UNIVERSIDADE ESTADUAL DA PARAÍBA

CAMPUS 1 – CAMPINA GRANDE

CENTRO DE CIÊNCIAS E TECNOLOGIA

CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ELEONILIA MONTEIRO RODRIGUES

ANÁLISE E PREVISÃO DA EPIDEMIA DA FEBRE AMARELA UTILIZANDO DADOS DA REDE SOCIAL TWITTER E REDES BAYESIANAS

CAMPINA GRANDE

2019

ELEONILIA MONTEIRO RODRIGUES

ANALISE E PREVISÃO DA EPIDEMIA DA FEBRE AMARELA UTILIZANDO DADOS DA REDE SOCIAL TWITTER E REDES BAYESIANAS

Trabalho de conclusão de curso apresentado ao curso de Graduação em Bacharelado em Ciência da Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharelado em Computação.

Área de concentração: Ciência de dados e Big Data

Orientador: Prof. Dr. Vladimir Costa de Alencar.

CAMPINA GRANDE
2019

R696a Rodrigues, Eleonilia Monteiro.

Análise e previsão da epidemia da febre amarela utilizando dados da rede social Twitter e Redes Bayesianas [manuscrito] / Eleonilia Monteiro Rodrigues. - 2019.

51 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2019.

"Orientação : Prof. Dr. Vladimir Costa de Alencar , Coordenação do Curso de Computação - CCT."

1. Febre amarela. 2. Mineração de texto. 3. Twitter. I. Título

21. ed. CDD 600

ELEONILIA MONTEIRO RODRIGUES

ANÁLISE E PREVISÃO DA EPIDEMIA DA FEBRE AMARELA UTILIZANDO DADOS DA REDE SOCIAL TWITTER E REDES BAYESIANAS

Trabalho de Conclusão de Curso de Graduação em Ciência da Computação da Universidade Estadual da Paraíba, como requisito à obtenção do título de Bacharel em Ciência da Computação.

Aprovada em 28 de Junho de 2019.

Prof. Dr. Vladimir Costa de Alencar (UEPB)
Orientador(a)

Prof. MsC. Isaque Lyra (UEPB)
Examinador(a)

Profa. Dra. Maria do Socorro Rocha Melo Peixoto (UEPB)
Examinador(a)

Dedico esse trabalho a Vera Lucio M. Simões e Valdemir Rodrigues Lima, pela dedicação, companheirismo e amizade.

AGRADECIMENTOS

Primeiramente quero agradecer a Deus por ter me dado forças para superar as dificuldades.

Quero agradecer de todo o coração aos meus pais Vera Lucia M. Simões e Valdemir R. Lima (*in memoriam*), pelo apoio incondicional e muito provavelmente sem vocês essa caminhada não teria sido realizada e principalmente concluída.

Quero agradecer aos meus irmãos em especial a Nubia Andreia F. M. C. Feitosa por seu apoio e por acreditarem em mim em todos os momentos vivenciados nesses anos de graduação.

Quero agradecer a Rubsmercio Correia F. da Silva que veio somar de forma bastante positiva durante essa jornada, tornando-se um parceiro de todas as horas, obrigada!

Quero agradecer a todos os meus amigos que de maneira direta ou não me ajudaram, não citarei nomes pra não correr o risco de esquecer-me alguém, mas saibam que de alguma forma todos vocês, sejam os que caminharam junto comigo durante todo esse percurso, sejam os que fizeram parte depois ou os que por força do destino tiveram que escolher caminhos de vida diferentes, todos de alguma forma foram essenciais tanto na minha formação acadêmica como pessoal.

Quero agradecer a todo corpo docente e colaboradores da UEPB, que foi a instituição onde iniciei a vida acadêmica, quero agradecer especialmente aos professores do curso de computação. Quero agradecer ao professor Vladimir C. de Alencar pela competência e dedicação nas orientações e leituras sugeridas ao longo dessa orientação.

“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo”.

(Albert Einstein)

RESUMO

Este trabalho teve como finalidade identificar e analisar focos da febre amarela no Brasil através do uso de técnicas de mineração de textos e algoritmos de classificação Bayesianos. Foi utilizado técnicas de mineração de texto para coleta e preparação dos dados minerados da rede social Twitter, no período entre 13 de março a 4 de julho de 2018. Para isso, foram verificadas as mensagens dos usuários da rede social que relatassem sintomas da doença ou ainda mensagens com alguma relação ao dado pesquisado. Para identificar os casos de febre amarela por meio da rede social Twitter, foi criado um modelo que teve uma taxa de acurácia de 93% de acerto no dataset de treino/teste. Depois de termos identificado e analisado os dados, foi possível correlacionar os dados recolhidos e verificados na rede social com as informações dos meios oficiais de notificação. Os resultados mostraram uma alta correlação entre os dados da rede social Twitter com os dados oficiais do Ministério da Saúde que foram de 81% e 78% para o estado e a cidade respectivamente, mostrando que a rede social Twitter pode ser usada para análise e prevenção de epidemias. De acordo com os resultados obtidos na análise de correlação, foi possível concluir que as informações contidas na rede social do Twitter podem ser usadas como fonte de dados para análise e previsão de epidemias.

Palavras-Chave: Febre Amarela. Mineração de texto. Twitter.

ABSTRACT

The purpose of this project was identifying and analyzing yellow fever occurrences in Brazil, using data mining and Naive Bayes Models. It was using data mining to collect and extract data by Twitter, in May 13h until June 4h of 2018. Yellow fever symptoms or related filtered the tweets. The model that identified yellow fever occurrences by tweet had 93% accuracy in train/test dataset. Identified and analyzed the tweet data, it has correlated with official data. The results showed that it had a high correlation between Twitter data and official data; it was 81% to the state data and 78% to the city data. This research concludes that Twitter is a powerful and safe social media to collect data to analyze and predict epidemics.

Keywords: Yellow Fever. Data Mining. Twitter.

LISTA DE FIGURAS

Figura 1 – Diagrama de fluxo do processo de Mineração de dados.....20
Figura 2 – Matriz BOW (Bag-of-Words).....22
Figura 3 – Representação de um SGBD chave-valor.....25
Figura 4 – Representação de um SGBD orientado a coluna 26
Figura 5 – Representação de um SGBD orientado a grafos.....27
Figura 6 – Representação de um SGBD orientado a documentos.....28
Figura 7 – Um documento JSON 30
Figura 8 – Exemplo envolvendo a pratica de esporte, baseado na condição meteorológica...35
Figura 9 – Um objeto JSON - Parte de uma informação recuperada de um tweet.....38
Figura 10 – Classificação dos tweets.....39
Figura 11 – Ocorrência de tweets por estado – Febre amarela 2018.....41
Figura 12 – Ocorrência de tweets por cidade – Febre amarela 2018 42

LISTA DE QUADROS

Quadro 1 - Vantagens e desvantagens do uso do MongoDB 32
Quadro 2 - Comparação entre Inferência Clássica e Inferência Bayesiana 33

LISTA DE TABELAS

Tabela 1 - Os 11 SGBDs mais utilizados 30
Tabela 2 - Resultados da análise de correlação entre os dados da SVS e do Twitter por UF . 43
Tabela 3 - Resultados da análise de correlação entre os dados da SVS e do Twitter por município 44

SUMARIO

1 INTRODUÇÃO 12

2 OBJETIVOS 14

2.1 Objetivos Gerais 14

2.2 Objetivos Específicos 14

3 A FEBRE AMARELA 15

4 FUNDAMENTAÇÃO TEÓRICA 17

4.1 A Mídia Social Twitter 17

4.2 Big Data *Analytics* 18

4.3 Mineração de texto 19

4.3.1 Determinação dos objetivos e metas do estudo 20

4.3.2 Exploração dos dados 20

4.3.3 Preparação dos dados e desenvolvimento e validação do modelo 20

4.3.4 Avaliação e Implantação dos Resultados 23

5 BANCOS DE DADOS NÃO-RELACIONAIS (NOSQL) 24

5.1 Chave-Valor 24

5.2 Orientado a Coluna 25

5.3 Orientado a Grafos 26

5.4 Orientado a Documentos 27

6 O MONGODB 29

7 O MÉTODO BAYESIANO 33

8 METODOLOGIA 36

8.1 O Uso do Twitter 36

8.2 Armazenamento dos tweets no MongoDB 37

8.3 Classificações dos tweets 39

9 RESULTADOS 41

10 CONCLUSÃO 45

REFERÊNCIAS 46

1 INTRODUÇÃO

Com o avanço das tecnologias, o aumento do uso das redes sociais (Twitter, Facebook, blogs entre outros) e o surgimento da Internet das Coisas (IoT), uma grande quantidade de dados são gerados e salvos em bancos de dados estruturados, semiestruturados ou não estruturados. Esse volume de dados tem despertado o interesse de diversas corporações, governos e estudiosos, no intuito de ajudar na tomada de decisões baseado na análise dos dados, concebendo assim, diversas oportunidades no comércio, inovação e políticas públicas de saúde (Hashem, 2015).

Segundo Data Science Academy (2018a), 90% dos dados gerados no planeta foram gerados nos últimos dois anos, destes dados, cerca de 80% são não estruturados ou estão em diferentes formatos, o que dificulta o gerenciamento e a análise. O aumento exponencial dos dados se deve ao advento da internet e de dispositivos como celulares, tablets, computadores entre outros.

A rápida expansão no volume de dados com características diversas vezes não estruturadas, fez com que os meios tradicionais de análise e gerenciamento não fossem suficientes. Segundo Data Science Academy (2018a), os modelos de análise de dados estruturados possuem limitações quando precisam tratar grande volume de dados, pois não é possível usar um banco de dados relacional (endereçar toda a informação em linhas e colunas de bancos de dados). Isso ocasionou o surgimento do conceito de Big Data que nos permite descobrir padrões e correlações nos dados, nos proporcionam conhecimento útil que permitem empresas, indústrias e governos, tomem as melhores decisões e ofereçam os melhores serviços e produtos (DAVENPORT, 2012).

O conhecimento é essencial para nos ajudar na tomada de decisões em diversas áreas, o que não é diferente no campo da medicina, logo existe a necessidade de investigar e automatizar métodos que possam auxiliar na obtenção de conhecimento (FREITAS, 2006).

Gonçalves (2012), afirma que a maior parte dos dados disponíveis no mundo não está em banco de dados relacionais, ou seja, se encontra digitalizado em forma de texto, por exemplo, livros, revistas, jornais, arquivos PDF, arquivos JSON, e-mails, dentre outros. Quando se trabalha com os dados textuais, deve-se levar em consideração, que não estão costumeiramente organizados em campos, como ocorre com as informações gravadas em bancos de dados tradicionais. Portanto, ao se comparar com as informações inseridas em SGBDs (Sistemas de Gestão de Base de Dados) relacionais, os dados textuais são mais difíceis de coletar, tratar, analisar e sumarizar. Isso determinou o aparecimento da mineração

de texto (text mining), que é uma subárea da mineração de dados, do inglês Data Mining (pratica de examinar grandes quantidades de dados à procura de padrões consistentes).

Para Aranha e Passos (2006), a mineração de texto é o processo que usa técnicas capazes de recuperar informações com o objetivo de extrair conhecimento. A tecnologia de mineração de texto usa Machine learning (aprendizado de máquina), que é um subcampo da inteligência artificial, que utiliza técnicas de análise de dados que automatiza o desenvolvimento dos modelos analíticos, ela pode ser usada em problemas quantitativos e qualitativos.

A UN Global Pulse (2015) realizou um estudo que acompanhava as atividades de países como Índia, Quênia, Nigéria e Paquistão, nas redes sociais Twitter e Facebook, com o objetivo de analisar as conversas relacionadas à vacinação no período de janeiro até dezembro de 2014. O projeto mostrou que é possível apoiar profissionais de saúde e campanhas de comunicação, através da análise de sentimentos, classificação de tópicos e análise de redes.

Alguns trabalhos como o de Alencar e Almeida (2016), já fizeram uso da mineração de texto e correlação dos dados, para ajudar a identificar casos suspeitos da febre Chikungunya no Brasil, a partir dos sintomas relatados por usuários da rede social Twitter. Dessa forma, os tweets foram classificados como: verdadeiro e falso. O resultado da classificação permitiu identificar casos suspeitos da febre Chikungunya no Brasil, através de técnicas de mineração de texto.

Desse modo, o objetivo do projeto foi criar um modelo de coleta de dados da mídia social Twitter, com a finalidade de identificar a doença tropical febre amarela, em municípios do Brasil, correlacionando os dados coletados dessa mídia social com os dados oriundos de fontes oficiais, no caso, Secretaria de Vigilância em Saúde, subsidiando informações e análises para a tomada de decisão e o planejamento de políticas públicas.

2 OBJETIVOS

2.1 Objetivos Gerais

Objetivo deste trabalho é identificar casos da doença tropical febre amarela no Brasil, através da coleta dados na rede social Twitter, usando técnicas Big Data *Analytics* para ajudar na tomada de decisões do Programa de Controle de Doenças Tropicais.

2.2 Objetivos Específicos

- Coleta de dados textuais (*tweets*) do Twitter referente à doença tropical febre amarela usando banco de dados NoSQL (Bancos de Dados Não-Relacionais).
- Utilizar mineração de texto para extrair informações sobre a doença tropical febre amarela tal como: sintomas, geolocalização, dentre outros dados.
- Distribuir geograficamente os casos novos da febre amarela, por município.
- Analisar a ocorrência de casos da febre amarela, correlacionando com caraterísticas clínicas, epidemiológicas, sociais e ambientais.
- Fornecer informações para subsidiar processos decisórios de gestão do Programa de Controle de doenças tropicais.

3 A FEBRE AMARELA

As doenças tropicais infecciosas que afetam principalmente, as regiões tropicais e subtropicais ou, mais seguidamente, são as que mais se espalham nos trópicos ou mais difíceis de prevenir e controlar. Elas podem prosperar em tais regiões por fatores biológicos, ecológicos, evolutivos e socias que apoiam níveis elevados de patogênicos e vetores, assim como os fatores sociais que dificultam o controle destas doenças, tendo um forte componente de subdesenvolvimento (CAMARGO, 2008).

A Organização Mundial da Saúde (OMS) desenvolveu um Programa Especial para Pesquisa e Treinamento em Doenças Tropicais (*Especial Programe for Researchand Training in Tropical Diseases* - TDR), o qual acompanha as doenças infecciosas negligenciadas, que afetam de maneira desproporcional populações pobres e marginalizadas, atualmente sendo objeto de interesse 16 doenças: Malária; Febre Amarela; Tripanossomíase africana; Dengue; Leishmaniose; Esquistossomose; Tuberculose; Doença de Chagas; Hanseníase; Filariose Linfática; Oncocercose; Ebola; Hepatite; Conjuntivite; Zika Vírus (OMS, 2017).

Conforme o registrado em Brasil (2017), entre o final de 2016 a março de 2017, ocorreu o maior surto de febre amarela observado em anos no Brasil, que envolveu principalmente os estados da região a sudeste, em particular Minas gerais e Espírito Santo. Nesse período, foram notificados 1.561 casos suspeitos, dos quais 448 casos foram confirmados, 263 foram descartados e os demais seguem em investigação. No mesmo período, foram confirmados 144 óbitos tendo a febre amarela como causa.

Conforme a Sociedade Brasileira de Infectologia (2017) a doença infecciosa febre amarela, ocorre em áreas de florestas tropicais da América do Sul e da África, e muitas vezes ocorrem em forma de surtos e epidemias com impacto em saúde pública. No caso dessa doença, a causa é o vírus do gênero *Flavivirus da* família *Flaviviridae*. Com relação à Febre Amarela Silvestre (FAS), o mosquito do gênero *Haemagogus Janthinomys* é o principal vetor ou reservatório e os hospedeiros naturais são os primatas não humanos (macacos). Os seres humanos não imunes podem participar como hospedeiro acidental, entrando em áreas enzoóticas. Na Febre Amarela Urbana (FAU), o homem é o único hospedeiro de importância epidemiológica, tendo o mosquito *Aedes aegypti* (fêmea infectada) como principal vetor e reservatório.

Segundo dados do Ministério da saúde a febre amarela é uma doença febril aguda, de curta duração e gravidade variável. Apresenta-se como infecções subclínicas e/ ou leves, até formas graves e fatais. O quadro clínico tem início abrupto com febre alta e pulso lento em

relação à temperatura (sinal de Faget), dor de cabeça intensa, náuseas e vômitos, calafrios, mialgias, que duram cerca de 3 dias, quando se observa remissão da febre e melhora dos sintomas, o que pode durar algumas horas ou, no máximo 2 dias. Dependendo do caso pode haver uma evolução para cura ou para a forma grave (período de intoxicação), onde há aumento da febre, diarreia e vômitos com aspecto de borra de café, desenvolvimento de insuficiência hepática e renal; além da coloração amarelada da pele e do branco dos olhos, manifestações hemorrágicas, oligúria, albuminúria e prostração intensa, além de comprometimento do sensório, que se expressa mediante alteração do estado de consciência com evolução para coma. O diagnóstico é clínico, epidemiológico e laboratorial, não havendo tratamento específico, apenas sintomático (Brasil, 2010).

Brasil (2010) considera como caso suspeito da febre amarela, o indivíduo que apresenta um quadro de febre aguda, de até 7 dias, com coloração amarelada da pele e da parte branca dos olhos e/ou manifestações hemorrágicas, que não tenha sido vacinado contra febre amarela ou com estado vacinal ignorado. Assim como o indivíduo com quadro febril agudo, residente ou que esteve em área com transmissão viral nos últimos 15 dias, não vacinado contra febre amarela. A doença é de notificação obrigatória em todo território nacional, assim como sua investigação epidemiológica de todos os casos, com a finalidade de detectar oportunamente a circulação viral para orientar as medidas de controle, reduzir a incidência e impedir a transmissão urbana.

Com isto, utilizamos a rede social Twitter como instrumento de coleta de informações referente a febre amarela no Brasil, e utilizamos técnicas de mineração de texto e Redes Bayesianas para identificação de possíveis casos da epidemia.

4 FUNDAMENTAÇÃO TEÓRICA

4.1 A Mídia Social *Twitter*

O Twitter é uma rede social de *microblogging* lançado em 2006 nos Estados Unidos usada por milhões de pessoas em todo o mundo. Como exposto por Vinha (2017) a mídia social Twitter inicialmente possibilitava aos seus usuários que se expressassem em textos de até 140 caracteres (*tweets*), mas essa quantidade de caracteres aumentou para 280 em meados de novembro de 2017. Além disso, possui suporte para mais de 37 idiomas.

Essa mídia social funciona por meio da criação de uma conta, que dará acesso a uma página onde o usuário poderá publicar mensagens, que, segundo Primo (2008), o Twitter convida os usuários a responder a pergunta "o que você está fazendo?", e é um serviço que permite que as pessoas façam breves narrativas. A rede social oferece aos usuários a liberdade para expressar o que desejam, desde mensagens simples às mais elaboradas, além disto, permite que acompanhem as novas publicações de outras pessoas através de um sistema de assinaturas, onde há possibilidade de escolher quais pessoas "seguir" e ser "seguido" por outras.

Segundo Primo (2008), a velocidade com que se publicam as mensagens (*tweets*) na rede, torna-se comum que o Twitter consiga divulgar notícias com muito mais agilidade que qualquer meio jornalístico tradicional. De acordo com Recuero e Zago (2009), as atualizações em sistemas do Twitter trazem informações como hora e data de postagem, que permite facilmente que os *tweets* sejam repassados de uma rede para outra. As postagens podem ser afirmações e observações simples sobre o cotidiano, mensagens mais elaboradas com links externos, anexar fotos, vídeos entre outros.

De acordo com o Twitter (2019a), o site/aplicativo oferece vários recursos que visam melhorar a experiência do usuário como o *retweet*, os *Trending Topics*, *hashtags* e as APIs (*Application Programming Interfaces*). O *retweet* possibilita republicar com rapidez uma mensagem de um usuário para seu perfil como forma de referenciar o autor original. Os *Trending Topics* ou "Assuntos do momento" são os assuntos mais comentados na rede social em tempo real. Os usuários utilizam as *hashtags* que são compostas pela palavra-chave do assunto antecedida pelo símbolo cerquilha *"#",* para tornar o conteúdo da postagem acessível a todos que tenham interesse.

Segundo a própria rede social Twitter (2019b), a plataforma oferece a empresas, desenvolvedores e usuários, acesso programático a dados públicos do Twitter por meio das

Interfaces de Programação de Aplicativos (APIs). Estas APIs permitem que os programadores desenvolvam softwares que se integram ao Twitter.

Presse (2019) afirma que o Twitter possui cerca de 330 milhões de usuários que frequentaram a rede social no período de janeiro a março de 2019, um aumento de nove milhões em relação ao trimestre anterior. Segundo dados do próprio Twitter (2019c), mais de 80% dos seus usuários usam dispositivos móveis, o que resulta em informações e interações realizadas com mais frequência e em muitas vezes instantaneamente. Além disso, a empresa tem mais de 35 escritórios em todo o mundo.

4.2 Big Data *Analytics*

Big Data hoje é uma realidade, Silva (2016) define como sendo um conjunto de dados cuja coleta, armazenamento, distribuição e análise necessitam de métodos e tecnologia avançadas, por causa da velocidade, variedade, complexidade e o grande volume de dados. Esse extraordinário crescimento significa que não devemos só compreender Big Data para conseguir elucidar as informações importantes, mas também devemos entender as possibilidades que a análise apropriada da Big Data pode trazer para as organizações (Taurion, 2013).

De acordo com Silva (2016), todos os tipos de comunicação moderna são digitais. A quantidade de pessoas que usam dispositivos digitais para se comunicar está na casa dos bilhões, alcançando à cobertura completa em várias partes do mundo. Com a grande demanda de informações que esses dispositivos têm gerado, tornou-se viável decidir as medidas de prevenção e opções de tratamento, através da análise dos dados. É uma importante oportunidade para países de dimensões continentais, explorar os dados originários da digitalização das comunicações, para auxiliar os sistemas de vigilância epidemiológicos tradicionais.

Analytics aplicado em Big Data é o processo de examinar grandes volumes de dados com intuito de descobrir padrões escondidos, correlações desconhecidas e outras informações úteis que podem ser usadas para a melhor tomada de decisão. Com a análise de Big Data, cientistas de dados e outras pessoas podem analisar grandes volumes de dados, que a análise tradicional e as soluções de *business intelligence* não conseguem analisar (ANKAM, 2016).

Segundo Loshin (2013), muitas organizações acumulam bilhões de dados com centenas de milhões de combinações de dados em várias memórias de dados e numerosos

formatos. Big Data *analytics* está em alta, pois é necessário para processar essa quantidade de dados para que se possa descobrir o que é importante e o que não é.

4.3 Mineração de texto

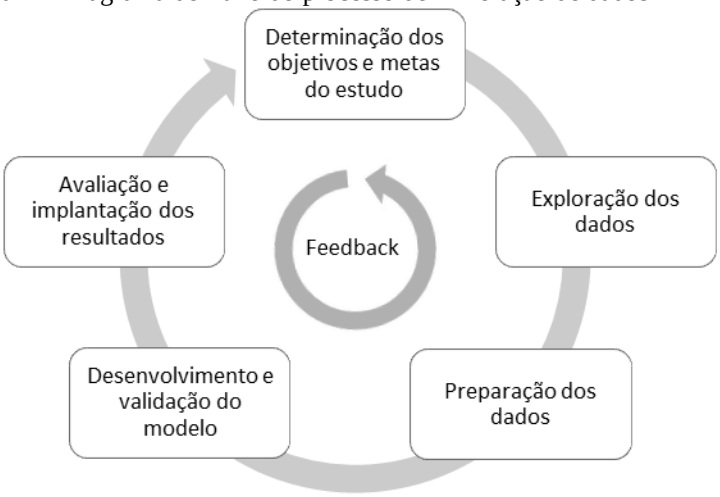
De acordo com Faro et al (2012) a mineração de texto, tem como o objetivo principal encontrar conhecimento oculto em textos e expor de forma lógica e concisa. Morais (2017) define a mineração texto como sendo o processo de descoberta de conhecimento, que usa análise e técnicas de extração de dados a partir de textos, ou seja, documentos em linguagem natural, que normalmente possuem pouca ou nenhuma estrutura de dados.

Segundo Morais (2007) a origem da mineração de texto está relacionada à área de Descoberta de Conhecimento em Textos (*Knowledge Discovery from Text* - KDT), que mostra como extrair informações relevantes de coleções de textos. O KDT é baseado no processo de mineração de textos, que abrange a recuperação de informação, análises textuais, extração de informação, clusterização, categorização, visualização e tecnologias de base de dados. Além disso, segundo Silva (2002), a mineração de texto envolve diversas áreas de conhecimento como: inteligência artificial, linguística, estatística, computação, aprendizagem de máquina e mineração de dados.

O processo de descoberta de conhecimento, de acordo com Miner et al. (2012), pode ser dividido em cinco partes básicas: determinação dos objetivos e metas do estudo, exploração dos dados, preparação dos dados, desenvolvimento e validação do modelo, avaliação e implantação dos resultados.

O diagrama de fluxo do processo de mineração dos dados (Figura 1) apresenta as cinco etapas básicas da mineração. Na imagem temos o feedback com uma seta no sentido anti-horário, indicando que pode haver um retorno em cada uma das fases visando melhorar o processo e realizar possíveis correções.

Figura 1 – Diagrama de fluxo do processo de Mineração de dados



Fonte: Alencar e Almeida (2016)

4.3.1 Determinação dos objetivos e metas do estudo

Determinar os objetivos e metas de estudo é o início do processo de mineração de texto, o que demanda conhecimento sobre a área do estudo onde se pretende realizar o processo de negócio. A mineração de texto tem como finalidade extrair informações úteis do grande volume de dados da pesquisa. Para que seja possível, a maioria das vezes faz-se necessário a participação de outros especialistas da área de estudo, com intuito de nos apresentar uma visão mais desenvolvida do assunto em questão, para que seja definido objetivos e metas (Miner et al., 2012).

4.3.2 Exploração dos dados

Para Miner et al. (2012) nesta fase é realizada a exploração da viabilidade e a natureza dos dados, onde se juga os dados usados são viáveis, acessíveis e aplicáveis em relação ao contexto do estudo. A exploração dos dados requer identificação da fonte dos dados, avaliação da acessibilidade e usabilidade dos dados. Além disso, é nessa fase que é realizada a coleta dos dados que serão empregados no estudo como também a avaliação quantitativa e qualitativa dos dados.

4.3.3 Preparação dos dados e desenvolvimento e validação do modelo

As etapas de preparação dos dados, desenvolvimento e validação do modelo são as realizações do que foi planejado anteriormente. Com relação à preparação dos dados, os fatores-chave para determinar o quanto o algoritmo de aprendizado maquina vai aprender depende da qualidade e quantidade de dados com informações significativas.

Conforme registrado por Madeira (2015), a mineração de texto tem como finalidade identificar padrões ou tendências em um grande volume de dados textuais, com intuito de extrair informações úteis.

Essa etapa pode ser dividida em três atividades: Formação do conjunto de treinamento (Preparação dos dados), pré-processamento dos dados e extração de conhecimento.

4.3.3.1 Preparação dos dados

De acordo com Morais (2007), a preparação dos dados é a primeira etapa do processo de descoberta de conhecimento. Nesta etapa são coletados os dados que constituem a base de textos que são relevantes para o problema em questão. Depois de coletados, os dados são organizados em um formato adequado para que sejam processados. A fase de preparação dos dados tem como objetivo tentar identificar semelhanças em função da morfologia ou do significado dos termos.

Para Alencar e Almeida (2016) a qualidade e a quantidade dos dados textuais são os elementos mais importantes para um projeto de mineração de texto, visto que, é por meio deles que será obtido o conhecimento útil. Entretanto, o conjunto de dados também pode ser preparado de forma separada manualmente e apresentado para a etapa de pré-processamento dos dados.

4.3.3.2 Pré-processamento dos dados

Esta etapa é executada após a formação do conjunto de dados, poderá ser usada para ajudar na escolha do melhor conjunto de dados, para extrair padrões e tendências. De acordo com Madeira (2015), a fase de pré-processamento dos dados tem como objetivo prover formatação e representação dos dados de forma sucinta e manipulável por algoritmos de extração de conhecimento.

Para converter o conjunto de dados minerados em uma estrutura de representação é comumente usado o modelo de saco de palavras (BOW - Bag of Words) também chamada de matriz de termos e documentos. A matriz BOW é usada no processamento de linguagem

natural para representar dados textuais, onde cada termo (coluna) representa um atributo (palavra) e cada linha representa um documento (item do documento ou frase) e o valor do campo representa a frequência com que os termos aparecem no documento (ANDRADE, 2015). Na Figura 2 é ilustrado um exemplo da matriz BOW.

Figura 2 – Matriz BOW (Bag-of-Words)

	Termo 1	**Termo 2**	**Termo...**	**Termo j**
Documento 1				
Documento 2				
Documento...				

Fonte: Alencar e Almeida (2016)

Para extrair informações do texto em linguagem natural, faz-se necessário usar algumas técnicas de pré-processamento: *tokenização*, *stemming*, *stopwords*, entre outras.

Tokenização é o primeiro passo da etapa de pré-processamento, que de acordo com Andrade (2015), esta técnica tem como finalidade dividir o texto em unidades pequenas. Cada unidade é chamada de *token* e que, geralmente podem ser palavras ou termos compostos, números e espaços e caracteres de pontuação. Essa técnica é realizada identificando os espaços e pontuações que costumam delimitar termos.

A técnica de *stemming* é o segundo passo da etapa de pré-processamento, utilizada para reduzir os termos em formas mais básicas, removendo variações morfológicas como afixo, vogais temáticas e desinências, com intuito de reduzir o número de termos da matriz BOW. Outra técnica aplicada nessa etapa é a remoção de *stopwords*, que são uma lista de termos, sem valor semântico. Os termos classificados como *stopwords* são artigos, preposições, conjunções, e termos geralmente usados no conjunto de dados que não acrescentam valor ao texto (ANDRADE, 2015).

O último passo da etapa de pré-processamento é o cálculo da relevância dos termos. Segundo Morais (2007), nem todos os termos de um documento tem o mesmo peso, alguns aparecem com mais frequência (com exceção das *stopwords*) e por isso, costumam ser mais importantes. Uma das formas mais comuns utilizadas para calcular a relevância dos termos é o tf-idf (*term frequency – inverse document frequency*). Esta medida tem como objetivo definir a importância dos termos dentro de cada coleção de documentos analisados.

4.3.3.3 Extração de conhecimento

Esta atividade é responsável pela aquisição do conhecimento por meio de um conjunto de dados, que nos permite identificar grupos de informações que estão correlacionadas. Para extrair o conhecimento da base de dados são utilizados os resultados obtidos na etapa de pré-processamento, em que são empregadas técnicas de mineração de texto para obter informações ou padrões a partir dos dados da matriz de termos e documentos.

4.3.4 Avaliação e Implantação dos Resultados

Esta fase envolve a avaliação e implantação dos resultados obtidos a partir do desenvolvimento e validação do modelo. Portanto, após o modelo ter sido avaliado quanto à qualidade e acurácia por meio de uma visão analítica dos dados, faz-se necessário avaliar a informação obtida e analisar o seu significado, que diversas vezes requer análise de especialistas da área e o uso de técnicas adequadas de visualização de informações.

Depois de passar por um processo de validação, temos a avaliação dos resultados, que pode ser feita de forma objetiva ou subjetiva. De acordo com Madeira (2015), a avaliação objetiva é realizada por intermédio de índices estatísticos que mensuram a qualidade dos resultados obtidos, já a avaliação subjetiva faz uso do conhecimento de especialistas da área.

É importante armazenar os dados coletados em um RGBDs não relacional (NoSQL - Bancos de Dados Não-Relacionais), pois são bancos de dados distribuídos, que foram projetados para armazenar grande volume de dados, por possuírem uma arquitetura mais escalável e eficiente que RGBDs relacionais. (DATA SCIENCE ACADEMY, 2018b).

5 BANCOS DE DADOS NÃO-RELACIONAIS (NOSQL)

De acordo com Toth (2011) com o enorme crescimento da quantidade de dados e informações na Web, a manipulação e processamento tornou-se um dos grandes desafios na área de Computação. Por se tratar de um conjunto de dados complexo e volumoso, que torna difícil realizar operações simples (inserções, buscas e alterações) de forma eficiente usando banco de dados relacionais. Dessa forma, surge o paradigma não tradicional NoSQL para resolver os problemas de flexibilidade, escalabilidade e alto desempenho.

Segundo Aniceto e Xavier (2014) o termo NoSQL foi utilizado pela primeira vez em 1998 para criar um banco de dados *open-source*, que omitia o uso da linguagem SQL. No ano de 2009, o termo foi retomado no evento "NoSQL *Meetup*", dessa vez o NoSQL foi usado referenciando um banco de dados não relacional.

Os bancos de dados não relacionais (NoSQL) surgiram como um paradigma não tradicional para lidar com grande volume de dados e para resolver desafios que surgiram com a chegada de implementações de Big Data, buscando alto desempenho e disponibilidade. Esse tipo de banco de dados tem como característica prover escalabilidade mais barata e menos trabalhosa, permitindo trabalhar com dados semiestruturados ou não estruturados, vindos de diversas fontes tais como: arquivos de blogs, websites, arquivos de multimídias, entre outros (DATA SCIENCE ACADEMY, 2018b).

De acordo com Amazon Web Services (2018a), para aplicações que exigem flexibilidade, escalabilidade, alto desempenho e são altamente funcionais, o NoSQL é ideal pois na maioria das vezes esse tipo de SGBDs proporcionam esquemas flexíveis, são arquitetados para serem facilmente escaláveis horizontalmente, foi melhorado para modelos de dados específicos e paradigma de acesso que proporcionam melhor performance. Além disso, disponibilizam APIs simples e tipos de dados altamente funcionais criados com foco em cada um de seus respectivos modelos de dados.

Bancos de dados NoSQL ofertados 4 categorias de banco de dados, que são respectivamente: chave-valor, orientado a documentos, orientado a coluna e orientado a grafos.

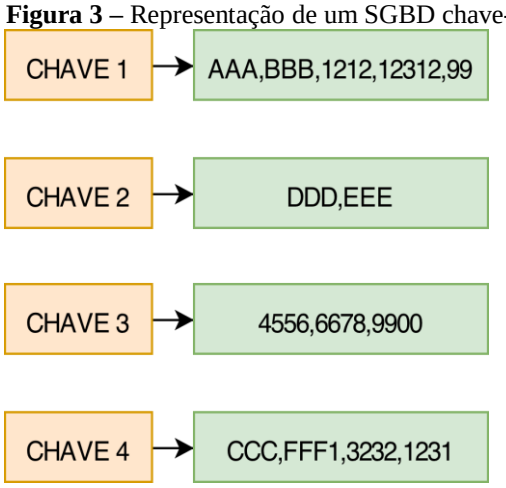
5.1 Chave-Valor

O modelo chave-valor é o mais simples e fácil de programar, também pode ser chamado de tabela *hash* distribuída. Para Aniceto e Xavier (2014), sua forma de armazenar os

dados se dá com base similar a um dicionário, na qual existe uma chave única e um indicador de um dado ou um item em particular. Essa forma de armazenamento é livre de esquemas, por ter uma estrutura simples. Novos dados podem ser inseridos em tempo real de execução, sem que ocorram conflitos com dados já existentes no banco.

A Amazon Web Services (2018a) descreve esse modelo como sendo bastante particionável que permite escalonamento horizontal. Geralmente usado em aplicações cache de conteúdo (grandes quantidades de dados e carregamento massivo), pesquisas rápidas, *logging* (registros de eventos importantes). Mas de acordo com Vardanyan (2011), não é eficiente quando está interessado em apenas consultar ou em atualizar parte de um valor. Podemos citar como exemplo de banco de dados chave-valor o Riak e Berkeley. A Figura 3 ilustra um exemplo do modelo chave-valor.

Figura 3 – Representação de um SGBD chave-valor



Fonte: Queiroz (2017)

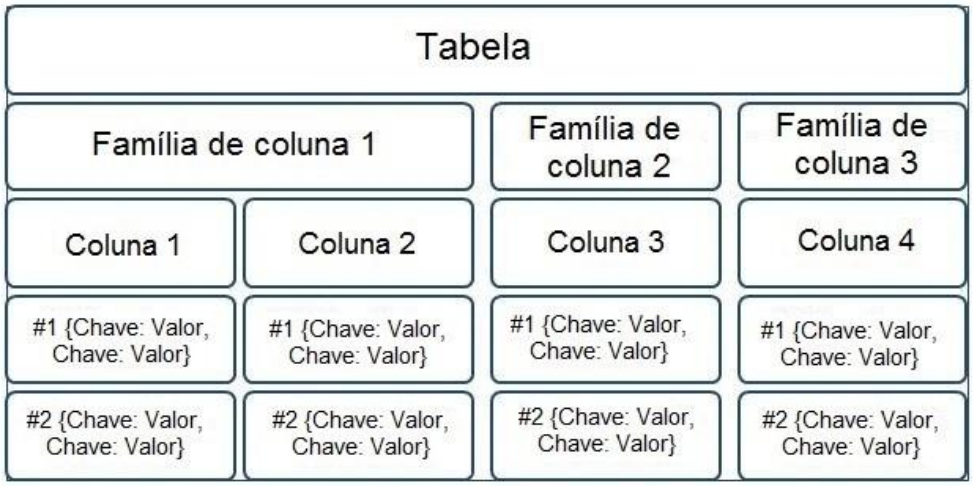
5.2 Orientado a Coluna

O modelo orientado a coluna é descrito por Souza et al (2014) como sendo um pouco mais complexo que o modelo chave-valor, nele mudamos o paradigma relacional de orientação a registro ou tuplas para orientação a atributos ou colunas (NoSQL). Neste modelo os dados são indexados por uma tripla (linha, coluna e *timestamp*), onde ainda existem chaves que identificam as linhas e colunas e o *timestamp* permite distinguir múltiplas versões de um mesmo dado.

De acordo com Lopes (2014) a forma que o banco de dados orientado a coluna é estruturado, permite que os atributos possam ser agrupados em famílias de colunas, modificando assim a organização dos dados e possibilitando o particionamento do banco,

podendo também serem adicionadas a qualquer momento novas colunas e linhas. Mas segundo Aniceto e Xavier (2014) um dos pontos negativos deste modelo é que ele possui uma menor flexibilidade em relação aos modelos chave-valor e orientado a documento, pois uma família de colunas não pode ser definida durante a execução. Podemos citar como exemplo bancos de dados nessa categoria HBase e Hypertable. Na figura 4 temos um exemplo de um banco de dados NoSQL orientado a coluna.

Figura 4 – Representação de um SGBD orientado a coluna



Fonte: Micreiros.com (2017)

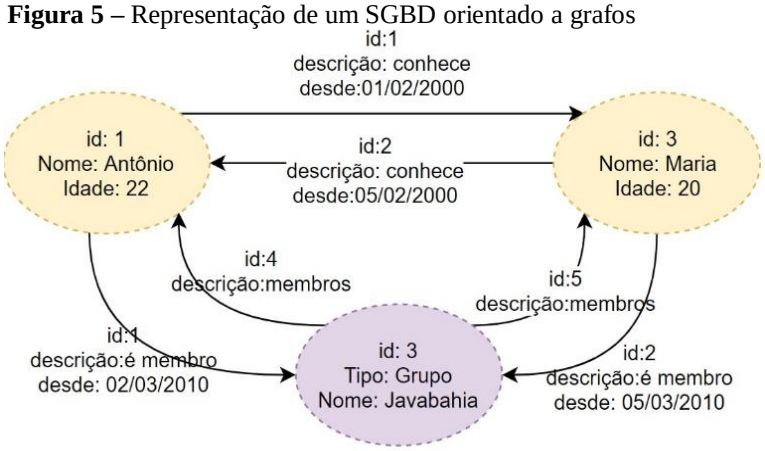
5.3 Orientado a Grafos

De acordo com Sousa (2015) o modelo orientado a grafos tem como finalidade armazenar relacionamentos entre os dados e é baseado no conceito de grafos. Este tipo de modelo possui três componentes básicos que são: os nós (são os vértices do grafo), relacionamentos (são as arestas) e as propriedades (ou atributos) dos nós e relacionamentos. Para Lopes (2014), esse modelo tem como objetivo uma performance aprimorada na gestão eficiente de bases onde os dados são fortemente ligados.

Segundo Amazon Web Services (2018a), o objetivo do banco de dados orientado a grafo é facilitar o desenvolvimento e a execução de aplicativos que trabalham com conjunto de dados conectados e podem ser usados em casos que incluem mecanismos de recomendação, detecção de fraudes, redes sociais e gráficos de conhecimento. Assim, de acordo com Diana e Gerosa (2010) “esse modelo também dá suporte ao uso de restrições sobre os dados, como restrições de identidade e de integridade referencial”, em contrapartida da maioria dos bancos de dados não relacionais que são mais flexíveis. Os bancos de dados

orientados a grafos mais populares são Neo4j e Giraph. Pode-se observar na Figura 5 um exemplo do modelo orientado a grafo.

Figura 5 – Representação de um SGBD orientado a grafos



Fonte: Queiroz (2017)

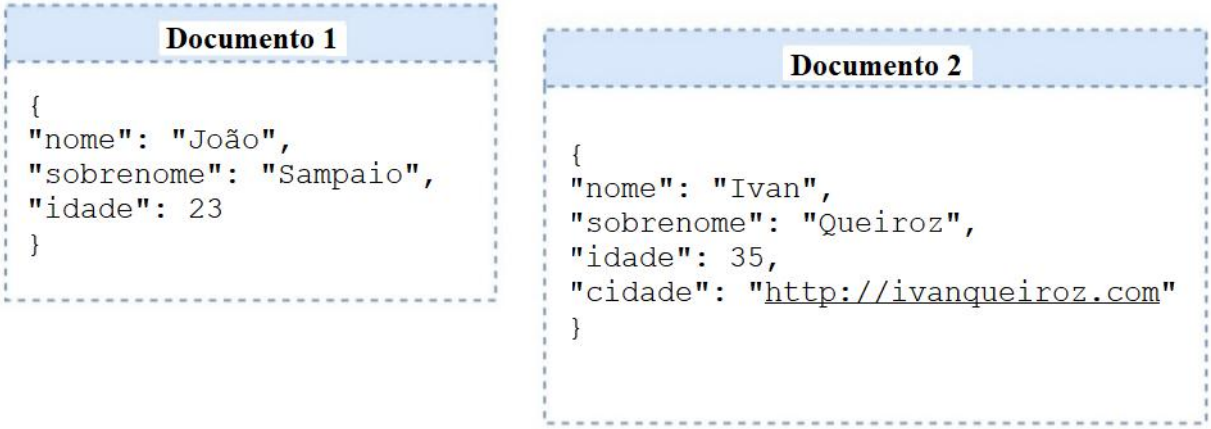
5.4 Orientado a Documentos

O modelo orientado a documento segundo Souza et al (2014) é formado por um conjunto de documentos e em cada documento existe pares de chave e valor, onde cada documento é um objeto que tem um identificador único dentro de toda a coleção. Este modelo tem como objetivo solucionar problemas causados por banco de dados relacionais, pois sua modelagem é mais flexível quanto a mudanças, é possível modificar a estrutura do documento quando achar necessário, sem que haja impactos para a aplicação, com o custo de riscos de inconsistências ou a existência de documentos obsoletos.

Para Diana e Gerosa (2010), esse tipo de banco de dados não possui esquema, tornando-o uma ótima opção para o armazenamento de dados semiestruturados. Esse tipo de modelo usa um formato de armazenamento que suporte um documento embutido em outro, como o JSON. Segundo Amazon Web Services (2018b), o banco de dados orientado a documento é usado em aplicações web, tolerante a dados incompletos e parecido com o modelo chave-valor. Bancos de dados populares nessa categoria são o MongoDB e o CouchDB. A Figura 6 ilustra um exemplo de banco de dados orientado a documento que usa o formato de armazenamento JSON.

O modelo NoSQL para armazenamento de dados usado no projeto, foi o modelo orientado a documento, visto que, os dados fornecidos pelo Twitter se adequam melhor ao formato orientado a documento, e utilizamos o SGBD MongoDB.

Figura 6 – Representação de um SGBD orientado a documentos



Fonte: Queiroz (2017)

6 O MONGODB

O MongoDB é um sistema gerenciador de banco de dados, que segundo w3big.com (2018) foi escrito na linguagem de programação C++, lançado em fevereiro de 2009 pela empresa 10gen, seu desenvolvimento durou quase dois anos, tendo iniciado outubro de 2007. O MongoDB é um banco de dados NoSQL de código aberto, gratuito que possui o sistema de armazenamento de arquivos distribuídos e suporta várias linguagens de programação como por exemplo: Ruby, Python, Java C++ PHP entre outras.

Para armazenar os dados coletados usamos o SGBD não relacional MongoDB, que de acordo com Chodorow (2013) é um banco de dados orientado a documento, que está destinado a ser um banco de propósito geral, flexível, de alta performance e facilmente escalável, que facilita o redimensionamento. Este banco substitui o conceito de "linhas" dos bancos de dados tradicionais por um mais flexível em "documentos". Não existe um esquema predefinido, ou seja, as chaves e valores. Os valores dos campos podem incluir outros documentos, *arrays* e *arrays* de documentos que não possui tamanho e tipo fixo. Como não tem um esquema fixo é possível adicionar ou remover campos quando achar necessário.

Para Lennon (2011) o que diferencia o MongoDB dos outros bancos de dados não relacionais (NoSQL), é a sua linguagem de consulta baseada em documento, que torna a transição de um banco de dado relacional para o MongoDB simples, pois realiza as consultas com muita facilidade. Os dados são armazenados dentro de documentos BSON (uma versão binária do JSON). O MongoDB é um banco de dados simples de usar e instalar, com binários e drives disponíveis para os principais sistemas operacionais e linguagens de programação.

Segundo Boaglio (2015) Crockford identificou uma prática usada pela empresa Netscape para solucionar problemas, como o aumento do tráfico de informações e aumento dos serviços no início do século. Crockford criou uma especificação e chamou de Notação de objetos JavaScript (JSON). A ideia é enviar informações de um servidor para web browser de forma mais simples, não sendo, mas necessário usar nenhum plugin para funcionar. O JSON suporta tipos de dados simples que são o *null*, *boolean*, *number*, *string*, *object* e *arrays.*

Na figura 7 é possível visualizar um exemplo que usa a sintaxe semelhante a JSON.

Figura 7 – Um documento JSON

```
{
"nome": "Maria",
"Idade": "18",
"sexo": "feminino",
"profissão": "dentista"
"grupos": ["esportes", "filmes", "series", "notícias"]
"contatos": {
        "telefone": "88 98888 9999"
        "email": "maria@gmail.com"
        }
}
```

Fonte: O autor (2019)

A Austrian IT Consulting (2019) apresenta um ranking dos bancos de dados mais utilizados, onde o MongoDB é líder no seguimento de bancos não relacional e está entre os cinco bancos de dados mais usados do mundo, sendo o único da categoria NoSQL (Tabela 1). Além de permitir criar, ler, atualizar e excluir dados, o MongoDB ainda nos fornece uma lista de recursos exclusivos cada vez maior, podemos citar por exemplo a agregação, a indexação, os tipos de dados especiais e o armazenamento de arquivo.

Tabela 1 - Os 11 SGBDs mais utilizados

Os 11 SGBDs mais utilizados							
Classificação			**DBMS**	**Modelo de banco de dados**	**Ponto**		
Abr 2019	**Mar 2019**	**Abr 2018**			**Abr 2019**	**Mar 2019**	**Abr 2018**
1	1	1	Oracle	SGBD relacional	1279,94	+0,80	-9,85
2	2	2	MySQL	SGBD relacional	1215,14	+16,89	-11,26
3	3	3	Microsoft SQL Server	SGBD relacional	1059,96	+12,11	-35,55
4	4	4	PostgreSQL	SGBD relacional	478,72	+8,91	+83,25
5	5	5	MongoDB	Loja de documentos	401,98	+0,64	+60,57
6	6	6	IBM DB2	SGBD relacional	176,05	-1,15	-12,89
7	↑8	↑9	Redis	Armazenamento de valor-chave	146,38	+0,25	+16,27
8	↑9	8	Elasticsearch	Motor de pesquisa	146,00	+3,21	+14,64
9	↓7	↓7	Microsoft Access	SGBD relacional	144,65	-1,55	+12,43
10	10	↑11	SQLite	Relacional	124,21	-0,66	+8,23
11	10	↓10	Cassandra	Armazenamento de coluna larga	123,61	+0,81	+4,52

Fonte: Austrian IT Consulting (2019)

De acordo com a documentação do MongoDB (2019) a operação de agregação permite processar registros de dados e retornar resultados computados. O MongoDB também

nos possibilita realizar a construção de agregações complexas de dados otimizando o desempenho. O NoSQL MongoDB nos fornece três tipos de agregações: *aggregation pipeline*, a função *map-reduce* (é implementação para agregação avançada), e os *single purpose aggregation methods*. A indexação suporta índices secundários genéricos, permitindo assim uma variedade de consultas eficientes. Os índices são considerados estruturas de dados especiais, que salvam uma pequena parte do conjunto de dados da coleção em um formato fácil de ler e percorrer.

Segundo Chodorow (2013) o MongoDB nos fornece tipos de dados especiais que suportam coleções do tipo *time-to-live* para dados que expiram em um determinado tempo de execução, como por exemplo as sessões. Este banco de dados também suporta coleções de tamanho fixo, que são úteis para manter dados recentes, pode-se citar como exemplo, os logs. Já o armazenamento de arquivos suporta o de grande volume de dados e metadados de arquivos. Podemos observar no Quadro 1 as vantagens e desvantagens da utilização do banco de dados NoSQL MongoDB.

Quadro 1 - Vantagens e desvantagens do uso do MongoDB

Vantagens	Desvantagens
❖ **Livre de esquemas** – Armazenamento de documentos sem a rigidez de uma padronização dos dados. ❖ **Facilmente escalável –** Base de dados pronta para operar em diversos nós distribuindo a carga de processamento. ❖ **Consistência –** você escolhe o nível de consistência dependendo do valor da informação. Mais rápido, garante a leitura, mais devagar, garante a escrita em todos os nós da rede antes de liberar o recurso**.** ❖ **Arquitetura Nexus –** Esta arquitetura incorpora os pontos fortes dos bancos de dados relacionais, juntamente com as inovações do NoSQL. ❖ **Linguagem de consulta expressiva –** este banco de dados é a única opção NoSQL que oferece uma linguagem de consulta expressiva, consistência forte e índices secundários.	❖ **Menos flexível em pesquisas –** não existe JOINs nativos no MongoDB pois devem ser utilizados na aplicação. ❖ **Informações atualizadas –** dependendo a forma da consistência escolhida a propagação de dados poderá gerar gargalos. ❖ **Alto uso de memória** – o banco usa mais memória por que armazenar nomes de chaves em cada par valores. ❖ **Transações** – o MongoDB não trata as operações como transações. Para alcançar escalabilidade e o desempenho esse banco de dados dispensa suporte de transações.

Fonte: Data Flair (2018) & Fellows (2016)

De posse dos dados coletados e trabalhados, para análise dos dados, utilizamos Redes Bayesianas para classificar os *tweets* e identificar a epidemia. As Redes Bayesianas são um método de modelagem e decisão, que usa raciocínio probabilístico, ou seja, sua metodologia é fundamentada em probabilidade, principalmente em probabilidade condicional (ARA-SOUZA, 2010).

7 O MÉTODO BAYESIANO

De acordo com Pena (2006), o filósofo Richard Price apresentou a Real Sociedade um artigo de Thomas Bayes que encontrou entre os papeis do matemático e reverendo presbiteriano que viveu no século XVIII na Inglaterra, artigo este com o nome 'Ensaio buscando resolver um problema probabilístico', que apresentava a demonstração do teorema de Bayes. Segundo Pena, Price acreditava-se que o artigo de Bayes apresentava prova da existência de Deus.

Para Bachega (2016) a teoria Bayesiana foi desenvolvida levando em consideração que o conhecimento do mundo é imperfeito e que os dados coletados estão saturados de ruídos, que pode nos levar conclusões erradas. Bachega também relata que esta teoria está sendo empregada em várias áreas como a bioestatística, a análise de riscos, a astrofísica, cosmologia entre outras.

"A ideia da Inferência Bayesiana é combinar a informação a priori com a função de verossimilhança. Esta combinação é feita através do "teorema de bayes", originando a distribuição "a posteriori"." (MANCUSO, 2010).

Método Bayesiano segundo Mancuso (2010) é outra possibilidade aos métodos clássicos de inferência, que é usada para estimar a probabilidade de um evento acontecer em determinada circunstância, usando assim uma estimativa *a priori* da probabilidade de sua ocorrência. A estatística bayesiana pode ser descrita de forma ainda mais sucinta, como o processo de diminuição da incerteza sobre o desconhecido que se baseia em dados estatísticos e em evidências previas. O Quadro 2 apresenta uma comparação sucinta entre a Inferência Clássica e a Inferência Bayesiana.

Quadro 2 - Comparação entre Inferência Clássica e Inferência Bayesiana

Inferência Clássica	**Inferência Bayesiana**
❖ O parâmetro desconhecido é considerado uma constante fixa.	❖ O parâmetro desconhecido é considerado uma variável aleatória que segue uma distribuição *a priori.*
❖ É considerada somente a informação amostral (através da verossimilhança).	❖ Podem-se considerar, na *priori*, informações de estudos anteriores, conhecimento pessoal, entre outros.
❖ Não se pode falar em probabilidade para as estimativas dos intervalos de confiança.	❖ Pode-se falar em probabilidade para estimativas dos intervalos de credibilidade.

Fonte: Extraído de Mancuso (2010).

Para Plentz (2003), os métodos Bayesianos permitem mostrar de forma numericamente, o grau de certeza sobre um evento, e manipulação conforme as regras estabelecidas na teoria de probabilidade. “Pois a teoria Bayesiana está fundamentada na teoria da probabilidade, sendo que a diferença básica está no enfoque não frequencista adotado pela teoria de Bayes. Na teoria da probabilidade, dados dois eventos A e B, é possível condicionar A a ocorrência de B” (PLENTZ, 2003).

$$P(A|B) = P\,(A \cap B) / P(B), \text{ se } P(B) > 0$$

Segundo Plentz (2003), a probabilidade de A dado B pode ser interpretada como a atualização da crença em A dado que a evidencia B tornou-se disponível. Em outras palavras, P(A|) representa à probabilidade do evento A (uma hipótese) condicionado à ocorrência de algum evento B (evidência). O autor Plentz ainda relata que o Teorema de Bayes apresentado pode ser reescrito com facilidade para obter a probabilidade posterior (*a posteriori* é um valor de probabilidade que foi visto usando-se informação adicional obtida posteriormente) de uma hipótese A, após a observação de alguma evidência A, dado a probabilidade *a priori* (é um valor de probabilidade inicial original alcançado antes que seja obtida qualquer informação) em A e a verossimilhança da observação B estar em A:

$$P(A|B) = (P(B|A)\,P(A)) / P(B); \text{ se } P(B) > 0$$

De acordo com Ray (2016), o teorema Bayesiano é uma fórmula matemática usada para calcular a probabilidade a posteriori de um evento, baseado em conhecimento a priori. Para entender como funciona o teorema de Bayes, Ray (2016) usou um exemplo que envolve esporte, relacionando a chance de um jogador praticar ou não esporte, baseado no clima ou tempo. Inicialmente foi fornecido um conjunto de dados de treinamento do clima e da correspondente variável-alvo onde precisamos classificar se os jogadores vão jogar ou não, baseado na condição meteorológica (Figura 8).

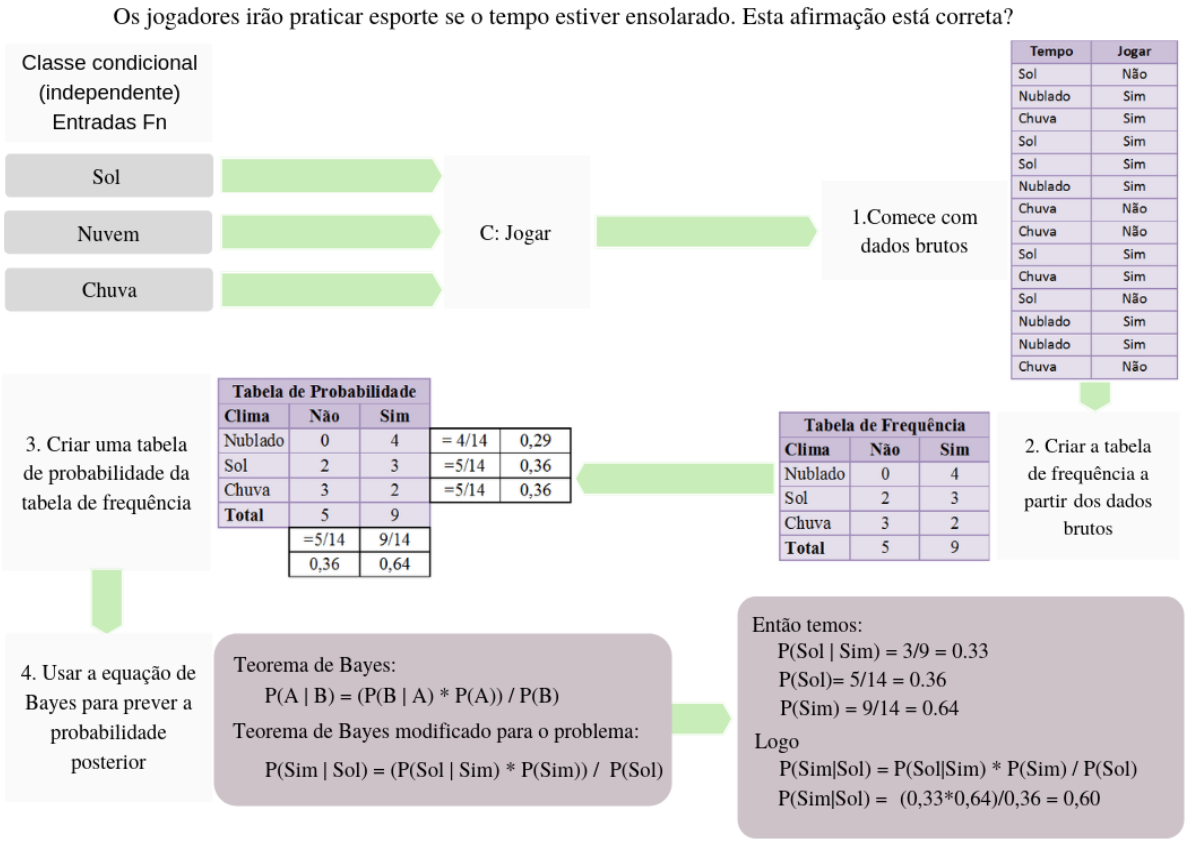
Problema: Os jogadores irão praticar esportes se o tempo estiver ensolarado. Esta afirmação está correta?

A Figura 8 apresenta as etapas para a solução do problema. Nela observar-se a tabela de frequência, que foi criada a partir dos dados brutos. Após criar a tabela de frequência, calculamos todas as probabilidades de cada combinação, com isso foi montado à tabela de probabilidade. Então, foi aplicado o teorema Bayesiano para calcular a probabilidade de os jogadores praticarem esportes se o tempo estiver ensolarado. Após substituirmos os valores

que encontramos na tabela de probabilidade na equação de Bayes, chega assim à chamada probabilidade a posteriori.

Portanto o raciocínio Bayesiano nos levou, de um modo simples, a concluir que a probabilidade *a posteriori* (ou seja, após o teste) dos jogadores praticarem esporte em um dia ensolarado é de 0,60 (60%), então podemos concluir que a afirmação está correta.

Figura 8 – Exemplo envolvendo a pratica de esporte, baseado na condição meteorológica.



Fonte: Adaptado de Glen (2019)

8 METODOLOGIA

8.1 O Uso do Twitter

A rede social Twitter pode ser usada para ver notícias em tempo real, medir as reações aos eventos atuais, encontrar links que tenham relação a tópicos específicos, entre outros, além de ser uma rica fonte de dados sobre os mais diversos assuntos (GRUS, 2016). Também podemos usar essa rede social para analisar tendências relacionadas a uma palavra-chave, analisar o sentimento relacionado a uma determinada marca ou obter feedback sobre assuntos e serviços (DATA SCIENCE ACADEMY, 2018b).

Para coletar dados do Twitter é necessário o uso de uma API (interface de programação de aplicativo), onde a API é um software que permite interação com o computador e serviços webs de forma rápida e simples. Grus (2016) afirma que diversos websites e serviços web disponibilizam **APIs** que possibilitam solicitar os dados em forma de estruturas, poupando assim o trabalho de extraí-los.

Inicialmente foi usado a biblioteca Twitter, que forneceu uma interface Python pura para API da rede social, que segundo Taylor (2017) facilita ainda mais o uso por programadores de Python e é exposta através da classe Twitter que suporta apenas a autenticação oAuth, já que os desenvolvedores do Twitter indicaram que oAuth é o único método que poderá ser suportado no futuro e que segundo Mitchell (2015) é uma das melhores bibliotecas do Twitter disponível para Python 3 (a versão do Python que usamos).

Para interagir com a API do Twitter foi necessário obter as credenciais, e para adquiri-las, foi preciso criar uma conta na rede social e depois acessá-la via site http://apps.twitter.com, fazer o *login* e em seguida foi criado uma nova aplicação que gerou automaticamente as duas primeiras chaves de acesso (*consumer key* e *consumer secret*), então foi criado também o token de acesso que gerou duas chaves de acesso (*Access Token* e *Access Token Secret*), que foram usadas no script Python.

A coleta de dados do Twitter foi realizada entre **13 de março de 2018** e se deu até o dia **7 de julho 2018**. Para isso foi utilizado filtro “febre amarela” para coleta dos dados. Através de comandos da linguagem Python foi possível salvar no banco de dados NoSQL MongoDB apenas *tweets* com coordenadas geográficas, em português e que foram "tuitados" sob o domínio “br”, visto que há vários países que utilizam português.

Foi feito um script em Python que, a cada sessenta segundos, capturasse *tweets*, isso fez com que fosse minerado *tweets* repetidos, dessa forma foi necessária verificar se o “id” do

tweet capturado já havia sido salvo no banco de dados, caso “id” não exista, o *tweet* era salvo no Mongodb.

8.2 Armazenamento dos tweets no MongoDB

O banco de dados NoSQL que usamos para armazenar os *tweets* coletados foi o MongoDB que é um banco de dados orientado a documentos, que substitui o conceito de “linhas” do banco de dados relacional por um modelo mais flexível em “documentos” e livre de esquemas. O MongoDB geralmente não possui estrutura em comum, essa característica é uma boa opção para armazenamento de dados semiestruturados (MEDEIROS, 2014).

Normalmente, um banco de dados orientado a documentos armazena dados nos formatos JSON ou XML. A compatibilidade com documentos no formato JSON torna mais fácil para os desenvolvedores realizar a serialização e o carregamento de objetos contendo propriedades e dados pertinentes. Os bancos de dados NoSQL foram projetados para aumentar a escala horizontal utilizando clusters distribuídos de hardware de baixo custo para aumentar o *throughput* sem aumentar a latência (Amazon Web Services 2018b).

Um objeto no formato JSON é representado por um conjunto de pares nome/valor iniciando e terminando com chaves. Cada nome (ou rótulo) é seguido por dois pontos e os pares nome/valor seguidos por vírgula, como mostra a Figura 9.

Figura 9 – Um objeto JSON - Parte de uma informação recuperada de um *tweet*


```
{
   "_id" : ObjectId("5b3d39e6e535430c7b8bff26"),
   "created_at" : "Tue Mar 13 19:57:57 +0000 2018",
   "id" : NumberLong(973649386491260929),
   "id_str" : "973649386491260929",
   "text" : "Faleceu um parente de uma amiga vítima de febre amarela, a situação tá punk. Quem ainda
não se vacinou, corra para… https://t.co/gm7Zo4vw3Y",
   "geo" : null,
   "lang" : "pt",
   "coordinates" : null,
   "retweeted" : "false",
   "place" : {
      "id" : "c83a3bc35870a7a1",
      "url" : "https://api.twitter.com/1.1/geo/id/c83a3bc35870a7a1.json",
      "place_type" : "city",
      "name" : "Santos",
      "full_name" : "Santos, Brasil",
      "country_code" : "BR",
      "country" : "Brazil",
      "contained_within" : [],
      "bounding_box" : {
         "type" : "Polygon",
         "coordinates" : [
            [
               [
                  -46.403227,
                  -23.99197
               ],
               [
                  -46.176966,
                  -23.99197
               ],
               [
                  -46.176966,
                  -23.73467
               ],
               [
                  -46.403227,
                  -23.73467
               ]
            ]
         ]
      },
      "attributes" : {}
      },
   }
```

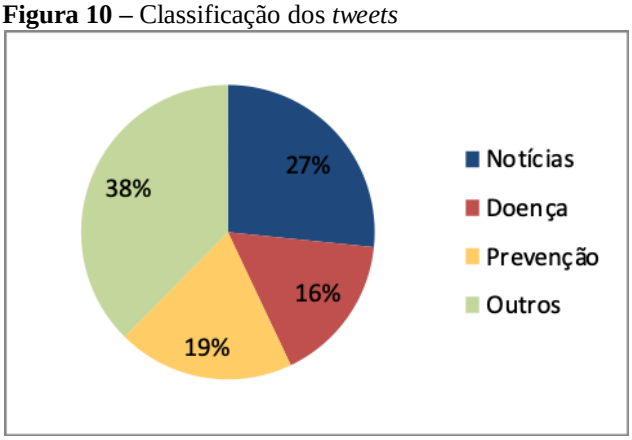

Fonte: O autor (2019)

8.3 Classificações dos tweets

Para formação do conjunto de treinamento do classificador, foram coletadas mensagens publicadas de usuários (*tweets*) postados no Brasil. Foram etiquetados manualmente um conjunto de **4.000** *tweets*. O período de coleta dos *tweets* destinado ao treinamento teve início 13 de março de 2018 até 7 julho de 2018.

Os *tweets* foram classificados em 4 grupos: **doença, prevenção, notícia e outros**, como visto no gráfico (Figura 10).

Figura 10 – Classificação dos *tweets*



Fonte: O autor (2019)

O primeiro grupo foi composto por **653** (corresponde a 16% dos *tweets* etiquetados) *tweets* rotulados como **doença**, foram consideradas mensagens que se tratavam de casos confirmados, sintomas e morte por febre amarela. Dentre os *tweets* selecionados pode-se citar como exemplo os seguintes: “Fui ver meu amigo que estava com febre amarela e mano ele superou metade já!” e “Peruíbe registra primeira morte por febre amarela na cidade”.

O segundo grupo classificado foi o de **prevenção**, neste caso foi levado em consideração *tweets* que falavam a respeito de vacinação, e cuidados tomados como prevenção, tais como evitar locais com alta incidência de febre amarela caso não se tenha tomado a vacina e outras formas de prevenção. Dos 4000 *tweets*, **779** (19% dos *tweets* rotuladas) foram rotulados como prevenção, dentre esses *tweet*s temos como exemplo: “Tomei a vacina da febre amarela” e “Minha mãe tomou vacina de febre amarela”.

O terceiro grupo é o de **notícias**, esse grupo foi composto por **1.065** (27% dos *tweets* coletados) mensagens, como por exemplo: “Vigilância em Saúde alerta população para urgência da vacinação” e “Macaco encontrado morto em Corupá não tinha febre amarela”.

O quarto grupo catalogado como **outros** foi composto por **1.503** *tweets*, que equivalente a 38% das mensagens rotuladas, está incluído neste grupo mensagens do tipo: “Estou com febre de amar ela” e “Aliança pra mim é igual a febre amarela, PASSA LONGE”.

Foi utilizado o **Método Bayesiano** para **classificar** os *tweets* com coordenadas (**781** *tweets*). Para realizar a classificação dos *tweets* com coordenadas foi usado os *tweets* que foram catalogados manualmente (**4.000** *tweets*) como evidência para classificar os *tweets* com coordenadas (**781** *tweets*).

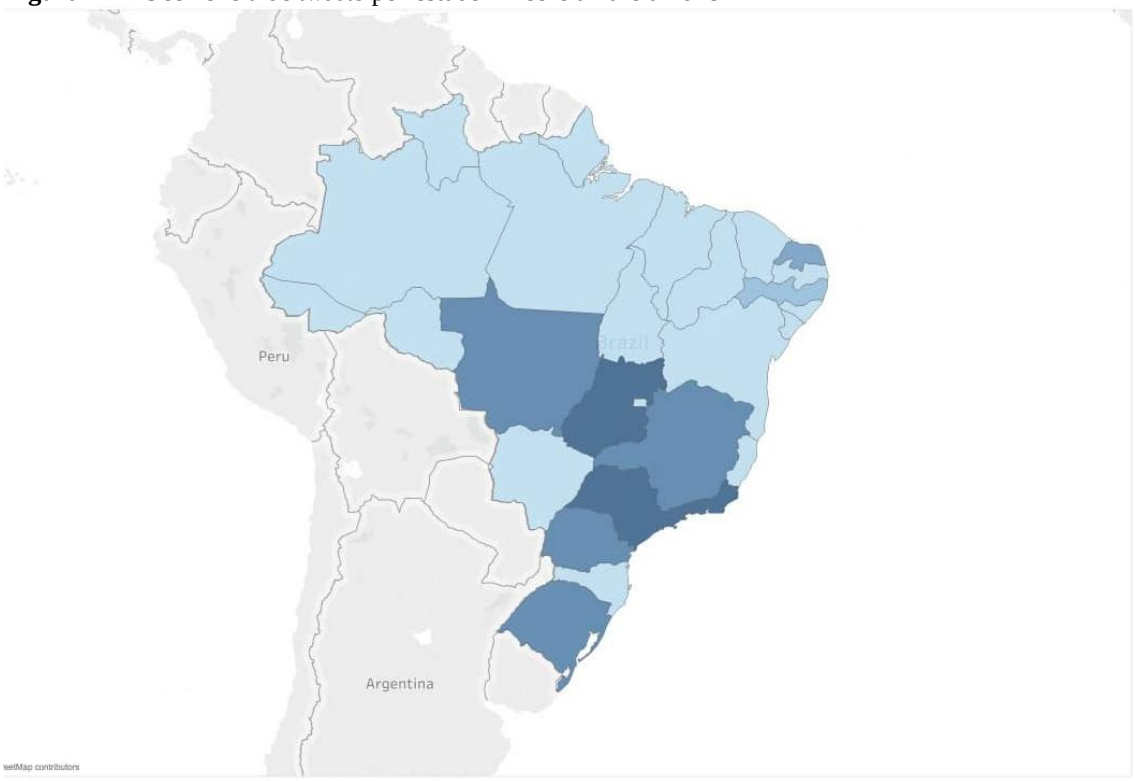
Foi utilizado validação cruzada (*Cross Validation*) para analisar e ajudar a desenvolver o modelo dos dados minerados durante a pesquisa. A validação cruzada foi usada depois de criarmos a estrutura de mineração e o modelo. Por fim foi calculado a acurácia, que é a proporção de predições corretas (soma de positivos verdadeiros e negativos verdadeiros). A matriz de confusão também conhecida como tabela de contingencia foi usada para validar o aprendizado supervisionado, comparando base de teste com o treinamento. Este tipo de tabela permite a visualização do desempenho do algoritmo de aprendizado utilizado (PRINA; TRENTIN, 2015).

9 RESULTADOS

Os resultados foram obtidos através da mineração de dados da rede social Twitter e do Ministério da Saúde através de boletins epidemiológicos, entre **13 de março de 2018 e 07 de julho de 2018.**

Verificou-se a ocorrência dos *tweets* por estado no mapa da Figura 11, vale salientar que as partes mais escuras são as que possuem maior ocorrência de *tweets* relacionado à febre amarela. Esse mapa (Figura 11) nos permite observar que os estados com mais mensagens relacionando a doença, foram tuitadas pelos estados do Rio de Janeiro e São Paulo (cor azul mais escuro), estados estes que estão entre os mais afetados pela febre amarela segundo o boletim do Ministério da saúde (BRASIL, 2018a, 2018b).

Figura 11 – Ocorrência de *tweets* por estado – Febre amarela 2018



Fonte: O autor (2019)

No mapa da Figura 12, é possível observar a ocorrência dos *tweets* por cidade, assim como no mapa de ocorrências por estado (Figura 11), os pontos mais escuros correspondem a um maior número de *tweets* relacionado à febre amarela no Brasil.

Figura 12 – Ocorrência de *tweets* por cidade – Febre amarela 2018

Fonte: O autor (2019)

Depois de minerarmos os *tweets* e classificarmos manualmente, criamos um modelo, realizarmos a validação cruzada, e medimos acurácia média do modelo resultando assim em um valor de **0.934.**

Como resultado da análise de correlação entre as variáveis, pode-se perceber na Tabela 2, que houve uma correlação de **0.8142** entre os estados analisados, com destaque os estados do Rio de Janeiro e São Paulo que segundo o boletim epidemiológico fornecido por Brasil (2018a, 2018b), tiveram o maior número de casos de Febre Amarela. O total de casos confirmados nesses estados é mais da metade das ocorrências da doença registradas no restante do país.

Tabela 2 - Resultados da análise de correlação entre os dados da SVS e do Twitter por UF

Resultados da análise de correlação entre os dados da SVS e do Twitter por UF		
	Total de Casos	
Unidade Federal	**SVS**	**Twitter**
Rio de Janeiro	152	255
São Paulo	179	137
Goiás	0	12
Rio Grande do Sul	0	10
Paraná	0	7
Minas Gerais	117	5
Mato Grosso	0	4
Rio Grande do Norte	0	4
Pernambuco	0	3
Pará	0	3
Santa Catarina	0	3
Paraíba	0	2
Espirito Santo	1	2
Correlação	**0.81429**	

Fonte: O autor (2019)

A Tabela 3 apresenta o resultado da análise de correlação dos casos de febre amarela no Brasil, obtidos a partir da análise de correlação dos dados da Secretaria de Vigilância em Saúde com os dados de *tweets* por município, percebe-se que houve uma correlação de 0,7788 nos municípios analisados.

Tabela 3 - Resultados da análise de correlação entre os dados da SVS e do Twitter por município

Resultados da análise de correlação entre os dados da SVS e do Twitter por município		
	Total de Casos	
Unidade Federal	**SVS**	**Twitter**
Guarulhos SP	15	5
Itatiaia RJ	5	4
Itatiaia RJ	5	4
São Paulo SP	5	3
São Jose dos Campos SP	3	3
Embu SP	1	3
São Lourenco da Serra SP	0	2
Santa Isabel SP	6	2
Juquitiba SP	3	2
São Sebastiao SP	3	2
Rio Claro RJ	4	1
São Sebastiao SP	3	1
Rio Piracicaba MG	0	1
Salto SP	0	1
Pirapora do Bom Jesus SP	0	1
Maua SP	0	1
Jarinu SP	0	1
Nova Granada SP	0	1
Presidente Venceslau SP	0	1
Salto de Pirapora SP	1	1
Caieiras SP	2	1
Paty do Alferes RJ	2	1
Jarinu SP	1	1
Itaquaquecetuba SP	0	1
São Fidelis RJ	0	1
Peruíbe SP	0	1
Correlação	**0.77889**	

Fonte: O autor (2019)

10 CONCLUSÃO

Através de técnicas de mineração de texto e de Classificação Bayesiana, este projeto teve como objetivo identificar e analisar os focos de febre Amarela no Brasil por meio de uma rede social, que no caso deste trabalho foi o Twitter, no período de **13 de março de 2018 e 07 de julho de 2018.** Para isso, foram verificadas as mensagens dos usuários da rede social Twitter que relatassem sintomas da doença ou ainda mensagens com alguma relação ao dado pesquisado.

O modelo criado para identificar casos de febre amarela através do Twitter, teve uma taxa de acurácia de 93% de acertos no *dataset* de treino/teste, as correlações dos dados coletados do **Twitter** com os dados oficiais do **Ministério da Saúde** foram de 81% e 78% para os estados e as cidades respectivamente, mostrando que esta rede social é um bom indicador para identificação de epidemias.

Atualmente, o Twitter é uma rede social que produz um grande volume de informações, de acordo com os dados fornecidos por Alencar e Almeida (2016) cerca de 783 mil novos *tweets* foram postadas todos os dias no Brasil, isso sem contar com os *tweets* sem informação geográfica.

Ao serem identificados e analisados os dados, foi possível correlacionar os dados colhidos e verificados na rede social com as informações dos meios oficiais de notificação. Isso foi constatado que os estados com mais mensagens relacionando a doença, foram tuitadas pelos estados do Rio de Janeiro e São Paulo, estados estes que estão entre os mais afetados pela febre amarela segundo o boletim oficial do Ministério da Saúde do ano de 2018. De acordo com os resultados obtidos na análise de correlação, foi possível concluir que as informações contidas na rede social do Twitter também podem ser usadas como fonte de dados para análise e previsão de epidemias.

Tendo em consideração os dados obtidos do Twitter, percebe-se um meio bastante rico em informações, o que torna apto a serem utilizado como fonte de pesquisa para diversas áreas, inclusive da saúde, mas é importante levar em consideração também a variação na popularidade da rede social, o que pode resultar também na facilidade ou não de se captar os dados necessários para futuras análises.

REFERÊNCIAS

ALENCAR, Vladimir; ALMEIDA, H. N. **Previsão de Epidemias através de Mídias Sociais**. 1. ed. Berlim, Alemanha: NEA, 2016.

ANDRADE, P. H. M. A. de. **Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos:** um Estudo da Automatização da Triagem de Denúncias na CGU. 2015, 54f. Dissertação (Mestrado Profissional em Computação Aplicada) – Universidade de Brasília, Brasília.

ANICETO, R. C.; XAVIER, R. F. **Um Estudo Sobre a Utilização do Banco de Dados NoSQL Cassandra em Dados Biológicos**. 2014, 50f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade de Brasília, Brasília.

ANKAM, Venkat. **Big Data Analytics**. Editora: Packt Publishing. Birmingham, Reino Unido, 2016.

ARANHA, Christian; PASSOS, Emmanuel. A Tecnologia de Mineração de Textos**. RESI-Revista Eletrônica de Sistemas de Informação**, v. 5, n. 2, p. 1-8, 2006. Disponível em: <http://periodicosibepes.org.br/index.php/reinfo/article/view/171/66>. Acesso em: 17 mar. 2019.

ARA-SOUZA, **A. Luiz. Bayesianas:** Uma Introdução Aplicada a Credit Scoring. In: 19º SINAPE: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 2010, São Pedro - SP. 19º SINAPE - Simpósio Nacional de Probabilidade e Estatística.

AUSTRIAN IT CONSULTING. **DB-Engines Ranking.** [S.l.:S.n], 2019. Disponível em: < https://db-engines.com/en/ranking>. Acesso em: 04 abr. 2019.

AWS. Amazon Web Services. **O que é o NoSQL?** [S.l.:S.n], 2018a. Disponível em: <https://aws.amazon.com/pt/nosql/>. Acesso em: 12 jul. 2018.

AWS. Amazon Web Services. **O que é um banco de dados de documentos?** [S.l.:S.n], 2018b**.** Disponível em: <https://aws.amazon.com/pt/nosql/document/>. Acesso em: 12 jul. 2018.

BACHEGA, Riis Rhavia Assis. **O que é a regra Bayes?** [S.l.]**,** 2016. Disponível em: < https://universoracionalista.org/o-que-e-a-regra-de-bayes/ >. Acesso em 1 out. 2018.

BOAGLIO, Fernando. **MongoDB**: Construa novas aplicações com novas tecnologias. São Paulo: Casa do Código, 2015.

BRASIL. Ministério Da Saúde**. Monitoramento do Período Sazonal da Febre Amarela Brasil – 2017/2018.** Informe n. 17, 2018a. Disponível em: <http://portalarquivos2.saude.gov.br/images/pdf/2018/marco/14/Informe-FA-17.pdf>. Acesso em: 9 jul. 2018.

BRASIL. Ministério Da Saúde. **Monitoramento do Período Sazonal da Febre Amarela Brasil – 2017/2018.** Informe n. 26, 2018b. Disponível em: <

http://portalarquivos2.saude.gov.br/images/pdf/2018/maio/18/Informe-FA-26.pdf>. Acesso em: 9 jul. 2018.

BRASIL. Ministério Da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. **DOENÇAS INFECCIOSAS E PARASITÁRIAS:** GUIA DE BOLSO. 8 ed. Brasília: Ministério da saúde, 2010.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Informe Especial Febre Amarela No Brasil Nº 01/2017**– DEVIT/SVS/MS. Brasília, 2017. Disponível em < http://portalarquivos.saude.gov.br/images/pdf/2017/marco/18/Informe-especial-COES-FA.pdf>. Acesso em: 03 ago. 2018.

CAMARGO, E. P. Doenças tropicais. **Estudos Avançados.** São Paulo, v. 22, n. 64, p. 95-110, Dc. 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142008000300007&lng=en&nrm=iso>. Acesso em: 19 dez. 2018.

CHODOROW, Kristina. **MongoDB:** The Definitive Guide. 2. ed. Sebastopol: O'Reilly Media 2013.

DATA FLAIR. [S.l.:S.n], 2018. Disponível em: <https://data-flair.training/blogs/advantages-of-mongodb/>. Acesso em: 20 set. 2018.

DAVENPORT T. H.; BARTH, P.; BEAN, R. How 'Big Data' Is Different. MIT Sloan. **MIT Sloan Management Review.** v. 54, n. 1, p. 21-24, 2012. Disponível em: <http://newvantage.com/wp-content/uploads/2012/10/MIT_Sloan_Review_How-Big-Data-Is-Different_Fall2012.pdf>. Acesso em: 05 set. 2018.

DIANA, Mauricio De; GEROSA, Marco Aurélio. **NOSQL na Web2.0: Um Estudo Comparativo de Bancos Não-Relacionais para Armazenamento de Dados naWeb2.0**. São Paulo, 2010. Disponível em: <https://www.ime.usp.br/~mdediana/nosql_wtdbd10.pdf>. Acesso em: 07 out. 2018.

DSA. DATA SCIENCE ACADEMY. **Big Data Fundamentos.** [S.l.:S.n], 2018a. E-book**.** Disponível em: <https://www.datascienceacademy.com.br/path-player?courseid=python-fundamentos&unit=56fa01d947d7ddf1938b456cUnit>. Acesso em: 04 dez. 2018.

DSA. DATA SCIENCE ACADEMY. **Python Fundamentos para Análise de Dados**. [S.l.:S.n], 2018b. E-book. Disponível em: <https://www.datascienceacademy.com.br/path-player?courseid=python-fundamentos&unit=56fa01d947d7ddf1938b456cUnit>. Acesso em: 04 nov. 2018.

FARO, Alberto; GIORDANO, Daniela; SPAMPINATO, Concetto. Combining literature text mining with microarray data: advances for system biology modeling. **Briefings in Bioinformatics**, v. 13, n. 1, p. 61-82, 2012. Disponível em: <https://academic.oup.com/bib/article/13/1/61/219461>. Acesso em: 02 abr. 2019.

FELLOWS, André. **Conhecendo o MongoDB**. [S.l.] 2016. Disponível em: <http://fellowsdevel.com/conhecendo-o-mongodb/>. Acesso em 20 set. 2018.

FREITAS, J. A. S. **Uso de Técnicas de Data Mining para Análise de Bases de Dados Hospitalares com Finalidades de Gestão**. 2006. 269f. Tese (Doutorado em Ciências Empresariais), Faculdade de Economia da Universidade do Porto. Porto.

GLEN, Stephanie. **Naive Bayes in One Picture**. [S.l.], 2019. Disponível em: <https://www.datasciencecentral.com/profiles/blogs/naive-bayes-in-one-picture-1>. Acesso em 20 mai. 2019.

GONÇALVES, Eduardo Corrêa. Mineração de Texto Conceitos e Aplicações Práticas. **Revista SQL Magazine**, v. 105, p. 31-44, 2012. Disponível em: <https://docit.tips/download/figura-12-stems-para-obter-o-stem-de-uma-palavra-a-preciso-utilizar_pdf>. Acesso em: 03 abr. 2019.

GRUS, Joel. **Data Science do Zero**. Tradução Welington Nascimento. Rio de Janeiro: Editora Alta Books, 2016.

HASHEM, I. A. T. et al. The rise of "big data" on cloud computing: Review and open research issues. **Information Systems** v. 47, p. 98-115, 2015.

LENNON, Joe. **Explore o MongoDB:** Saiba por que esse sistema de gerenciamento de bancos de dados é tão popular. [S.l.], 2011. Disponível em: <https://www.ibm.com/developerworks/br/library/os-mongodb4/index.html>. Acesso em: 8 nov. 2018.

LOPES, J. M. **Um Estudo Comparativo Entre Bancos De Dados Considerando As Abordagens Relacional E Orientada A Grafo.** 2014. 90f. Dissertação (Tecnologias da Informação e Comunicação) - Universidade Federal de Santa Catarina, Araranguá.

LOSHIN, David. **Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph**. Editora Morgan Kaufmann. Burlington, Massachusetts, USA. 2013.

MADEIRA, R. O. de C. **Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais.** 2015. 66f. Dissertação (Mestrado em Modelagem Matemática da Informação), Fundação Getúlio Vargas, Escola de Matemática Aplicada, Rio de Janeiro.

MANCUSO, Aline Castello Branco. **Métodos Bayesianos em Metanálise**. 2010, 82f. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do sul, Porto Alegre.

MEDEIROS, H. **Introdução ao MongoDB.** [S.l.], 2014. Disponível em:< https://www.devmedia.com.br/introducao-ao-mongodb/30792>. Acesso em: 9 jul. 2018.

MICREIROS.COM. **Tipos de bancos de dados NoSQL**. [S.l.:S.n] 2017. Disponível em: < http://micreiros.com/tipos-de-bancos-de-dados-nosql/ >. Acesso em: 30 nov. 2018.

MINER, G.; Delen, D.; Elder, J. et al. **Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications**. 1ª. ed. Waltham, MA: Academic Press, 2012.

MITCHELL, Ryan. **Web Scraping com Python**. Tradução: Aldir José Coelho Corrêia da Silva. São Paulo: Novatec Editora Ltda, 2015.

MONGODB. Documentação do MongoDB. [S.l.:S.n]. Disponível em: <https://docs.mongodb.com/>. Acesso em: 23 jan. 2019.

MORAIS, E. A. M. **Contextualização de Documentos em Domínios Representados por Ontologias Utilizando Mineração de Textos.** 2007, 113f. Dissertação (Mestrado em Sistema de informação), Universidade Federal de Goiás, Goiânia.

OMS. Organização Mundial da Saúde. **Integrating neglected tropical diseases into global health and development: fourth WHO report on neglected tropical diseases**. Geneva: World Health Organization; [S.l.:S.n], 2017. Disponível em: < http://apps.who.int/iris/bitstream/10665/255011/1/9789241565448-eng.pdf?ua=1>. Acesso em: 19 dez. 2018.

PENA, Sérgio Danilo. Bayes: o 'cara'! **Revista Ciência Hoje**, v. 38, n. 228, p. 21-29, 2006. Disponível em: <http://dreyfus.ib.usp.br/bio5706/penabayes.pdf>. Acesso em 5 dez. 2018.

PLENTZ, Rafael Dobrachinsky. **Redes Bayesianas Para Análise De Comportamento Aplicada A Telefonia Celular**. 2003 57f. Dissertação (Mestrado em Ciências da Computação Área de Concentração Sistemas de Computação) – Universidade Federal de Santa Catarina, Florianópolis.

PRESSE, France**. Lucro do Twitter mais do que triplica no 1º trimestre.** [S.l.], 2019. Disponível em**:** <https://g1.globo.com/economia/tecnologia/noticia/2019/04/23/lucro-do-twitter-mais-do-que-triplica-no-1o-trimestre.ghtml>. Acesso em: 01 mai. 2019**.**

PRIMO**,** Alex. A cobertura e o debate público sobre os casos Madeleine e Isabella: encadeamento midiático de blogs, Twitter e mídia massiva. **Revista Galáxia**, São Paulo, n. 16, p. 43-59, 2008**.**

PRINA, Bruno Zucuni; TRENTIN, Romario. **GMC**: Geração de Matriz de Confusão a partir de uma classificação digital de imagem do ArcGIS®. Santa Maria, 2015. Disponível em: <http://www.dsr.inpe.br/sbsr2015/files/p0031.pdf >. Acesso em: 29 abr. 2019.

QUEIROZ, Ivan. **O que devo saber sobre NoSQL?** 2017. Disponível em <http://blog.ivanqueiroz.com/2017/01/o-que-devo-saber-sobre-nosql.html>. Acesso em: 30 ago. 2018.

RAY, Sunil. **Fundamentos dos Algoritmos de Machine Learning (com código Python e R).** [S.n] 2016. Disponível em:<https://www.vooo.pro/insights/fundamentos-dos-algoritmos-de-machine-learning-com-codigo-python-e-r/**>.** Acesso em: 19 ago. 2018.

RECUERO, Raquel; ZAGO, Gabriela**.** Em busca das "redes que importam": redes sociais e capital social no Twitter. **Revista Líbero**, São Paulo – v. 12, n. 24, p. 81-94, 2009. Disponível em: <http://seer.casperlibero.edu.br/index.php/libero/article/view/498/472>. Acesso em: 10 set. 2018.

SBI. Sociedade Brasileira de infectologia. Associação Médica Brasileira. **FEBRE AMARELA - INFORMATIVO PARA PROFISSIONAIS DE SAÚDE.** São Paulo, 2017. Disponível em: <https://www.infectologia.org.br/admin/zcloud/125/2017/02/FA_-_Profissionais_13fev.pdf>. Acesso em: 08 ago. 2018.

SILVA, E.M. **Descoberta de Conhecimento com o uso de Text Mining:** Cruzando o Abismo de Moore. 2002. 174 f. Dissertação (Mestrado em Gestão do Conhecimento e da Tecnologia da Informação), Universidade Católica de Brasília. Brasília.

SILVA, F. A. B. da. **Big Data e Nuvens Computacionais: Aplicações em Saúde Pública e Genômica.** Jornal Health Informatics. São Paulo – v. 8, n. 2, p. 73-79, 2016. Disponível em: <http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/336/0>. Acesso em: 03 set. 2018.

SOUSA, Gonçalo da Cruz Pereira e. **Document-Based Databases: Estudo Comparativo no Âmbito das Bases de Dados NoSql**. 2015, 111f. Dissertação (Mestrado Integrado em Engenharia e Gestão de Sistemas de informação) - Universidade do Minho Escola de Engenharia, Braga, Portugal.

SOUZA, Alexandre Morais de; PRADO, Edmir P. V.; FANTINATO, Violeta Sun1 Marcelo. **Critérios para Seleção de SGBD NoSQL: o Ponto de Vista de Especialistas com base na Literatura**. 2014. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2014/0012.pdf>. Acesso em: 24 set. 2018.

TAURION, Cezar. **Big Data**. 1ª ed. São Paulo. Editora: Brasport Livros e Multimídia Ltda., 2013.

TAYLOR, Mike. **Python Twitter**. [S.n], 2017. Disponível em: < https://github.com/bear/python-twitter/wiki >. Acesso em: 02 nov. 2018.

TOTH, Renato Molina. **Abordagem NoSQL – uma real alternativa**. [S.n], 2011. Disponível em: < https://dcomp.sor.ufscar.br/verdi/topicosCloud/nosql_artigo.pdf>. Acesso em: 20 mar. 2019.

TWITTER. **Apresentação do Twitter para Empresas.** [S.l:S.n], 2019a. Disponível em: <https://business.twitter.com/pt/basics/intro-twitter-for-business.html>. Acesso em: 01 mar. 2019.

TWITTER. **Segmentação por dispositivo.** [S.l.:S.n], 2019c. Disponível em:< https://business.twitter.com/pt/targeting/device-targeting.html>. Acesso em: 17 mar. 2019.

TWITTER. **Sobre as APIs do Twitter.** [S.l.:S.n], 2019b. Disponivel em: <https://help.twitter.com/pt/rules-and-policies/twitter-api>. Acesso em: 15 mar. 2019.

UN GLOBAL PULSE. **Understanding Immunisation Awareness and Sentiment Through Social and Mainstream Media**. Global Pulse Project Series, n.19, 2015.

VARDANYAN, M. **Escolhendo a ferramenta certa para o banco de dados NoSQL**. [S.l.], 2011. Disponível em: <https://www.monitis.com/blog/picking-the-right-nosql-database-tool/>. Acesso em: 27 ago. 2018.

VINHA, Felipe. **Faça o download do Twitter e comunique-se em 280 caracteres de qualquer lugar**. [S.l.], 2017. Disponível em: <https://www.techtudo.com.br/tudo-sobre/twitter.html>. Acesso em: 17 dez. 2018.

W3BIG.COM. **MongoDB curso**. [S.l.:S.n]. Disponível em: <http://www.w3big.com/pt/mongodb/mongodb-intro.html>. Acesso em 8 set. 2018.