evidente que estas não são as únicas análises possíveis para estes dados, se acrescentarmos ao banco de dados os comentários dos leitores e os números de compartilhamentos de cada matéria, poderíamos estudar a posição da população sobre determinados assuntos ou se um tema gera mais polêmica do que outro. O estudo das fontes utilizadas para estruturação da página é outro fator que pode ser incorporado em trabalhos futuros. O termo em negrito ou com colaração diferente do padrão, pode dar uma ideia da importância daquela palavra para o contexto da matéria. Assim, concluímos que a aplicação da metodologia frente ao cenário estudado foi muito satisfatória, entretanto, ainda podemos acrescentar diversos pontos que tornariam a interpretação dos dados ainda mais poderosa.

CONSIDERAÇÕES FINAIS

O objetivo principal deste trabalho foi a apresentação de uma metodologia que englobasse técnicas de mineração de textos e *web crawler*, aplicando-a para interpretação de um cenário real. Os dados da aplicação foram recuperados através da seção de economia do jornal O Globo (edição digital) e foi escolhido devido sua frequência de publicação, número de acessos e impacto sobre a população. Apesar de, inicialmente "comum", artigos que envolvam estas temáticas ainda são muito focados na análise de redes sociais e, portanto, optou-se pelo desenvolvimento de uma pesquisa sob uma ótica diferenciada.

O desenvolvimento de rastreadores web para a coleta e estruturação dos dados se mostrou muito eficiente, possibilitando a busca por um número maior de conteúdos e propiciando a replicação da pesquisa, fato que não seria viável caso tivessemos que buscar manualmente, ou dependecemos do próprio portal, para coletar as matérias.

Como tarefa de mineração, optou-se pela técnica de agrupamento não-hierárquico *k-means*, visto sua maior efetividade para grandes conjuntos de dados. Todavia, para a otimização desta escolha (tempo de processamento e análise) empregamos o método hierárquico de Ward com os termos mais frequentes. Esta ação se mostrou muito eficaz, sugerindo que entre 5 e 8 grupos poderiam ser uma boa estimativa para a aplicaç]ao do *k-means*.

Em relação aos resultados obtidos, observamos que a escolha por $k = 6$ retornou uma estratificação bastante satisfatória, organizando as matérias em grupos muito específicos quanto ao assunto principal abordado em cada publicação. Podemos ressaltar dentre estes assuntos: a tragédia ambiental ocorrida em Mariana - MG e seus impactos ambientais; discussões acerca da integridade e ou competência de personalidades públicas; a saída do banqueiro André Esteves da liderança do banco BTG Pactual; e os impactos e influenciadores das variações na bolsa de valores.

O desenvolvimento da aplicação nas linguagens R e Python também foi outro fator relevante da pesquisa. Uma vez que o Python apresenta pacotes muito mais robustos e com maior facilidade de manipulação dos dados, o usamos para o desenvolvimento do algoritmo de rastreamento e estruturação do banco. As análises foram realizadas através de pacotes específicos do software R. Apesar de não ser uma boa prática o desenvolvimento de estudos em liguagens diferentes, deste modo podemos aproveitar o que havia de melhor em cada um, facilitando o desenvolvimento e melhorando a qualidade da pesquisa.

Observamos que muitas ações podem ser incorporadas dentro da metologia apresentada, inclusive, evidenciamos no Capítulo 4 alguns pontos que poderiam ser incorporados no estudo da aplicação e poderiam enriquecer a análise do cenário, tais como: análise das fontes utilizadas na matéria para identificar termos mais relevantes; estudo dos comentários dos leitores para ter um panorama sobre a opinião dos leitores acerca de

cada assunto; e a análise dos compartilhamentos de cada matéria, que poderia evidenciar assuntos mais ou menos polêmicos.

Uma vez que a metodologia apresentada é bastante genérica, apesar do desenvolvimento de cada algoritmo depender da coletânea que se deseja estudar, é uma proposta futura sua aplicação para a análise de outras esferas da sociedade. A exemplo, podemos empregar estas mesmas técnicas para auxiliar advogados e juízes quanto a coleta, estratificação e análise dos principais conteúdos abordados em um grupo de jurisprudências. Esta ação seria muito pertinente, visto que exitem uma infinidade de jurisprudências disponibilizadas na web e a análise, uma a uma, pode tomar um tempo considerável de profissionais desta área. Além disto, poderiamos estudar o cenário economico de outros países. Nesta proposta, o desenvolvimento seria muito similar ao apresentado no Capítulo 4, no entanto, recuperando informações dos principais jornais e revistas de cada nação, por exemplo.

REFERÊNCIAS

ALBUQUERQUE, M. A.; FERREIRA, R. L. C. et al. Estabilidade em análise de cluster - estudo de caso em ciências florestais. *Sociedade de Investigações Florestais*, v. 30, n. 2, 2006.

AMO, S. d. *Análise de Clusters*: Introdução. [S.l.]: UFU, 2003. 9 p. Notas de Aula.

ARAUJO, J. M. P. *Processo de Descoberta de COnhecimento em Dados Não-Estruturados*: Estudo de caso para a inteligência competitiva. Dissertação (Mestrado em Informática) — Universidade Católica de Brasília, 2007.

ASSIS, A. K. T.; RAVANELLI, F. M. M. Reflexões sobre o conceito de centro de gravidade nos livros didáticos. *Ciência & Ensino*, v. 2, n. 2, Junho 2008.

BALL, G. H.; HALL, D. J. *Promenade - an On-line Pattern Recognition System*. [S.l.]: Defense Technical Information Center, 1967.

BATISTA, M. J. *Análise de Associação Aplicada ao Mapeamento Genético de Doenças.* Dissertação (Mestrado em Estatística) — Instituto de Matemática e Estatística, Universidade de São Paulo, 2006.

BELBIN, L. The use of non-hierarquical allocation methods for clustering large sets of data. *Australian Computer Journal*, v. 19, n. 1, Fevereiro 1987.

BERRY, M. W.; CASTELLANOS, M. *Survey of text mining: Clustering, Classification, and Retrieval*. 2. ed. [S.l.]: Springer, 2007.

BRAGA, T. M. *Uma Ferramenta de Mineração de Texto em Bancos de Dados de um Hospital Universitário Utilizando Decomposições Matriciais*. Dissertação (Mestrado em Engenharia Elétrica) — Universidade Federal do Rio de Janeiro, 2011.

BRAMER, M. *Principles of data mining*. [S.l.]: Springer, 2007. v. 131.

CADIMA, J. *Apontamentos de Estatística Multivariada*. [S.l.]: Departamento de Matemática, ISA - Universidade Técnica de Lisboa, 2010. 197 p. Notas de Aula.

CAMILO, C. O.; SILVA, J. C. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. [S.l.]: Instituto de Informática, Universidade Federal de Goiás, 2009. 29 p. Relatório Técnico.

CARRILHO, J. R. J. *Desenvolvimento de uma Metodologia para Mineração de Textos.* Dissertação (Mestrado em Engenharia Elétrica) — Pontifícia Universidade Católica do Rio de Janeiro, 2007.

CASTILLO, C.; BAEZA-YATES, R. *Web Crawler*. [S.l.]: Universitat Pompeu Fabra, 2010. 40 p. Notas de Aula.

CHAKRABARTI, S. *Mining the Web: Discovering knowledge from hypertext data*. [S.l.]: Elsevier, 2002.

CHEN, H. et al. Automatic concept classification of text from electronic meetings. *Communications of the ACM*, ACM, v. 37, n. 10, p. 56–73, 1994.

CIOS, K. J. et al. *Data Mining: A Knowledge Discovery Approach*. [S.l.]: Springer, 2007.

CORREA, G. N.; MARCACINI, R. M.; REZENDE, S. O. *Uso da mineração de textos na análise exploratória de artigos científicos*. [S.l.]: Instituto de Ciências Matemáticas e Computação, Universidade de São Carlos, 2012. 36 p. Relatório Técnico.

DEVMEDIA. *HTML Básico - códigos HTML*. 2014. Disponível em: ⟨http://www.devmedia.com.br/html-basico-codigos-html/16596⟩. Acesso em: Novembro 2015.

DONI, M. V. *Análise de Cluster*: Métodos hierárquicos e de particionamento. Monografia (Graduação em Sistemas de Informação) — Faculdade de Computação e INformática, Universidade Presbiteriana Mackenzie, 2004.

EVANS, R. Varieties of learning. *AI Game Programming Wisdom*, Charles River Media, Hingham MA, v. 2, 2002.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, v. 39, n. 11, p. 27–34, 1996.

FAYYAD, U. M. Data mining and knowledge discovery in databases: applications in astronomy and planetary science. In: AAAI PRESS. *Proceedings of the thirteenth national conference on Artificial intelligence-Volume 2*. [S.l.], 1996. p. 1590–1592.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996.

FILHO, L. A. d. S. *Mineração de Regras de Associação Utilizando KDD e KDT: Uma Aplicação em Segurança Pública*. Dissertação (Mestrado em Ciências da Computação) — Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, 2009.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: An overview. *AI magazine*, v. 13, n. 3, p. 57, 1992.

FREITAS, J. A. B. *Análise de Cluster da Lisozima*. [S.l.]: Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2006. 46 p. Relatório Científico.

FáVERO, L. P. et al. *Análise de Dados*: Modelagem multivariada para tomada de decisões. 2. ed. [S.l.]: Elsevier, 2009.

GLOBO, O. *Estudo da EMC prevê que volume de dados virtuais armazenados será seis vezes maior em 2020*. 2014. Disponível em: ⟨http://oglobo.globo.com/sociedade/tecnologia/estudo-da-emc-preve-que-volume-de-dados-virtuais-armazenados-sera-seis-vezes-maior-em-2020⟩. Acesso em: Setembro 2015.

HAN, J.; KAMBER, M. *Data Mining*: Concepts and techniques. 2. ed. [S.l.]: Elsevier Inc., 2006.

HUANG, Z. Clustering large data sets with mixed numeric and categorical values. In: SINGAPORE. *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD)*. [S.l.], 1997.

HäRDLE, W.; SIMAR, L. *Applied Multivariate Statistical Analysis*. 1. ed. [S.l.]: Springer, 2003.

JOHNSON, R. A.; WICHERN, D. W. et al. *Applied Multivariate Statistical Analysis*. 6. ed. [S.l.]: Pearson Prentice Hall Inc., 2007.

JONES, K. S. *Readings in information retrieval*. [S.l.]: Morgan Kaufmann, 1997.

LAROSE, D. T. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2014.

LOPES, C. C. *Um Sistema de Apoio à Tomada de Decisão no Acompnhamento do Aprendizado em Educação a Distância*. Dissertação (Mestrado em Informática) — Centro de Ciências e Tecnologia, Universidade Federal de Campina Grande, 2003.

LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of research and development*, IBM, v. 2, n. 2, p. 159–165, 1958.

MACHADO, A. P. et al. Mineração de texto em redes sociais aplicada à educação a distância. *Colabor@-A Revista Digital da CVA-RICESU*, v. 6, n. 23, 2010.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14.

MAIMON, O.; ROKACH, L. *Data mining and knowledge discovery handbook*. 2. ed. [S.l.]: Springer, 2005.

MANNING, C. D. et al. *Introduction to information retrieval*. [S.l.]: Cambridge university press Cambridge, 2008. v. 1.

MARCUS, M. P. *Theory of syntactic recognition for natural languages*. [S.l.]: MIT press, 1980.

MORAIS, E. A. M.; AMBRóSIO, A. P. L. *Mineração de Textos*. [S.l.]: Instituto de Informática, Universidade Federal de Goiás, 2007. 30 p. Relatório Técnico.

PAES, V. C. *Crawler de Faces na Web*. Dissertação (Mestrado em Ciência e Tecnologia da Informação) — Universidade Federal de Itajubá, 2012.

PRASS, F. S. *Estudo Comparativo entre Algoritmos de Análise de Agrupamentos em Data Mining*. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Santa Catarina, 2004.

REIS, T. *Algoritmo rastreador web especialista nuclear*. Tese (Doutorado) — Universidade de São Paulo, 2013.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informação da FSMA*, n, v. 7, p. 7–21, 2011.

SECOM. *Pesquisa Brasileira de Mídia 2015: hábitos de consumo de mídia pela população brasileira.* [S.l.]: Presidência da República - Secretaria de Comunicação Social, 2014.

SHARMA, S. *Applied Multivariate Techniques*. [S.l.]: John Wiley & Sons Inc., 1996.

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.

UFRJ. *Curso HTML - Resumo*. 2015. Disponível em: ⟨http://www.nce.ufrj.br/ginape/cursohtml/conteudo/introducao/resumo.htm⟩. Acesso em: Novembro 2015.

WITTEN, I. H. Text mining. Chapman & Hall/CRC Press, 2005.

WIVES, L. K. *Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estrturadas Usando Técnicas de "Clustering"*. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul, 1999.

ANEXO A – *Stoplist* utilizadas para pré-processamento da coletânea

Tabela 9 - Lista de *Stopwords*

#	Termo	#	Termo	#	Termo	#	Termo	#	Termo
1	de	45	sem	89	dela	133	hajam	177	temos
2	a	46	mesmo	90	delas	134	houvesse	178	tém
3	o	47	aos	91	esta	135	houvéssemos	179	tinha
4	que	48	seus	92	estes	136	houvessem	180	tínhamos
5	e	49	quem	93	estas	137	houver	181	tinham
6	do	50	nas	94	aquele	138	houvermos	182	tive
7	da	51	me	95	aquela	139	houverem	183	teve
8	em	52	esse	96	aqueles	140	houverei	184	tivemos
9	um	53	eles	97	aquelas	141	houverá	185	tiveram
10	para	54	você	98	isto	142	houveremos	186	tivera
11	com	55	essa	99	aquilo	143	houverão	187	tivéramos
12	não	56	num	100	estou	144	houveria	188	tenha
13	uma	57	nem	101	está	145	houveríamos	189	tenhamos
14	os	58	suas	102	estamos	146	houveriam	190	tenham
15	no	59	meu	103	estão	147	sou	191	tivesse
16	se	60	às	104	estive	148	somos	192	tivéssemos
17	na	61	minha	105	esteve	149	são	193	tivessem
18	por	62	numa	106	estivemos	150	era	194	tiver
19	mais	63	pelos	107	estiveram	151	éramos	195	tivermos
20	as	64	elas	108	estava	152	eram	196	tiverem
21	dos	65	qual	109	estávamos	153	fui	197	terei
22	como	66	nós	110	estavam	154	foi	198	terá
23	mas	67	lhe	111	estivera	155	fomos	199	teremos
24	ao	68	deles	112	estivéramos	156	foram	200	terão
25	ele	69	essas	113	esteja	157	fora	201	teria
26	das	70	esses	114	estejamos	158	fôramos	202	teríamos
27	à	71	pelas	115	estejam	159	seja	203	teriam
28	seu	72	este	116	estivesse	160	sejamos	204	,%.
29	sua	73	dele	117	estivéssemos	161	sejam	205	,%,
30	ou	74	tu	118	estivessem	162	fosse	206	$
31	quando	75	te	119	estiver	163	fôssemos	207	ser
32	muito	76	vocês	120	estivermos	164	fossem	208	sobre
33	nos	77	vos	121	estiverem	165	for	209	deve
34	já	78	lhes	122	hei	166	formos	210	nesta
35	eu	79	meus	123	há	167	forem	211	feito
36	também	80	minhas	124	havemos	168	serei	212	vem
37	só	81	teu	125	hão	169	será	213	disse
38	pelo	82	tua	126	houve	170	seremos	214	ter
39	pela	83	teus	127	houvemos	171	serão	215	porque
40	até	84	tuas	128	houveram	172	seria	216	pode
41	isso	85	nosso	129	houvera	173	seríamos	217	ainda
42	ela	86	nossa	130	houvéramos	174	seriam	218	vai
43	entre	87	nossos	131	haja	175	tenho		
44	depois	88	nossas	132	hajamos	176	tem		

ANEXO B – Algoritmo desencolvido para recuperação de dados em páginas HTML

```
from bs4 import BeautifulSoup
from datetime import date
import requests
import urllib
import csv


# FUNCOES

def BuscaPosts(linkStart, saida, total, Start = True):

    soup = BeautifulSoup(linkStart.content)

    for i in range(len(soup.find_all('article'))):
        link = soup('article')[i]('h2')[0]('a')[0].get('href')
        autor = soup('article')[i]('span')[0].string
        data = soup('article')[i]('time')[0].string[:10]
        titulo = soup('article')[i]('h2')[0].string
        saida.append((link, autor, data, titulo))

    if len(saida) < total:
        linkNext = soup('div', attrs={'class':'pag'})[0]('p',
           attrs={'class':'next'})[0]('a')[0].get('href')
        linkNext = requests.get(linkNext)
        BuscaPosts(linkNext, saida, total, Start = False)

    return saida

# buscando links dos posts com autor, titulo e data

r_start = requests.get('http://blogs.oglobo.globo.com/miriam-
   leitao/')

posts = []
BuscaPosts(r_start, posts, 500)

#for i in range(len(posts)):
    #print posts[i]
```

```
37
38 print len(posts)
39
40 # buscando texto dos posts cujo links ja foram apanhados
41
42 lista_posts = csv.writer(open('lista_posts.csv', 'w'))
43
44 for i in range(len(posts)):
45     t = posts[i][0]
46     t = requests.get(t)
47     soup_t = BeautifulSoup(t.content)
48     data = posts[i][2]
49     dt = date(int(data[6:]), int(data[3:5]), int(data[:2]))
50     autor = posts[i][1]
51     titulo = posts[i][3]
52     texto = soup_t('div', attrs = {'class': 'entry'})[0]('p')[2:]
           # retirando primeiros elementos que se referem a class tag
           e clas author
53     texto = texto[:len(texto)-4] # retirando dois ultimos
          paragrafos sempre em branco e compartilhar e comentar
54     txt = '' # elemento para adicionar o texto dos posts
55     for j in range(len(texto)):
56         txt = ' '.join([txt, texto[j].text.encode('utf8')])
57     lista_posts.writerow([data.encode('utf8'), autor.encode('utf8
          '), titulo.encode('utf8'), txt])
```