UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANA CAROLINA BRAS COSTA

Análise de sentimentos em nível de sentença a partir de dados extraídos do Twitter utilizando o *framework* Apache Ignite

São Luís
2017

ANA CAROLINA BRAS COSTA

Análise de sentimentos em nível de sentença a partir de dados extraídos do Twitter utilizando o *framework* Apache Ignite

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, **como parte dos requisitos necessários** para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Prof.ª Dr. Simara Vieira da Rocha

São Luís

2017

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Costa, Ana Carolina Bras.
Análise de sentimentos em nível de sentença a partir de dados extraídos do Twitter utilizando o framework Apache Ignite / Ana Carolina Bras Costa. - 2017.
62 f.

Orientador(a): Simara Vieira da Rocha.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís, 2017.

1. Análise de sentimentos. 2. Apache Ignite. 3. Big Data. 4. MapReduce. 5. Twitter. I. Rocha, Simara Vieira da. II. Título.

ANA CAROLINA BRAS COSTA

Análise de sentimentos em nível de sentença a partir de dados extraídos do Twitter utilizando o *framework* Apache Ignite

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, **como parte dos requisitos necessários** para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em: 14 / 07 / 17

BANCA EXAMINADORA

Prof.ª Simara Vieira da Rocha, Dra.

(Orientadora)

Prof. Geraldo Braz Júnior, Dr.

(Membro da Banca Examinadora)

Prof. Carlos Eduardo Portela Serra de Castro, M.Sc.

(Membro da Banca Examinadora)

Dedico essa monografia a meus avós, Raimundo e Raimunda Costa, Célio (in memoriam) e Maria Aparecida Bras. Essa conquista também é de vocês.

AGRADECIMENTOS

À minha família, pelo amor incondicional oferecido, que mesmo eu estando longe de vocês, sempre pude contar com o apoio e carinho.

Aos meus pais, Osvaldo e Meire, por todo sacrifício, paciência e confiança depositados em mim durante todos esses anos. Me considero uma pessoa de sorte por ser fruto de duas pessoas maravilhosas, honestas, trabalhadoras e que sempre e, incondicionalmente, torceram pelo meu sucesso. Amo vocês.

À minha orientadora, Prof.ª Dra. Simara, pela paciência, competência, dedicação e conhecimento compartilhado comigo.

Aos meus amigos, principalmente os da turma 2010.1, por sempre se fazerem presentes e me apoiarem nas horas mais difíceis. Foi um privilégio estudar com vocês.

Ao meu namorado Jonatas, pelas cobranças, ajuda e apoio incondicional durante a reta final do curso que foram imprescindíveis para a realização desse trabalho.

Aos professores que me marcaram durante meu período nesta instituição, a quem eu tenho um grande estimo, Zair Abdelouahab (in memoriam), Carlos Eduardo Portela e Elivaldo Macêdo. Eu sou imensamente grata por toda amizade e conhecimento compartilhados comigo.

"You must give up the life you planned in order to have the life that is waiting for you."

Joseph Campbell

RESUMO

O exponencial crescimento dos dados gerados por usuários de mídias digitais deu origem ao conceito abstrato de *Big Data*. Nos últimos anos, as redes sociais foram as principais responsáveis por esse fenômeno. O Twitter é uma delas, ele funciona como um *microblogging* onde as pessoas podem expressar suas opiniões no limite de 140 caracteres. Diversas organizações e empresas começaram a prestar atenção nas informações e *insights* que esse tipo de conteúdo pode conter. Através da análise de sentimentos é possível analisar os sentimentos contidos em opiniões expressadas de forma textual. Por exemplo, é possível classificar, como positivo ou negativo, o que os consumidores de uma determinada marca estão comentando sobre ela nas redes sociais. Este trabalho propõe uma metodologia para análise de sentimentos de *tweets* em nível de sentença utilizando o *framework* Apache Ignite. Para isso, será realizado um estudo de caso sobre o relançamento do Super Nintendo, demonstrando o uso de processamento paralelo e do modelo de programação MapReduce na tarefa de análise de sentimentos. Os resultados da classificação dos *tweets* demostraram a eficácia da metodologia proposta.

Palavras-Chave: *Big Data*. Análise de Sentimentos. Twitter. Apache Ignite. MapReduce.

ABSTRACT

The exponential growth of data generated by digital media users created the abstract concept of Big Data. Recently, social networks were the main responsible for this phenomenon. Twitter is one of them, it works like a microblogging where people can express their opinions at the limit of 140 characters. Several organizations and companies have begun to pay attention to the information and insights that this type of content may contain. Through the sentiment analysis it is possible to analyze the feelings contained in opinions expressed in textual form. For example, it is possible to classify, as positive or negative, what customers of a particular brand are commenting about it on social networks. This work proposes a methodology for sentiment analysis in sentence-level of tweets using the Apache Ignite framework. For this, a case study will be carried out on the relaunch of Super Nintendo, demonstrating the use of parallel processing and the MapReduce programming model in the task of sentiment analysis. The results of the tweets classification demonstrated the effectiveness of the proposed methodology.

Keywords: Big Data. Sentiment Analysis. Twitter. Apache Ignite. MapReduce.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

A popularização da internet fez com que os *blogs*, fóruns, grupos de discussões e, principalmente as redes sociais *online*, se tornassem veículos de comunicação onde as pessoas cada vez mais compartilham suas opiniões, crenças, sentimentos e experiências. No primeiro trimestre de 2016, estimou-se que a rede social Facebook atingiu a marca de 1,65 bilhões de usuários ativos por mês (G1, 2016). Dessa forma, o conteúdo produzido por esses usuários gera um grande volume de dados que, segundo Cisco Visual Networking Index (2017), contribuem diretamente para a estimativa de 1,2 zettabyte[1] gerados pela internet só no ano de 2016. Sendo assim, temos à disposição uma imensa quantidade de dados de opinião em formato digital (CHEN et al., 2014), logo se faz necessário o desenvolvimento de novos métodos e técnicas voltadas para esse conjunto que é denominado *Big Data*.

A análise de sentimentos ou mineração de opinião surgiu com o intuito de extrair de opiniões, através do processamento de linguagem natural, informações úteis que podem ser aplicadas em vários setores. Para Liu (2012), análise de sentimentos é definida como uma área de pesquisa que investiga as opiniões das pessoas para diferentes matérias: produtos, eventos, organizações. O autor ainda destaca que sistemas de análise de sentimentos encontraram aplicações em quase todos os negócios e também no domínio social. Dessa forma, as grandes empresas passam a utilizar as opiniões extraídas para tomadas de decisão (SILVA JUNIOR, 2015).

O principal objetivo da análise de sentimentos é identificar sentimentos e emoções contidos em um determinado texto (GRAÇA NETO, 2016). Essa análise eleva seu grau de dificuldade de acordo com a fonte e com o número de palavras nela contida. A má-formatação do texto se torna uma barreira (SOUSA, 2012), devido a isso, muitos trabalhos utilizam a rede social Twitter, onde o usuário pode publicar apenas 140 caracteres por vez, o que torna mais fácil a manipulação e formatação. No trabalho de Santos (2016) é feita uma análise baseada em *tweets*[2] de língua inglesa, sobre como as pessoas se sentiam em relação a *Black Friday*. O estudo cruzou a quantidade de vezes que determinada marca apareceu com avaliações negativas dos usuários. Dados como esses, podem ser muito pertinentes, com aplicação, por exemplo, na estratégia de *marketing* de uma empresa.

[1] Equivalente a um sextilhão de bytes.
[2] Nome utilizado para designar as publicações feitas na rede social do Twitter.

A versatilidade no uso de análise de sentimentos possibilita a aplicação desta para áreas além do comércio, como por exemplo, na política. Os autores França e Oliveira (2014) analisaram o grau de aprovação da população brasileira quanto aos protestos que ocorreram no Brasil entre junho e agosto de 2013. O estudo feito com *tweets* e constatou aprovação positiva, o que foi validado por notícias vinculadas nos meios de comunicação. Na área da saúde, Araújo et al. (2012) questiona o porquê de poucos trabalhos de análise de sentimento, e ressalta os benefícios que este poderia trazer, como pode gerar mapeamentos de ocorrência de doenças e análise da reação das pessoas ante informações e/ou campanhas veiculadas na mídia. Essas informações poderiam impactar significativamente na divulgação e no *marketing* de temas em saúde coletiva.

Na discussão de Araújo et al. (2012) também é levantada a dificuldade de achar trabalhos em português e que utilizem a língua portuguesa na análise de sentimentos. A carência desse tipo de estudo foi constatada durante o levantamento da bibliografia, muitos trabalhos de análise de sentimentos focam em textos escritos na língua inglesa pela grande quantidade de ferramentas, bases de conhecimento léxicas e ontológicas disponíveis para esse idioma.

Nesta monografia será abordada a análise de sentimentos em nível de sentença utilizando o *framework* Apache Ignite e o modelo de programação MapReduce como ferramentas. Esses conceitos serão desenvolvidos no próximo capítulo. A proposta trata-se de uma abordagem diferente das encontradas na literatura, uma vez que, na área de análise de sentimentos há poucos trabalhos que utilizem o conceito de processamento paralelo em memória. O uso de um *framework* visa simplificar a tarefa de análise de sentimentos, a tornando mais simples e eficiente.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral deste trabalho consiste em desenvolver uma metodologia para análise de sentimentos em nível de sentença utilizando o *framework* Apache Ignite, a partir de dados extraídos do Twitter. Dessa forma, será possível demonstrar como essa tecnologia pode otimizar o processamento de grandes volumes de dados de forma simples, rápida, mesmo em *hardware* com pouca capacidade de processamento. Além disso, ela permite que grandes volumes de informações, ou *streaming* de informações textuais, possam ser analisados utilizando uma estrutura computacional já existente. Isto posto, o mesmo pode ser utilizado por instituições públicas ou privadas de qualquer porte.

1.2.2 Objetivos Específicos

- Examinar os conceitos envolvidos em *Big Data*, análise de sentimentos, principalmente no que diz respeito a análise em nível de sentença;
- Analisar o *framework* Apache Ignite para análise de sentimentos; e,
- Demonstrar o uso da metodologia proposta por meio de um estudo de caso.

1.3 Trabalhos Relacionados

As metodologias para análise de sentimentos, envolvem, em sua maioria, técnicas de aprendizado de máquina e de léxico. A abordagem proposta por Kolchyna et al. (2015) fez uma análise em nível de sentença em *tweets*, comparou a eficácia desses dois métodos e propôs uma combinação deles, visando uma maior acurácia dos resultados. O grau de positividade e negatividade das palavras contidas em cada *tweet* foi analisado. No método léxico, isso se deu através de cálculos feitos com a pontuação de cada palavra, que são atribuídos por meio de um dicionário léxico. Já no método de aprendizado de máquina o *tweet* foi previamente sinalizado como positivo ou negativo e partir desse ponto foram aplicados classificadores como *SVM* e *Naïve Bayes*. A combinação resultou na melhora de 5% nas classificações.

Seguindo a mesma linha, mas usando a língua portuguesa, o trabalho de Oliveira (2013) comparou os resultados de análises de sentimentos feitas a partir *tweets*, comentários do YouTube e do Mercado Livre. A abordagem léxica do autor utilizou o dicionário *SentiWordNet*[3] para a atribuição da polaridade. Entretanto, os resultados apontaram baixa acurácia na classificação dos *tweets,* com média de 46.4% de acertos. Além de apresentar muitos falsos negativos. Já a classificação dos comentários do YouTube teve acurácia média de 85.67% e as do Mercado Livre 76%. O autor atribuiu baixo resultado dos *tweets* ao tipo de linguagem, uma vez que no Twitter, os usuários tendem a ser a usar linguagem mais informal do que em comentários, o que dificulta a classificação.

[3] http://sentiwordnet.isti.cnr.it/

O trabalho conduzido por Moreira et al. (2016) utilizou diferentes dicionários léxicos para análise de sentimentos e comparou os seus respectivos resultados com uma análise manual feita pelos autores. Uma amostra de textos foi coletada da rede social Facebook e, enquanto as ferramentas utilizadas classificaram mais de 40% dos comentários como neutros, a análise dos autores indicou que 71% dos comentários eram positivos. Na classificação foram identificadas divergências na categorização da intensidade dos sentimentos por cada método. Nenhuma abordagem, incluindo a realizada pelo pesquisador, foi considerada por eles eficiente, uma vez que o maior nível de acurácia obtido foi inferior a 70%.

O trabalho de Sousa (2014) propôs um método para análise de sentimentos em nível de documento utilizando o *framework* GridGain. A metodologia aplicada é próxima da que é proposta neste trabalho. O método apresentado faz uso de um dicionário léxico e permite determinar a polaridade de uma pesquisa de opinião, como um todo, em positiva ou negativa. O processamento em paralelo e distribuído dos dados apresentou-se como uma ferramenta poderosa ao ser aliada com análise de sentimentos. Contudo, a metodologia possui limitações quanto ao tipo de análise de sentimentos aplicada. A análise em nível de documento não abrange aspectos específicos acerca da opinião expressada, uma vez que se trata de uma análise geral.

1.4 Organização do Trabalho

Além deste capítulo introdutório, este trabalho contém mais três capítulos e referências bibliográficas.

No capítulo 2 apresenta os conceitos fundamentais deste trabalho detalhando as propriedades e etapas *Big Data*, plataformas *open source* utilizadas para análise de *Big Data*. Além das definições e tipos de opinião, desafios, níveis e técnicas para análise de sentimentos.

O capítulo 3 aborda a metodologia utilizada, sendo esta aplicada em um estudo de caso envolvendo o console Super Nintendo. Os resultados da análise de sentimentos proposta serão discutidos, além de ser feita uma comparação com os trabalhos relacionados.

Finalmente no capítulo 4, apresenta-se a conclusão acerca das contribuições apresentadas neste monografia e sugestões para trabalhos futuros.

2 FUNDAMENTOS TEÓRICOS

Este capítulo apresenta os temas fundamentais para compreensão dos métodos utilizados no desenvolvimento deste trabalho. Aborda-se *Big Data*, suas propriedades e etapas, e as plataformas e *frameworks* utilizados para análise de *Big Data*: MapReduce, Apache Haddop, Apache Ignite. Além das definições de análise de sentimentos e de opinião, tipos de opinião, desafios da análise de sentimentos, níveis de análise de sentimentos e técnicas para análise de sentimentos.

2.1 *Big Data*

O recurso mais valioso do mundo não é mais petróleo e sim dados (THE ECONOMIST, 2017a). Essa afirmação é reforçada quando a IDC (2011) declara que vivemos em um "universo digital" cercados por mídias e conteúdos geradores de grandes volumes de dados potencialmente valiosos. Além disso, as projeções feitas por ela apontam que, entre 2005 e 2020, o universo digital crescerá de 130 exabytes[4] para 40 trilhões de gigabytes, sendo gerado mais de 5.200 gigabytes por cada homem, mulher e criança no mundo. Os fluxos desses dados criam novos paradigmas, novas infraestruturas, novos negócios, além de novas economias. Em outras palavras, estamos lidando com *Big Data*, um conceito abstrato que revoluciona a forma como analisamos esses dados e suas influências nas mais diversas áreas (GANDOMI e HAIDER, 2014).

As fotos, vídeos e postagens compartilhadas pelos usuários das redes sociais, assim como as informações produzidas por companhias áreas e as milhões de compras realizadas nos comércios eletrônicos a cada hora, são exemplos de fontes de dados que, de acordo com The Economist (2017b), podem ser caracterizadas como *Big Data.* A velocidade com que esses dados crescem faz com esses sejam, na maioria das vezes, desestruturados. Segundo Chen et al. (2014), *Big Data* refere-se a conjuntos grandes de dados desestruturados que ultrapassam a capacidade das ferramentas convencionais de banco de dados para capturar, armazenar, gerenciar e analisar. O autor deixa claro que a definição de *Big Data* ainda está em discussão, por esta ser ampla e subjetiva, uma vez que não é claro o quão grande um conjunto de dados deve ser para ser considerado como tal.

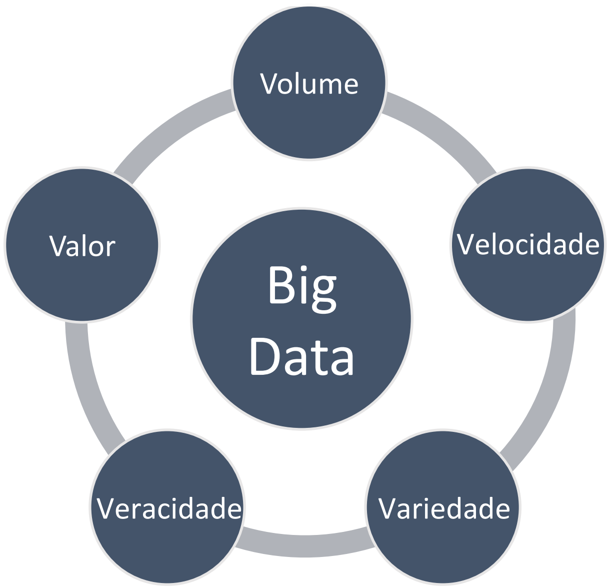
[4] 10^{18} bytes

Sagiroglu e Sinanc (2013) definem *Big Data* como um termo para conjuntos maciços de dados que possuem uma grande, variada, complexa estrutura, com dificuldades de armazenamento, análise e visualização. O processo aplicado nesses conjuntos, que revela padrões e correlações antes não percebidas, é chamado de mineração de dados. As informações extraídas pela mineração de dados são úteis para empresas e/ou organizações com o intuito de ganhar conhecimentos mais profundos e obter vantagem sobre a concorrência. Desta forma, *Big Data* tornou-se uma área de pesquisa em evidência que tem chamado cada vez mais atenção pelas suas aplicabilidades. A IDC (2016) prevê que até 2018 haja um crescimento de 26,4% por ano do mercado de soluções em *Big Data*, ultrapassando cerca de 6 vezes a taxa de crescimento do mercado global de tecnologia da informação.

2.1.1 Propriedades do *Big Data*

Para Gandomi e Haider (2014), *Big Data* pode ser caracterizado pelas três seguintes propriedades: variedade, velocidade e volume. Essa definição que é conhecida como três “Vs”, foi reformulada e, foram adicionados um quarto v, referente a veracidade, e um quinto v, referente a valor. Essa mudança, segundo Ward e Barker (2013), foi com o objetivo de evidenciar a incerteza acerca dos dados e ressaltar a importância do *Big Data* no cenário mundial, reforçando a necessidade de seu estudo e utilização. A seguir cada uma dessas propriedades serão detalhadas. Na Figura 1 pode ser observado a representação dos 5 vs.

Figura 1 – 5 Vs do *Big Data*



Fonte: Adaptado de (DEMCHENKO et al., 2013)

- **Volume:** o volume é a primeira e mais evidente propriedade do *Big Data.* De acordo com Jain (2016), *Big Data,* primeiro e acima de tudo, tem de ser "*big*". O tamanho nesse caso refere-se ao volume gerado por *e-mails*, *tweets*, fotos, vídeos, dados de sensores e etc., que produzimos e compartilhamos a cada segundo. Dessa forma, torna-se cada vez mais complicado armazenar e analisar grande volume de dados usando a tecnologia de banco de dados tradicional. Manyika et al. (2011) prevê que ao longo do tempo o tamanho dos volumes de dados classificados como *Big Data* irão aumentar, e que sua definição irá variar de acordo com a área e as ferramentas de *software* disponíveis.

- **Velocidade:** o avanço cada vez mais rápido e contínuo da tecnologia alterou a percepção de tempo e distância da pessoas, em parte, graças a velocidade de transmissão via fibra ótica. Para IRMA (2016) o fluxo de dados agora é quase tempo real e a janela de atualização foi reduzida a frações de segundos. As informações estão acessíveis pelos mais diversos meios de comunicação, principalmente pela internet e redes sociais. Um exemplo pontual é a rapidez com que as notícias são divulgadas, algo divulgado ontem já é considerado antigo. Gandomi e Haider (2014) declaram que a velocidade como propriedade de um *Big Data* diz respeito à rapidez com que os dados são gerados, armazenados e tratados. Dada a crescente popularidade dos *smartphones*, os varejistas, por exemplo, lidam com centenas de milhares de fontes de dados *streaming* que demandam análises em tempo real. Os sistemas tradicionais de gerenciamento de dados não são capazes de lidar com enormes fluxos de dados instantaneamente. As novas tecnologias de *Big Data* permitem o processamento em tempo real desses grandes volumes de dados, pois se os dados não processados a tempo, logo perder seu valor.

- **Variedade:** uma vez que os dados podem ser provenientes de diferentes fontes, eles podem apresentar diferentes formatos. Demchenko et al. (2013) cita que essa diversidade de dados inclui dados estruturados, semiestruturados e não estruturados. Os dados estruturados são facilmente interpretados, mas os dados não estruturados são aleatórios e difíceis de analisar. Os dados semiestruturados não estão em conformidade com os campos fixos, mas contém *tags* para separar os elementos de dados. Dessa forma, se faz necessário a organização dos mesmos de forma unificada, para assim torná-los significativos.

- **Veracidade:** a veracidade em *Big Data* refere-se ao nível de confiabilidade dos dados gerados. Para Schroeck e Smart (2012) manter a qualidade destes é imprescindível da mesma forma que é um desafio. Os autores ainda destacam que alguns conjuntos de informações são inerentemente incertos, como por exemplo: os sentimentos. Ao lidar com esses tipos de dados, é necessário reconhecer e abraçar essa incerteza, uma vez que ela contém informações valiosas. As eleições presidenciais de 2012 no México foram um exemplo da importância da veracidade. O Twitter ficou repleto de contas falsas que poluíram a discussão política, introduziram *hashtags*[5], e geraram muitos dados que, à primeira vista, teriam importância e que, na verdade, não tinham nenhum valor. *Big Data* é tão vasto que os problemas de qualidade são uma realidade. Um em cada três líderes empresariais não são de confiar nas informações que eles usam para tomada de decisões, logo para eles confiarem nas informações provenientes de uma plataforma *Big Data,* prover veracidade é primordial (ZIKOPOULOS et al., 2012).

- **Valor:** o v de valor, segundo Gandomi e Haider (2014), foi introduzido pela Oracle[6] e refere-se ao baixo valor que um *Big Data* têm em relação ao seu volume quando é apenas um conjunto de dados bruto. Esse valor só se eleva através da análise e aplicação dos dados. O valor se revela o principal "v", pois o poder que uma análise de *Big Data* tem de influenciar um mercado está diretamente ligada ao valor atrelado às informações dele. Se faz necessário que as empresas compreendão claramente os custos e benefícios da aplicação de uma solução de *Big Data* e o peso que ela terá em seus negócios (MARR, 2014).

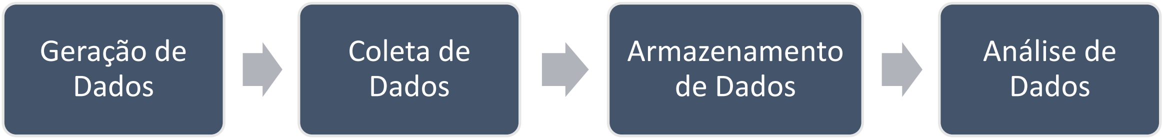
2.1.2 Etapas do *Big Data*

As etapas envolvidas na extração das informações de um grande volume de dados são conhecidas como cadeia de valor de *Big Data*. O fluxo das informações é descrito por Cavanillas et al. (2016) como uma série de etapas necessárias para gerar valor e *insights* úteis para tomada de decisão. De forma geral, essa cadeia de valor é comumente apresentada como na Figura 2. Como pode-se observar, é formada por 4 etapas: Geração de Dados; Coleta de Dados; Armazenamento de Dados; e, Análise de Dados. Contudo, no trabalho desenvolvido por Curry et al. (2014) é acrescentada mais uma etapa chamada *"Data Curation"*. Ela assegura as propriedades de valor e veracidade abordadas no subtópico anterior. O autor destaca o uso de pessoas chamadas "curadores de dados", eles têm a responsabilidade de assegurar que os dados sejam dignos de confiança, acessíveis, reutilizáveis e adequado ao seu propósito. Descrita a seguir está a cadeia de valor proposta por HU et al. (2014).

---

[5] *Hashtags* são compostas pela palavra-chave do assunto antecedida pelo símbolo cerquilha (#).
[6] Oracle *Corporation* é uma empresa multinacional de tecnologia e informática especializada no desenvolvimento e comercialização de *hardware* e *softwares* e de banco de dados.

Figura 2 – Cadeia de valor *Big Data*



Fonte: Adaptado de (HU et al., 2014)

O início da cadeia de valor, chamado de geração de dados, corresponde ao processo de produção das informações. Essas podem vir de redes sociais, informações de rede, informações de faturamento, bolsa de valores, etc. (HU et al., 2014).

A coleta de dados é o processo de aquisição, coleta, filtragem e limpeza de dados. Segundo Cavanillas et al. (2016) esta é a etapa que possui um dos maiores desafios em termos de requisitos de infraestrutura, pois essa deve suportar a aquisição de grande volume de dados além de fornecer latência baixa e previsível tanto na captura de dados, quanto na execução de consultas.

O armazenamento de dados é a persistência e o gerenciamento de dados de forma a satisfazer as necessidades das aplicações que requerem acesso rápido aos dados. Os Sistemas de Gestão de Bancos de Dados Relacionais (RDBMS) têm sido a principal, e quase única, solução para o paradigma de armazenamento por quase 40 anos. No entanto, o ACID[7], propriedades que garantem as transações nos bancos de dados, não têm flexibilidade no que diz respeito às mudanças no esquema e nem tolerância a falhas, principalmente quando os volumes de dados e a complexidade crescem. Logo o torna inadequados para cenários de *Big Data*. As tecnologias como *NoSQL*, *Hadoop File System*, *Google File System* (abordados no tópico 2.2) foram projetadas com o objetivo de escalabilidade em mente e apresentar uma ampla gama de soluções baseadas modelos de dados (CAVANILLAS et al., 2016).

Como pode ser visto no trabalho de HU et al. (2014), a análise de dados se preocupa em tornar os dados brutos adquiridos passíveis de interpretação para serem usados em tomadas de decisão. Ela envolve a exploração, transformação e modelagem dos dados com o objetivo de destacar dados relevantes, sintetizando e extraindo informações ocultas úteis com alto potencial de um ponto de vista empresarial. O tópico 2.2 abrange as principais ferramentas para análise de dados.

---

[7] Atomicidade, Consistência, Isolamento e Durabilidade

2.2 Plataformas e Frameworks *Open Source* para Análise de *Big Data*

Miller e Mork (2013) afirmam que diretamente proporcional ao crescimento dos dados, está o interesse por analisá-los. Essa afirmação é reforçada quando Chen et al. (2014) diz que a análise de dados é a etapa mais importante na cadeia de valor de *Big Data*. Uma vez que dela sai as informações processadas que irão proporcionar aos gestores a capacidade de tomar decisões em relação ao seu empreendimento. O crescente interesse em *Big Data* fez com surgissem várias ferramentas e técnicas voltadas para esta etapa. Nessa seção, serão apresentadas algumas alternativas *open source*. A seguir, serão abordados os seguintes *frameworks* e plataformas: MapReduce, Apache Hadoop e o Apache Ignite.

2.2.1 MapReduce

O MapReduce é um modelo de programação elaborado pela Google para processamento de grandes volumes de dados paralelamente sobre uma estrutura de *clusters* distribuídos. Baseia-se nas primitivas *map* e *reduce* comumente utilizadas em programação funcional. As suas principais características consistem em alta escalabilidade, suporte a tolerância a falhas e o balanceamento de carga entre os *clusters* (DEAN e GHEMAWAT, 2004).

De acordo com Dean e Ghemawat (2004), os programas baseados no MapReduce possuem duas funções básicas: *map* e *reduce*. A função *map* recebe um par chave/valor, os combina de acordo com critério estabelecido pelo usuário, e gera um conjunto intermediário de dados no mesmo formato. Em seguida, em uma função *sort*, são reunidos todos os valores intermediários que possuem a mesma chave intermediária *I.* Esses são passados como parâmetro para a função *reduce* que agrupa os valores das chaves.

Figura 3 – Pseudocódigo de contagem de palavras

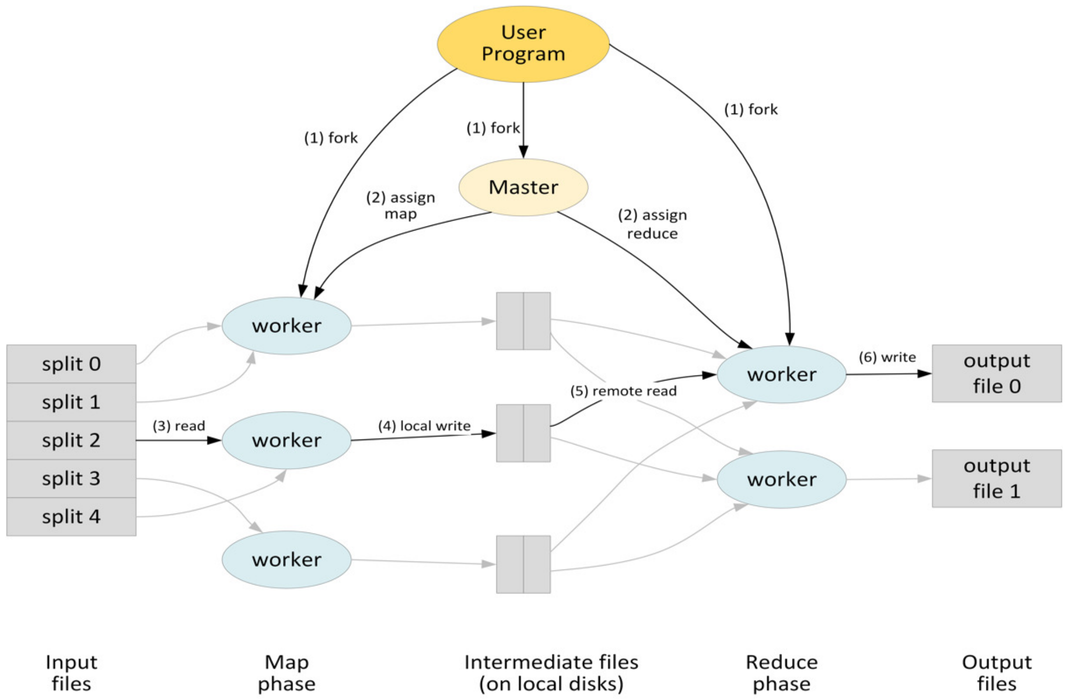
```
map(String key, String value):
  // key: document name
  // value: document contents
  for each word w in value:
    EmitIntermediate(w, "1");

reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
  int result = 0;
  for each v in values:
    result += ParseInt(v);
  Emit(AsString(result));
```

Fonte: (DEAN e GHEMAWAT, 2004)

Na Figura 3 temos um exemplo de um pseudocódigo onde a ocorrência de palavras com a letra “w” em um determinado documento é contada. A função *map* teve como entrada a chave “nome do documento/conteúdo” e para cada palavra com “w” presente no conteúdo do documento, a função *map* passa para uma função *sort* uma nova chave palavra/“1”. A função *reduce* recebe uma palavra e a lista de vezes que essa apareceu e as conta (DEAN e GHEMAWAT, 2004).

Figura 4 – Visão global de uma execução MapReduce



Fonte: (DEAN e GHEMAWAT, 2004)

A execução de uma tarefa MapReduce é apresentada de forma global pela Figura 4 e pode ser descrita no seguinte passo a passo enumerado por Paiva (2011):

1. A biblioteca MapReduce divide os dados de entrada em M pedaços e em seguida são feitas várias cópias do programa em um *cluster* de computadores.
2. Uma das cópias do programa é definida como *master* e as demais como *workers*. A função do *master* é atribuir tarefas para os *workers*. Essas podem ser tarefas de *map* ou de *reduce*, o que varia de acordo com a necessidade.

3. Aquele *worker* que for atribuído com uma tarefa de *map,* lê o conteúdo de um dos M pedaços, os processa e interpreta os pares chave/valor. Os pares intermediários de chave/valor resultantes da função são armazenados em memória.
4. De tempos em tempos os pares de dados dos *buffers* são gravados em disco em uma das R regiões divididas pela função de particionamento. As localizações desses pares de dados no disco são repassadas ao *master*, que por sua vez as informa para os *workers* que estão responsáveis pela tarefa de *reduce*.
5. Ao ser notificado pelo *master,* o *worker* de *reduce* busca os dados do disco local dos *workers* de *map*. Uma vez lidos esses dados, esses são ordenados pelas suas chaves intermediárias, para que todas as ocorrências de uma mesma chave sejam agrupadas.
6. O *worker* de *reduce* percorre os dados intermediários já ordenados e, para cada chave encontrada, ele passa a chave e os valores intermediários para a função de *reduce* previamente definida pelo usuário. A saída de cada função de *reduce* é anexada a um arquivo de saída para aquela partição de *reduce.*
7. Após completadas todas as tarefas de *map* e *reduce*, o *master* retorna ao programa do usuário. O resultado final fica disponível em R arquivos (um para cada operação de *reduce*).

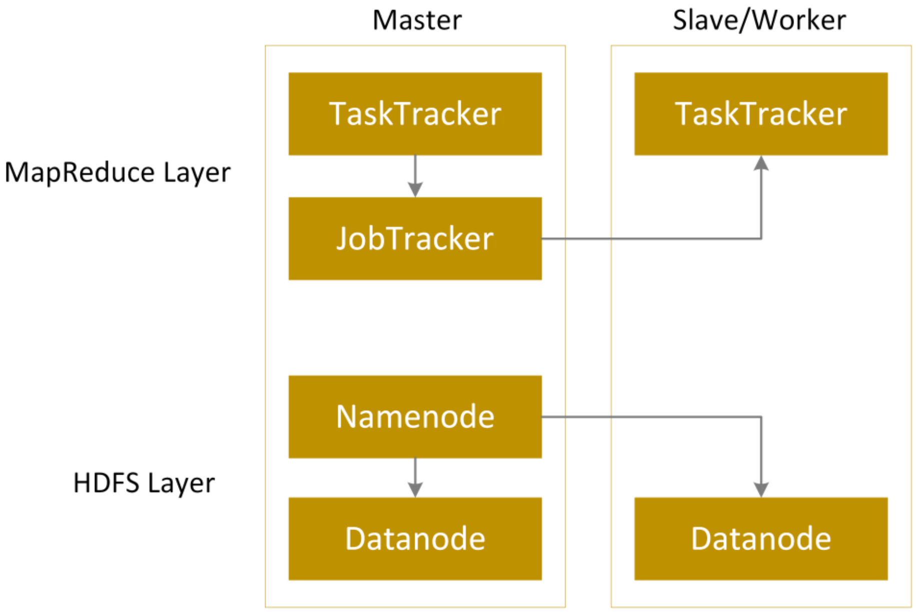
A Google utiliza o modelo de programação MapReduce em múltiplas aplicações da empresa. Dentre elas podemos citar mineração de dados, aprendizado de máquina, geração de dados para o serviço de busca na *web*, clusterização de problemas para o Google *News*, entre outras (DEAN e GHEMAWAT, 2004). Apesar do modelo de programação proposto pela Google ser público*,* a implementação que eles desenvolveram é *software* proprietário. Mesmo assim, o modelo de programação MapReduce inspirou o desenvolvimento de *frameworks open source* baseados nele. Por exemplo, o Apache Hadoop e o Apache Ignite, que serão abordados nos próximos subtópicos.

2.2.2 Apache Hadoop

O projeto Apache Hadoop abrange *softwares open source* para processamento de grandes volumes de dados, de forma confiável, escalável e distribuída. O projeto inclui quatro módulos: Hadoop Common, contém os utilitários para suporte dos outros módulos Hadoop; Sistema de Arquivos Distribuídos Hadoop (HDFS), um sistema de arquivos distribuído que provê acesso de alto rendimento; Hadoop YARN, um *framework* responsável pelo gerenciamento do *cluster* e a programação de tarefas; e, Hadoop MapReduce: Um sistema baseado em YARN para processamento de dados em paralelo (APACHE HADOOP, 2017).

Assim como o MapReduce, o Apache Hadoop possui uma estrutura que permite o processamento distribuído de grandes conjuntos de dados entre clusters de computadores usando modelos de programação simples. Ambos partilham, basicamente, a mesma arquitetura, baseada nas primitivas de *map* e *reduce*. As maiores diferenças são o sistema de arquivos, que no caso da Google é utilizado o GFS (*Google File System*), além de que o Apache Hadoop conta com vários módulos, é *open source* e fornece uma API[8] para escrita das funcionalidades de *map* e *reduce* em outras linguagens de programação além de Java.

Figura 5 – Arquitetura interna do Apache Hadoop



Fonte: (APACHE HADOOP, 2017)

Observando a Figura 5 temos a estrutura do *framework* MapReduce do Hadoop. Esse é dividido em duas camadas. Na camada do *Hadoop Distributed File System* (HDFS), o conjunto de dados de entrada são distribuídos entre o conjunto de máquinas em rede usando o HDFS. Ele é um sistema de arquivos distribuídos que para aumentar a confiabilidade, cria cópias dos blocos de dados e os redistribui em computadores próximos ao *cluster*. Pera persistir os dados, os discos locais das máquinas conectadas na rede são usados, permitindo assim que as outras máquinas na rede tenho acessado aos dados. O HDFS possui uma arquitetura do tipo

[8] *Application Programming Interface* que significa interface de programação de aplicativos

mestre/escravo. Um cluster HDFS apresenta dois tipos de nós, são eles: um *namenode* (nó de nome) e múltiplos *datanodes* (nós de dados). O *namenode* gerencia o *namespace* do sistema de arquivo, administra todos os arquivos e diretórios, mapeia os arquivos e blocos que estão sendo utilizados. Um cliente acessa o sistema de arquivos através da comunicação com *namenodes* e *datanodes.* Os datanodes armazenam e recuperam os blocos quando eles são solicitados pelo usuário ou pelo *namenode*. Apresentam periodicamente um relatório ao *namenode* com as listas de blocos que estão armazenando dados (APACHE HADOOP, 2017).

O sistema de arquivo não funcionará sem o *namenode*. Se os *namenodes* em execução forem apagados, todos os arquivos no sistema de arquivos estariam perdidos, pois não haveria outra maneira de acessar os *datanodes* senão pelos *namenodes.* Em razão disso, o Hadoop prover dois mecanismos de tolerância a falhas, sendo eles: o *backup* de estados persistentes do sistema de arquivos e a utilização de *namenodes* secundários (APACHE HADOOP, 2017).

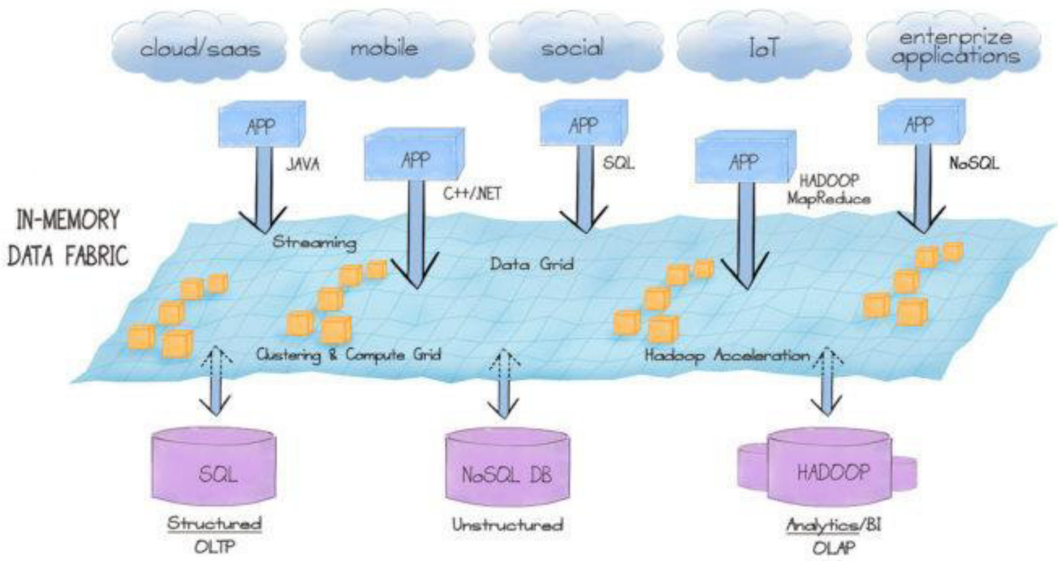
Na camada MapReduce há um único *JobTracker* e um número de processos *TaskTracker*. O *JobTracker* executa na mesma máquina que o *Namenode*. Os usuários enviam seus *jobs* MapReduce para o *JobTracker*, que divide a tarefa entre as máquinas presentes no *cluster*. Cada uma delas executa um processo *TaskTracker*, que por sua vez se comunica com o *JobTracker*, que designa uma tarefa *map* ou *reduce* quando possível. Hadoop pode ser organizado para executar múltiplas simultâneas tarefas *map* em nós simples. Em sistemas de núcleos múltiplos essa é um grande benefício, já que permite o uso total dos núcleos (SOUSA, 2014).

2.2.3 Apache Ignite

O Apache Ignite é um *framework* distribuído, integrado e de alta performance baseado em *Java Virtual Machine* (JVM), que oferece rapidez e escalabilidade ilimitada para o processamento de grandes conjuntos de dados (GRIDGAIN, 2017). Anteriormente conhecido como GridGain, ele surgiu através da doação da *GridGain Systems* em 2014 - como é explicado por Dern (2015) - para a *Apache Software Foundation* sob o projeto *open source* Apache Ignite. A versão 1.5 do Apache Ignite corresponde o GridGain Enterprise Edition 7.5.

Uma das principais das vantagens do Apache Ignite é que, mesmo utilizando um *hardware* de baixo custo, ele provê processamento de alta performance. Isso é possível graças a sua API, que repassa o processamento dos dados para *grids* em memória. O conceito por trás disso tudo é o *In-Memory Computing* que proporciona o gerenciamento de grandes volumes de dados mais rapidez e com economia de recursos. No paradigma tradicional de armazenamento, os dados ficam armazenados em discos rígidos e, para acessá-los, é necessário que a memória do computador as localize para assim executar uma ação. Esse procedimento acaba criando um gargalo, principalmente ao se tratar de processamento em *Big Data*. No *In-memory computing*, os dados ficam alojados na RAM[9], os tornando disponíveis instantaneamente para acesso.

Figura 6 – Visão geral do *In-memory Computing*



Fonte: (GRIDGAIN, 2017)

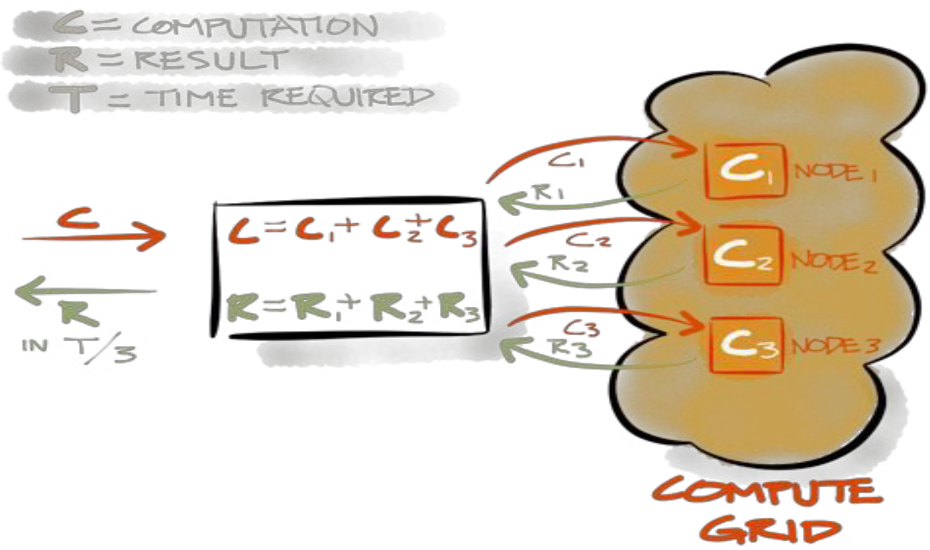
Além disso, a API do Apache Ignite se mostra muito versátil, como pode ser observado na Figura 6, suportando a utilização de diferentes bancos de dados, como por exemplo, RDBMS, *NoSQL* e *Hadoop Data Stores*; diversas plataformas em nuvem, como AWS e Microsoft Azure, ou ambiente híbrido; e, ainda múltiplas linguagens de programação como SQL, C++, .Net, Java, Scala, Groovy, PHP e Node.js. Na camada central pode ser observado alguns conceitos – abordados a seguir - como: *Streaming*, *Hadoop Acceleration, Data Grid* e *Compute Grid*, esses somados com o amplo suporte de diferentes tecnologias, fazem do Apache Ignite uma plataforma muito mais versátil do que outras do mesmo segmento.

De acordo com GridGain (2017), *Streaming* ou *In-Memory Streaming* é utilizado em conjuntos de dados de larga escala em tempo real em aplicações onde as formas tradicionais de processamento e armazenamento não são rápidas o suficiente. Dentre suas principais aplicações

[9] RAM, do inglês *Random Access Memory* (memória de acesso aleatório)

está a análise de dados por intervalo de tempo. Por exemplo: "Quais são os 10 produtos mais populares do último semestre?". Por sua vez, o *Hadoop Acceleration* melhora o funcionamento da tecnologia Hadoop existente adicionando análise em tempo real permitindo o processamento rápido de dados. Fora que possui a tecnologia *"plug and play"* onde a instalação e funcionamento do Hadoop não requer mudança de código. Segundo (IVANOV, 2016) o *Hadoop Acceleration* é compatível com o sistema de arquivo do Hadoop e provê *In-Memory* MapReduce otimizado. Por esta razão, pode ser observado um desempenho até 100 vezes mais rápido. Os conceitos abordados a seguir compõem a parte mais importante do Apache Ignite para o desenvolvimento deste trabalho.

Figura 7 – Modelo de execução do *In-Memory Compute Grid*

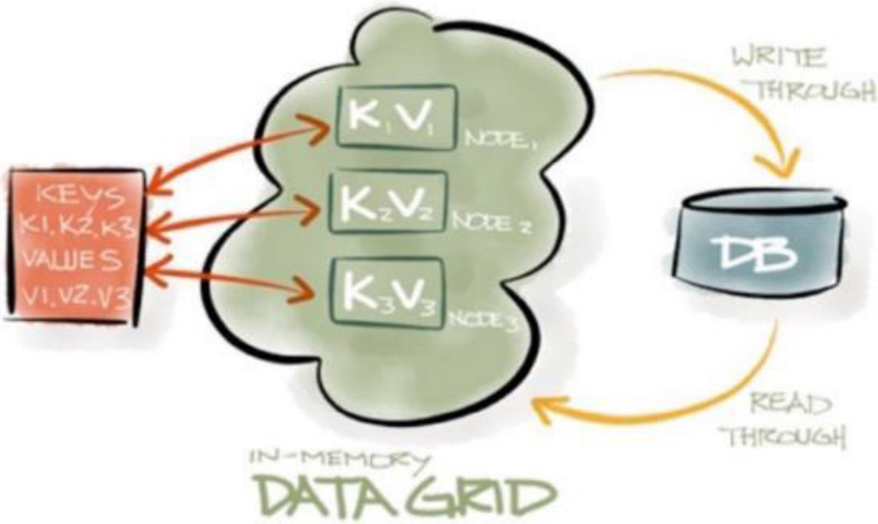


Fonte: (APACHE IGNITE, 2017)

Na Figura 7, observa-se o funcionamento do Apache Ignite MapReduce rodando em um esquema de *In-Memory Compute Grid* (IMCG). A arquitetura do MapReduce possui dois tipos de nós, mestre e trabalhadores. O nó mestre transforma uma tarefa em múltiplas sub tarefas, as balanceia e distribui para os *clusters* trabalhadores disponíveis. Esta primeira etapa é chamada de tarefa *mapping*. Em seguida, a execução é feita em paralelo e as sub tarefas são agrupas, o que é chamada etapa *reduce*. O resultado é repassado ao usuário por meio da interface da aplicação. Entretanto, segundo (GRIDGAIN, 2017), o Apache Ignite IMCG apresenta melhores resultados de processamento em computações com ciclo de vida curto, como computações que levem menos de 100 milissegundos. Isso acarreta em melhor tempo de processamento e uso dos recursos.

Apache Ignite *In-Memory Data Grid* (IMDG), Figura 8, é um sistema de armazenamento de dados em memória RAM, organizados em um esquema chave/valor (*In-Memory Key-Value*) similar a um *hash map*. O Apache Ignite IMDG é baseado em *caching* distribuído e seu diferencial para outros sistemas de armazenamento, como por exemplo o HDFS, é que um conjunto de dados pode ser atualizado e consultado ao mesmo tempo e em tempo real. A Figura 8 mostra um IMDG com um conjunto de chaves de {k1, k2, k3} onde cada chave corresponde a um nó diferente. O componente de banco de dados externo não é obrigatório. Caso haja um, o IMDG irá se conectar automaticamente para ler ou escrever no banco de dados (APACHE IGNITE, 2017).

Figura 8 – Modelo de execução do *In-Memory Data Grid*



Fonte: (APACHE IGNITE, 2017)

A diferença entre o sistema de arquivos do Apache Ignite IMDG e um sistema de arquivos tradicional é que seu modelo é baseado em domínio de usuário, realizando *caching* dos dados diretamente da interface. No mais, o Apache Ignite IMDG inclui transações ACID, *failover*, balanceamento de carga avançado e suporte extensivo ao SQL nativo para consultas, inclusive *joins* distribuídos. Utilizando memória em vez de disco, o Apache Ignite é até 1 milhão de vezes mais rápido do que os bancos de dados tradicionais (APACHE IGNITE, 2017).

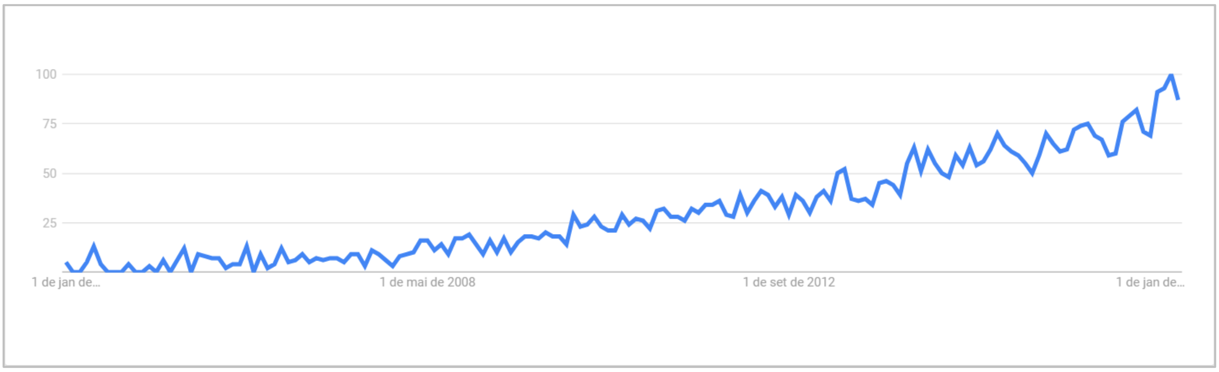
A plataforma GridGain/Apache Ignite é utilizada em mais de 500 organizações. Segundo GridGain (2017), Sony, Embrapa, Apple, Canon são apenas alguns exemplos de grandes empresas que utilizam *framework* em suas aplicações. A versatilidade da plataforma permite o seu uso em áreas como: análise de sentimentos, plataformas de comércio eletrônico, análise de modelos de investimento, processamento geoespacial, *games multiplayer online*, educação à distância, etc. Os principais benefícios de sua utilização são a diminuição do tempo de processamento, não precisar investir em *hardware* de alto custo, escalabilidade e processamento em tempo real. Diante das características abordadas nessa seção e por ser um *framework open source* baseado em Java, compatível com as mais diversas tecnologias e com arquitetura solidificada na comunidade, o Apache Ignite se mostrou a melhor opção para o desenvolvimento deste trabalho.

2.3 Análise de Sentimentos

Os vários tipos de conteúdo que são produzidos pelo consumidor final são chamados de conteúdo gerado pelo usuário, do inglês *user-generated content*. No trabalho de Almashrae et al. (2016), esse termo refere-se ao conteúdo que o consumidor produz espontaneamente sobre um produto ou assunto, por meio de *websites*, *blogs*, sites de comentários, redes sociais. Segundo Moens et al. (2014), esse conteúdo pode ser fotos, vídeos ou textos, que remetem a uma sensação ou sentimento, como por exemplo: felicidade, frustração, tristeza, indignação e que, podem ser usados como fontes de informações tanto para as empresas quanto para outras pessoas, clientes e colaboradores, agências ou até mesmo jornais e mídias em geral. Dito isso, no campo de análise textual, surge a análise de sentimentos, também conhecida como mineração de opinião, que é um campo de estudo que analisa textualmente esses conteúdos postados, determinado assunto ou alvo em um conjunto de documentos, focando em entender seus sentimentos, atitudes e emoções para convertê-los em informações úteis (LIU, 2012).

A análise de sentimentos surgiu para extrair informações de textos utilizando técnicas de mineração de dados e processamento de linguagem natural. O objetivo da análise de sentimentos é desenvolver um modelo computacional que seja capaz de identificar emoções por trás das opiniões dos usuários, e isso é possível através da classificação das opiniões como positivas ou negativas (GRAÇA NETO, 2016). A análise de sentimentos tem se tornado cada vez mais aplicada e pesquisada, na Figura 9 pode ser observado o crescimento nas pesquisas do termo "*Sentiment Analysis*" no buscador Google. Os primeiros trabalhos sobre o tema datam o ano de 2002 e foram realizados por Pang et al. (2002) e Turney (2002), de lá para cá, cada vez mais as pessoas têm buscado por opiniões *online*, tanto para produtos como para serviços, com intuito de aproveitar as melhores oportunidades além de se prevenir de fraudes.

Figura 9 – Pesquisa de Janeiro/2004-Janeiro/2017



Fonte: Google Trends (2017)

Esse aumento de interesse ocorreu com a popularização das redes sociais. Na verdade, elas são o principal objeto de pesquisa da análise de sentimentos. De acordo com Benevenuto et al. (2015), as opiniões contidas nas redes sociais permitem compreender e explicar diversos fenômenos sociais complexos, além de prevê-los. Além disso, a análise de sentimentos auxilia empresas das mais diversas áreas e organizações a entender como seus clientes pensam e consomem, buscando a opinião deles a respeito da qualidade de seus produtos ou serviços (LIU, 2012). Embora, a análise de sentimentos possa ajudar a descobrir o que os clientes pensam sobre determinada marca, é necessário entender que há uma grande diferença entre o que as pessoas estão dizendo e porquê de elas estarem dizendo. Uma ferramenta de análise de sentimentos deve tentar responder tanto o "o que" e o "porquê", pois se ela simplesmente processar a opinião e não ter a capacidade de responder a essas perguntas, a informação coletada não terá impacto nenhum (ZIKOPOULOS et al., 2012).

2.3.1 Definição de Opinião

O dicionário do Aurélio (2017) define a palavra “opinião” como uma “manifestação das ideias individuais a respeito de algo ou alguém”. No entanto, para a análise de sentimentos, a definição de opinião se estende também a sua estrutura. O trabalho feito por Silva (2015) explora o seguinte exemplo para ilustrar a estrutura de uma opinião: “ (1) Eu comprei um iPhone alguns dias atrás; (2) Ele parecia ser um ótimo celular; (3) O *touch screen* era realmente bom; (4) A qualidade de voz era clara também; (5) Porém, minha mãe ficou furiosa comigo por não ter avisado a ela antes de ter feito a compra dele; (6) Ela também achava que o celular era muito caro e queria que eu o devolvesse para a loja”. Alguns elementos podem ser extraídos das opiniões expressadas nesse exemplo, um deles é o sentimento: positivo, como as letras (2), (3) e (4), ou negativo, como as letras (5) e (6). Além disso, é possível identificar que cada uma das sentenças possui alvos, ou seja, um elemento central a que se refere. Por exemplo, na sentença (3) é o *touch screen*.

Além desses elementos, outros são formalizados por Liu (2012) e podem ser identificados no exemplo acima:

- **Entidade:** uma entidade pode ser uma marca, instituição, pessoa, produto, serviço, evento, etc. Ela é associada com um par: e(T, W), onde T é a hierarquia de componentes e subcomponentes e W é o conjunto de atributos de e. No texto, o iPhone é considerado uma entidade;
- **Aspecto:** os aspectos de uma entidade são os componentes e atributos de e, no caso, o conjunto de componentes do iPhone seriam a tela, botões, carregador, fone de ouvido. Já o conjunto de atributos seriam elementos como peso, tamanho e cor;
- **Nome de aspecto e Expressão de aspecto:** o nome de aspecto refere-se ao nome que um usuário associa a um aspecto, enquanto uma expressão de aspecto é a real palavra ou frase que apareceu no texto indicando um aspecto. No texto, um aspecto do iPhone tem é o nome de câmera frontal, mas há outras expressões que o representam, tais como câmera, câmera de selfie;
- **Nome de entidade e Expressão de entidade:** um nome de entidade é o nome de uma entidade atribuído por um usuário, já uma expressão de entidade é uma palavra real ou frase que apareceu no texto indicando uma entidade; e,
- **Dono da opinião:** é o locutor que expressou a opinião, podendo este ser uma pessoa ou uma organização.

Dito isto, podemos afirmar que uma opinião é uma quíntupla, (e_i, a_{ij}, s_{ijkl}, h_k, t_l), onde e_i corresponde o nome de uma entidade, a_{ij} é um aspecto de e_i, s_{ijkl} representa o sentimento no aspecto a_{ij} da entidade e_i, h_k é o dono da opinião e t_l é o tempo em que a opinião é expressada por h_k (LIU, 2012). O sentimento s_{ijkl}, pode ser classificado como positivo, negativo, ou neutro, às vezes, por ser expressado por níveis de intensidade. Isso é visto, por exemplo, em escala de em pesquisas de opinião, onde têm opção de “concordo totalmente”, “concordo parcialmente”, ambas se enquadram em positivo mas possuem intensidades diferentes.

2.3.2 Tipos de Opinião

Existem dois tipos principais de opiniões: opiniões regulares e opiniões comparativas. Como explicado por Dubey e Gupta (2016) as opiniões regulares referem-se a uma opinião expressada diretamente a uma entidade ou um aspecto da entidade e essas se subdividem em dois grupos: opiniões diretas e indiretas. A maioria dos estudos sobre opiniões regulares focam em opiniões diretas. Por exemplo, a sentença “O pôr do sol visto da praia é o mais encantador”, nela entidade é bem definida e opinião clara, sendo assim mais simples de lidar. Em contrapartida das opiniões indiretas dificultam a identificação da entidade principal por levar em consideração mais de uma. Esse tipo de situação é apontado como um dos principais desafios de uma análise e está desenvolvido no próximo tópico.

Uma opinião comparativa, expressa uma relação de semelhanças ou diferenças entre a mesma, duas ou mais entidades. Ela se subdivide em opiniões explícitas e implícitas. Uma opinião explícita é a uma afirmação subjetiva que dá uma opinião regular ou comparativa, por exemplo, “O sorvete da Häagen-Dazs é o melhor que o do Ben & Jerry’s”. Uma opinião implícita é uma afirmação objetiva que implica uma opinião regular ou comparativa. Em geral, expressa um fato desejável como por exemplo, “A velocidade do processador e a resolução da tela do smartphone S6 são melhores do que IPhone 6; no entanto, o corpo de metal do IPhone 6 é mais atraente do que S6”. A grande maioria dos pesquisadores optam por trabalhar com opiniões explicitas, uma vez que são mais facilmente detectadas e classificadas que opiniões implícitas (PATEL et al., 2015).

2.3.3 Desafios da Análise de Sentimentos

No trabalho de Kolkur et al. (2015) são apontados alguns dos principais desafios da análise de sentimentos: sarcasmo, negação, comparações. Todavia, é pertinente acrescentar um ponto considerado por Silva (2013) que é conhecimento de mundo.

- **Sentimentos implícitos e sarcasmos:** as sentenças de um texto podem apresentar sentimentos implicitamente, uma vez que pode não haver a presença clara destes através das palavras. No exemplo: "Esse jejum intermitente é estranho, como que alguém consegue passar um dia inteiro sem comer nada?", percebe-se que essa sentença não carrega explicitamente um teor negativo através das palavras apesar de ela possuir essa polaridade.

O mesmo acontece com o sarcasmo, onde intenção do autor da opinião, com jogo de palavras, é expressar o contrário do que está escrito.

- **Comparações:** muitas das opiniões são expressadas através de comparação de entidades. Determinar a polaridade das frases comparativas pode ser um desafio, por não conseguir definir qual é a entidade principal. Por exemplo, “A câmera do iPhone 7 é melhor que a do Samsung S8”. O sentimento está centrado na palavra positiva "melhor", mas se o objeto analisado é o iPhone 7, a opinião é classificada como positiva, se o objeto de análise for o Samsung S8 é negativa. Logo é importante identificar para qual das entidades do contexto a opinião dada é direcionada em vista da classificação correta.

- **Negação:** o maior desafio da análise de sentimentos é lidar com as negações. Isso se deve ao fato de que nem sempre elas estão explícitas, exigindo considerar, o contexto em que foi empregada a negação para saber se ela realmente se caracteriza como uma. Considere os seguintes exemplos: (1) “Eu não gosto de manga”; (2) “O filme da mulher maravilha é sensacional, porém não teve nenhum efeito 3D”; (3) “Eu não apenas gostei da jogabilidade do novo jogo *Tomb Raider*, como adorei os gráficos”. Em (1), a negação estava explícita, onde é possível identificar o operador de negação na frase. No caso de (2) é preciso perceber que o operador de negação (porém) não está alterando a polaridade da primeira sentença. O exemplo (3) representa o caso mais complicado. A polaridade da primeira sentença não é alterada em vista da presença do operador, apesar te de o “não” o autor faz uma comparação positiva.

- **Conhecimento de mundo:** cada linguagem tem suas peculiaridades, expressões regionais e, principalmente, gírias. Logo, se faz necessário algum conhecimento de mundo seja adicionado as ferramentas que analisam sentimentos. O exemplo “O novo visual do Chevrolet Ônix ficou ralado” retrata um sentimento negativo, porém, para identificá-lo, é preciso ter um conhecimento das gírias e levar em consideração os regionalismos e a língua falada. Esta é uma tarefa complicada, que não sendo levada em conta, pode levar a uma classificação errada e comprometer a análise.

2.3.4 Níveis de Análise de Sentimentos

A análise de sentimentos pode ser aplicada em três níveis e segundo Liu (2012), esses são: sentença, aspecto, documento. Entretanto, é possível achar na literatura trabalhos que abordam outros níveis como contexto e usuário. A importância da área de pesquisa assim como o quão aprofundado se deseja os resultados, determinam em que nível será aplicada a análise de sentimentos.

- **Análise em nível de sentença:** analisa e determina se cada frase do documento analisado expressou uma opinião positiva, negativa ou neutra. De acordo com Bongirwar (2015) neste nível a análise está intimamente relacionada com a classificação da subjetividade, é possível distinguir sentenças objetivas, que expressam fatos, de sentenças subjetivas que expressam pontos de vista e opiniões subjetivas. No entanto, ser classificada como uma sentença subjetiva não significa necessariamente que está expressando um sentimento, assim como sentenças objetivas podem implicar opiniões. Dentre as principais aplicações que utilização este nível de análise encontra-se os trabalhos que utilizam o Twitter como fonte de dados. Kolchyna et al. (2015) faz uma análise a nível de sentença onde é analisado o grau de positividade e negatividade de cada palavra contida em um *tweet*. Ao utilizar esse nível, a análise é mais aprofundada.

- **Análise em nível de aspecto e entidade:** tem sido bastante estudado no contexto de revisões de produtos e serviços, uma vez que, realiza uma análise que ao invés de focar nas construções de linguagem (sentenças, documentos, parágrafos), constrói a análise com base no alvo da opinião (BECKER e TUMITAN, 2013). A crescente demanda por ferramentas que processem opiniões no meio corporativo onde as empresas necessitam avaliar as opiniões sobre seus produtos, foi o que levou trabalhos como o de Kauer (2016) a serem desenvolvidos. A metodologia aplicada por ele visou reduzir o número de atributos utilizados para a classificação de textos, desenvolvendo um método que identifica expressões que mencionem aspectos e entidades em um texto, utilizando ferramentas de processamento de linguagem natural combinadas a algoritmos de aprendizagem de máquina.

- **Análise em nível de documento:** Ferilli et al. (2016) destaca que a classificação é dada ao documento como um todo, expressando um sentimento positivo ou negativo. Este tipo de análise é adequado quando o documento trata de uma única entidade, por exemplo, um documento que forneça opiniões sobre o seriado *House of Cards*[10]. Onde será dito se o documento é apenas positivo ou negativo em relação a sua entidade. Pang e Lee (2004) mostraram pela primeira vez que uma análise em nível de sentença poderia melhorar o desempenho de uma análise em nível do documento. Um exemplo disso é o trabalho de Farra et al. (2010), onde foi classificado o sentimento de um documento em árabe, investigando

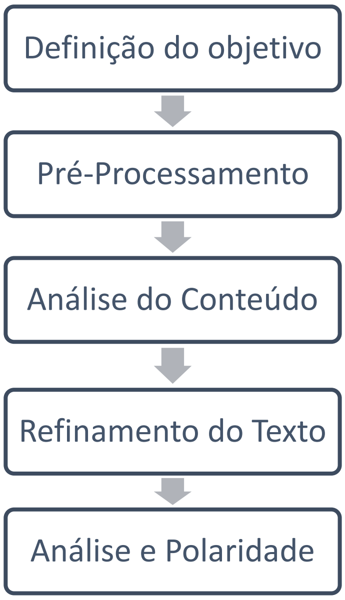
[10] Seriado de drama político produzido pela empresa Netflix

as polaridades de sentenças individuais para extração do sentimento geral. Cada documento foi dividido em tempo de execução em um número *n* de pedaços. A maioria dos documentos não representam apenas um único ponto de vista, normalmente são compostos por vários. Ao reduzir um documento inteiro a única polaridade, perde-se muito da análise de sentimentos individual que pode vir a conter os melhores *insights*.

2.3.5 Técnicas para Análise de Sentimentos

Os passos envolvidos em uma análise de sentimentos estão, em um alto nível, representados na Figura 10. Após, a fase de pré-processamento é possível escolher entre três tipos de técnicas para a classificação de opiniões, essas são: aprendizado de máquina, léxico e abordagem híbrida (MADHOUSHI et al., 2015). Classificadores como *Support Vector Machine* (SVM) e *Naïve Bayes* estão entre as abordagens mais populares de aprendizado de máquina. Já as abordagens léxicas utilizam, principalmente, partes da língua falada e dicionários como, por exemplo, o *SentiWordNet*. A terceira técnica combina aprendizado de máquina e abordagem léxico (GODSAY, 2015). A seguir, as duas principais técnicas serão detalhadas:

Figura 10 – Etapas da análise de sentimentos



Fonte: Adaptado de (GODSAY, 2015)

- **Técnicas baseadas em léxico**: Segundo Taboada et al. (2011), as técnicas baseadas em léxico fazem a classificação baseada em valores positivos e negativos dos sentimentos. Com esta abordagem é usado um dicionário de palavras que associa uma palavra a um valor de sentimento positivo ou negativo. De um modo geral, Jurek et al. (2015) sintetiza que as abordagens baseadas em léxico, representam um pedaço de mensagem de texto como um saco de palavras. A partir daí os valores de sentimento do dicionário são atribuídos a todas as palavras ou frases positivas e negativas dentro da mensagem. Depois, uma função de combinação, como soma ou média, é aplicada para fazer a previsão final em relação ao sentimento geral da mensagem. Além de um valor de sentimento, o aspecto do contexto local de uma palavra é geralmente levado em consideração, como negação ou intensificação. Por exemplo "Hoje está quente" é diferente de "Hoje está muito quente". Existem três formas de construir um dicionário léxico de acordo com Pang e Lee (2008): construção manual, métodos baseado em dicionário e métodos baseado em *corpus*. O primeiro é a forma mais trabalhosa, uma vez que é necessário o usuário manualmente colete e atribua um peso a cada palavra ou uma pontuação, indicando o quão positiva ou negativa é esta. Na abordagem baseada em dicionário, as palavras são buscadas em um dicionário, como por exemplo o *SentiWordNet* e é recuperado uma pontuação positiva e negativa atribuído para cada palavra. Técnicas baseadas em *corpus* por sua vez começam com um dicionário predefinido de palavras positivas e negativas, e depois usam contagens de palavras ou outras medidas de incidência de palavras e frequência para classificar todas as opiniões. Esse tipo de abordagem permite a produção de palavras de opinião com relativamente alta acurácia e ajuda a encontrar um domínio especifico para as palavras de opinião e sua orientação (RICE e ZHORN, 2013).

- **Técnicas baseadas em Aprendizado de Máquina**: esses tipos de técnica são conhecidos por técnicas de classificação supervisionada. Isso significa que é necessário dois conjuntos de documentos, um conjunto de treino e outro conjunto de teste, os métodos de classificação de texto que utilizam a abordagem de aprendizado de máquina podem ser divididos em métodos de aprendizagem supervisionados e não supervisionados. Os métodos supervisionados utilizam uma grande quantidade de documentos de treinamento rotulados. Os métodos não supervisionados são usados quando é difícil encontrar esses documentos de treinamento rotulados (MEDHAT et al., 2014). Os classificadores mais utilizados em técnicas de aprendizado segundo Schauwen (2010) são: *Naïve Bayes* (NB), *Maximum Entropy* (ME) e *Suport Vector Machine* (SVM). De modo geral, as etapas de aprendizado de máquina são iniciadas com a coleta do conjunto de dados que serão de treino, em seguida, esse é submetido ao classificador escolhido para treino. Após essa etapa, o próximo passo é escolher as *features*, ou seja, quais características são mais importantes e terão mais peso na análise. O conjunto de treino tem a função de fazer o classificador aprender a diferenciação de características de documento e o conjunto de teste é usado para testar a performance do classificador em um conjunto de dados maior. (SCHAUWEN, 2010).

Este capítulo discorreu sobre os principais pontos que envolvem *Big Data,* o *framework* Apache Ignite, MapReduce e análise de sentimentos. Foram expostas as definições, propriedades, finalidades de cada um deles.

No próximo capítulo eles serão aplicados no desenvolvimento da metodologia de análise de sentimentos proposta, através de um estudo de caso.

3 ESTUDO DE CASO

A empresa japonesa de jogos eletrônicos Nintendo anunciou no dia 26/06/2017 que relançará o console Super Nintendo com 21 jogos na memória (G1, 2017). O anúncio provocou diversas reações e também muitos comentários nas redes sociais, especialmente no Twitter. Dessa forma, os *tweets* relacionados a Nintendo e ao lançamento do console são o tema escolhido para ser feita a análise de sentimentos em nível de sentença. No intuito de apresentar a metodologia proposta para este trabalho, esse capítulo mostrará, através do desenvolvimento de um estudo de caso sobre o Super Nintendo, o passo a passo da aplicação dos conceitos apresentados e explicados no capítulo anterior. Os softwares e hardwares utilizados neste trabalho estão descritos no tópico 3.1 e, no tópico 3.2, a metodologia adotada será explicada. Logo após, as etapas da metodologia serão abordadas e detalhadas individualmente.

3.1 Software e Hardware utilizados

Cada uma das etapas da metodologia requereu diferentes softwares, como o trabalho utiliza o conceito *open source,* esses são gratuitos e/ou de código aberto. A captura dos dados da base foi feita através do uso da linguagem de programação Python e utilizando o ambiente de desenvolvimento Spyder, disponível no site do Git Hub[11]. Os dados foram exportados para uma planilha em formato .csv que é compatível tanto com Excel como com Libre Office Calc.

A ferramenta de análise de sentimentos foi desenvolvida na linguagem de programação Java, utilizando a IDE Eclipse Neon, disponível no site da *Eclipse Foundation*[12]. A incorporação do *framework* Apache Ignite no código foi feita através da API Java obtida no site oficial[13]. A versão do *framework* utilizada neste trabalho foi a 1.9.0 de março de 2017.

O hardware utilizado no desenvolvimento e testes foi um *ultrabook* DELL, com processador Core Intel I7 de quarta geração, 16 GB de memória RAM, 1TB de HD[14] e sistema operacional Windows 10.

[11] https://github.com/spyder-ide
[12] http://www.eclipse.org/download/
[13] https://ignite.apache.org/download.cgi
[14] HD, do inglês *Hard Drive* (disco rígido)

3.2 Metodologia Proposta

A metodologia proposta para o desenvolvimento deste trabalho está ilustrada na Figura 11. As etapas são: a aquisição dos dados; o pré-processamento; a análise de sentimentos; e, a avaliação dos resultados.

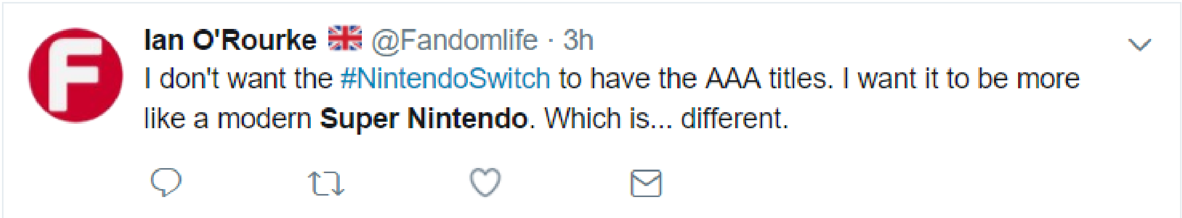
Figura 11 – Etapas da metodologia proposta

Fonte: COSTA (2017)

3.2.1 Aquisição dos Dados

Nesta etapa foi escolhido o Twitter como a fonte da base de dados. Ele é uma rede social onde os usuários postam suas ideias e compartilhem informações em tempo real em forma de *tweet* (Figura 12). De acordo com Zhang (et al., 2011), o Twitter, pode ser considerado como uma valiosa fonte online de opiniões. As razões que levaram a essa escolha dessa rede social como fonte de dados, foram primeiramente, por conta do *tweet* ser de no máximo 140 caracteres, o que força o usuário ser objetivo em sua mensagem. Além de que a análise de sentimentos no Twitter é rápida, eficiente e uma boa forma de estudar a opinião pública para o *marketing* comercial e/ou estudos sociais. Por exemplo, uma empresa pode ter *feedback* imediato sobre um novo produto ou anúncio no mercado avaliando as opiniões das pessoas no Twitter.

Figura 12 – Exemplo de *tweet*



Fonte: Twitter (2017)

O Twitter disponibiliza aos desenvolvedores o site chamado *Twitter for Developers*[15] onde encontra-se documentação de como ter acesso aos *tweets*, registro do desenvolvedor (onde é disponibilizada uma chave de autorização), além de duas APIs para recuperação de dados, a Twitter *search* API e a *streaming* API. A Twitter s*earch* API, recupera postagens recentes de usuários a partir de requisições HTTP, usando o método GET através do endereço http://search.twitter.com/search.json?q=“parâmetro de busca”. A sua principal limitação é o fato da busca retornar apenas mensagens recentes, postadas em um período de máximo 7 dias anteriores a data da busca. Dependendo do termo buscado, não se consegue uma amostra grande de *tweets.* Além disso, ela não permite a realização de buscas complexas, não podendo passar mais que um parâmetro por requisição.

O *streaming* dos *tweets* se mostrou mais vantajoso para este trabalho pois as buscas são atualizadas em tempo real, é possível colocar mais de um parâmetro de busca, além de proporcionar acesso a uma maior quantidade de *tweets.* Sendo assim, a coleta dos *tweets* ocorreu através do uso da biblioteca Python chamada Tweepy[16] acessando a API de *streaming* do Twitter. Os *tweets* foram coletados no dia 26 de junho de 2017 no período da tarde, por cerca de 15 minutos, totalizando 1.689 *tweets*. Os termos de pesquisa foram “Super Nintendo” e “SuperNintendo”, segundo o Twitter (2017), os *tweets* que contenha o termo de pesquisa, mesmo estando em maiúsculo ou minúsculo ou em *hashtag*, são recuperados. Logo, nos resultados que foram coletados, há termos como “#SuperNintendo”, “#Nintendo”, “Super Nintendo”.

A escolha do tema foi baseada nos *Trending Topics*, que é uma lista, atualizada em tempo real, dos assuntos que estão em alta no Twitter. A Figura 13 mostra um *print* retirado do site https://trends24.in que mantém histórico do *Trending Topics* ao longo das horas. Ela é composta por *hashtags* e também assuntos gerais. A linguagem escolhida dos *tweets* foi inglês, uma vez que os assuntos que estavam em alta, e que tinham uma quantidade relevante de *tweets* em português, se tratavam de política ou não eram interessantes de serem usados em uma análise de sentimentos. Dessa forma, se optou por utilizar *tweets* em inglês o que coincidiu com o anúncio da Nintendo que estava sendo comentado, majoritariamente, em inglês.

[15] https://dev.twitter.com/
[16] http://www.tweepy.org/

Figura 13 – *Trending Topics* do dia 26/06/2017

Fonte: https://trends24.in/brazil/

Os *tweets* são recuperados em formato JASON e, como pode ser visto na Figura 14, este possui diversos atributos como data de criação, id, texto, *url*, localização, nome do usuário, etc., porém, nem todos esses atributos são necessários para a análise de sentimentos, então o algoritmo desenvolvido filtrou apenas o atributo "text" e "lang" e converteu para arquivo em formato .csv.

Figura 14 – Exemplo JSON de um *tweet*


```
"{
                "created_at":"Mon Jun 26 18:12:08 +0000 2017",
                "id": 879401911631454209,
                "id_str": "879401911631454209",
                "text":"RT @Press_Quit: Usually we link articles in the day, but this can't wait! Nintendo announces the SNES
                Mini! (all detail in link)",
                "source":"<a href="http://twitter.com/download/iphone" rel=\nofollow"\>Twitter for iPhone"</a>,
                "truncated": false,
                "in_reply_to_status_id": null,
                "in_reply_to_status_id_str": null,
                "in_reply_to_user_id":null,
                "in_reply_to_user_id_str":null,
                "in_reply_to_screen_name":null,
                "user":{
                                "id":1483518709",
                                "id_str": "1483518709",
                                "name": "DrinkABeer&PlayAGame",
                                "screen_name": "ABeerAndAGame",
                                "location":null,
                                "url": "http://www.drinkabeerandplayagame.com/",
                                "description": "We drink beers, we play games, we review both and like to talk about it.
                                Podcast is on iTunes: our YT - https://m.youtube.com/channel/UCkEXtBQymCMPmRdkLIzNCtA",
                                "protected": false,
                                "verified": false,
                                "followers_count": 7812,
                                "friends_count": 3364,
                                "listed_count": 170,
                                "favourites_count": 67873,
                                "statuses_count": 27670,
}
```

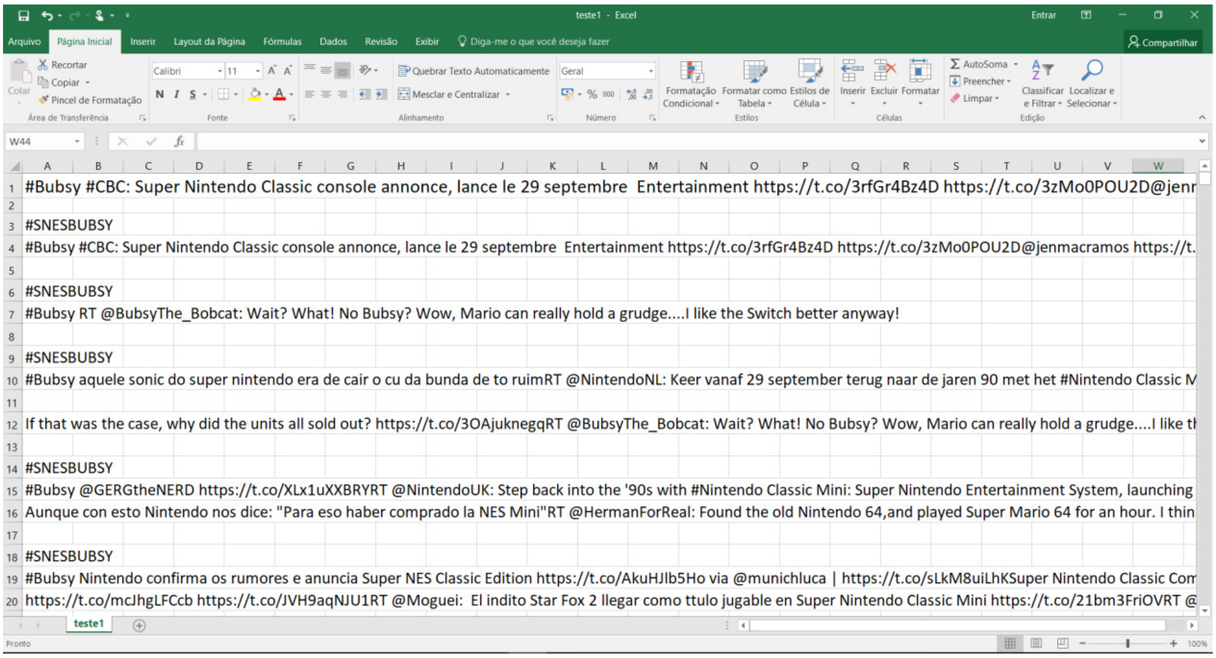

Fonte: COSTA (2017)

3.2.2 Pré-Processamento

Nesta etapa é executado um pré-processamento da base de dados. Isso é necessário pois, as técnicas baseadas em léxico, requerem que os dados que serão analisados estejam em formato textual para que esses sejam comparados com um dicionário de palavras que expressa sentimentos. Uma vez que a palavra identificada no dicionário ela é associada a uma pontuação. Dito isso, é necessário retirar do *tweets* partes textuais que atrapalhem essa comparação. Na Figura 15 é possível ver a base sem formatação e a seguir estão descritos os elementos eliminados.

A API *streaming* retorna *tweets* em diversas linguagens. Posto que a linguagem escolhida para a análise de sentimentos foi o inglês, aqueles *tweets* que foram capturados e que não possuíam o atributo "*lang*" igual a "en" foram descartados.

Figura 15 – Base não formatada



Fonte: COSTA (2017)

Uma redução dos dados foi feita devido à grande quantidade de *retweets* presentes na base. Esse termo se refere a um *tweet* republicado, normalmente ele vem antecedido da expressão "RT". Dessa forma, para não comprometer os resultados ao analisar *tweets* repetidos, todos aqueles precedidos da expressão "RT" foram removidos.

Alguns *tweets* apresentaram caracteres em formato Unicode, assim como sinais de pontuação e *emoticons* que, para a metodologia adotada, não são relevantes. Dessa forma, todos esses caracteres foram removidos através do uso de expressões regulares.

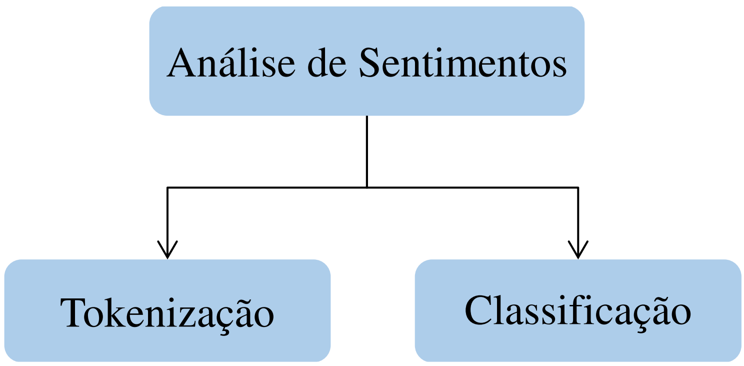
Um problema bastante comum encontrado foi o uso de abreviações e acrônimos para se referirem a determinadas palavras. Por exemplo, “lol” e “omg” que significam "*lots of laughs*" e “*oh my god*” respectivamente. Desse modo, cada abreviação e acrônimo encontrado que esteja presente na lista elaborada por Lee (2015) foi substituído por seu respectivo significado. Dessa maneira a palavra pode vir a ser encontrada no dicionário.

Por fim, foi retirada uma amostra de 94 *tweets* da base e esses foram classificados manualmente como positivos ou negativos. A expressão “pos” foi adicionada na frente daqueles considerados positivos e “neg” nos considerados negativos. Essa classificação é importante para discussão e comparação dos resultados além de testar se o algoritmo implementado foi eficiente.

3.2.3 Análise de Sentimentos

Nesta etapa será analisado e identificado, em nível de sentença, o sentimento contido em *tweets* sobre o anúncio do relançamento do Super Nintendo. Isso se dará através do uso do *framework* Apache Ignite. A técnica aplicada foi a baseada em léxico, introduzida no subtópico 2.3.5. Ao final da análise as opiniões estarão polarizadas, ou seja, estarão classificadas como positivas ou negativas. Para melhor compreensão do método aplicado, ilustrada na Figura 16, a etapa de análise de sentimentos foi subdividida em duas partes: tokenização e classificação.

Figura 16 – Subdivisões da etapa de Análise de Sentimentos



Fonte: COSTA (2017)

A adaptação da lógica da análise de sentimentos ao modelo de programação MapReduce visto no subtópico 2.2.1 permite a criação de um algoritmo mais simples e intuitivo, além de prover processamento paralelo e distribuído. A primeira parte da etapa de análise de sentimentos proposta, chama-se tokenização e corresponde a tarefa *map* do modelo de programação MapReduce e a parte chamada de classificação à tarefa *reduce*. A seguir o passo-a-passo adotado em cada uma delas será detalhado.

3.2.3.1 Tokenização

Os nós da *grid* do *framework* Apache Ignite são inicializados ao executar no ambiente Windows, o comando "ignite.bat" passando um arquivo XML de configuração. Esse arquivo contém detalhes como: endereço do host, caminho dos *jars*[17], método de carregamento das classes, etc. Os nós trabalhadores devem ser inicializados antes da execução da aplicação, na Figura 17 está demostrado a inicialização de um nó. Por sua vez, o nó mestre é instanciado no Eclipse através da aplicação e ele automaticamente detectará os outros nós e distribuíra as tarefas.

Figura 17 – *Script* de um nó do Apache Ignite no ambiente Windows

```
Prompt de Comando - ignite.bat  examples/config/example-ignite.xml
Microsoft Windows [versão 10.0.15063]
(c) 2017 Microsoft Corporation. Todos os direitos reservados.

C:\Users\Carol Bras>cd c:/apache/1.9.0/bin/

c:\Apache\1.9.0\bin>ignite.bat examples/config/example-ignite.xml
[23:40:02]    __________  ________________
[23:40:02]   /  _/ ___/ |/ /  _/_  __/ __/
[23:40:02]  _/ // (7 7    // /  / / / _/
[23:40:02] /___/\___/_/|_/___/ /_/ /___/
[23:40:02]
[23:40:02] ver. 1.9.0#20170302-sha1:a8169d0a
[23:40:02] 2017 Copyright(C) Apache Software Foundation
[23:40:02]
[23:40:02] Ignite documentation: http://ignite.apache.org
[23:40:02]
[23:40:02] Quiet mode.
[23:40:02]   ^-- Logging to file 'C:\Apache\1.9.0\work\log\ignite-bc38d4f2.0.log'
[23:40:02]   ^-- To see **FULL** console log here add -DIGNITE_QUIET=false or "-v" to ignite.{sh|bat}
[23:40:02]
[23:40:02] OS: Windows 10 10.0 amd64
[23:40:02] VM information: Java(TM) SE Runtime Environment 1.8.0_121-b13 Oracle Corporation Java HotSpot(TM) 64-Bit Serv
er VM 25.121-b13
[23:40:02] Configured plugins:
[23:40:02]   ^-- None
[23:40:02]
[23:40:03] Message queue limit is set to 0 which may lead to potential OOMEs when running cache operations in FULL_ASYNC
 or PRIMARY_SYNC modes due to message queues growth on sender and receiver sides.
[23:40:04] Security status [authentication=off, tls/ssl=off]
[23:40:17] Performance suggestions for grid  (fix if possible)
```

Fonte: COSTA (2017)

[17] Um arquivo compactado usado para distribuir um conjunto de classes Java

O ambiente estando pronto e a aplicação inicializada, o dicionário léxico é carregado em memória. O dicionário escolhido foi o *SentiWordNet* 3.0, ele atribui a cada termo do dicionário *WordNet*[18] duas pontuações de sentimento: positividade, negatividade. Abaixo, na Figura 18, temos uma das 117.659 linhas que compõem o dicionário.

Figura 18 – Exemplo de linha do dicionário *SentiWordNet*

```
a 00035779 0.25 0.125 brisk#3 very active; "doing a brisk business"
```

Fonte: *SentiWordNet* (2017)

Os elementos que compõe a linha representam: o primeiro é o *Part-of-Speech* (POS), em português parte do discurso, que se refere ao tipo gramatical da palavra. O *SentiWordNet* trabalha com cinco tipos: “a”, adjetivo, “n”, substantivo, “r”, advérbio, “s”, adjetivo satélite e “v”, verbo; o segundo elemento é o ID, um identificador único do *synset* (nome dado os verbetes do dicionário); o terceiro é a pontuação positiva; o quarto é a pontuação negativa; o quinto elemento é s*ynset* em si, caso ele possua sinônimos dentro dicionário, elas também aparecem na mesma linha; em seguida está o seu significado; e, por último o emprego em uma frase.

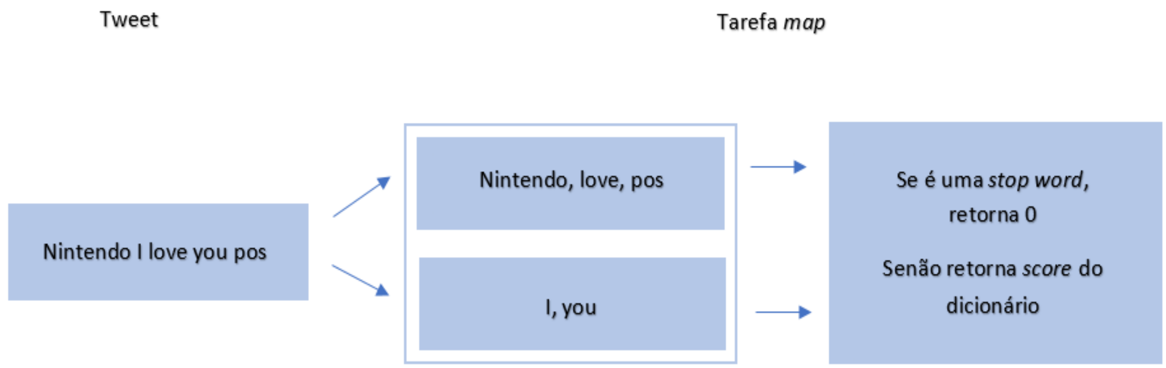
Desses elementos foram extraídos o POS, as pontuações e *synsets* e armazenados em uma estrutura *hashmap*. A pontuação final da palavra é a subtração da pontuação positiva pela negativa. Se uma palavra aparecer mais de uma vez no dicionário, com tipos gramaticais diferentes e, consequentemente, pontuações diferentes, a aplicação desenvolvida soma as pontuações de todos os tipos gramaticais que a palavra têm e as divide pelo número de ocorrências no dicionário léxico. A razão pela qual foi feito esse cálculo foi porque a aplicação desenvolvida não prevê determinar a função da palavra na sentença.

Feito isso, são carregadas em uma lista as palavras vazias, em inglês *stop words.* Elas são palavras que possuem função de conexão em uma sentença, como artigos, pronomes, preposições, etc. Essas palavras, apesar de terem bastante ocorrência nos textos, não afetam o sentimento final da sentença. A lista de *stop words* usada neste trabalho está disponível neste site: http://www.ranks.nl/stopwords. Normalmente esse passo é feito na etapa de pré-processamento, porém ao adaptar a lógica, viu-se que era mais vantajoso fazê-lo na tarefa *map* pois ela depende da palavra estar tokenizada.

[18] http://wordnet.princeton.edu/

A base de dados está organizada com um *tweet* por linha. A aplicação lê uma por vez e a encapsula em um objeto juntamente com a dicionário léxico e a lista de *stop words*. Se somente a linha fosse passada como parâmetro para a função *map*, apenas o nó mestre conseguiria ter acesso ao dicionário e a lista de *stop words*. Sendo assim, passa-se o objeto criado para ocorrer a tokenização e a associação a pontuação em todos os nós da *grid*. Na Figura 19, o processo da tarefa *map* está ilustrado.

Figura 19 – Tarefa *map*

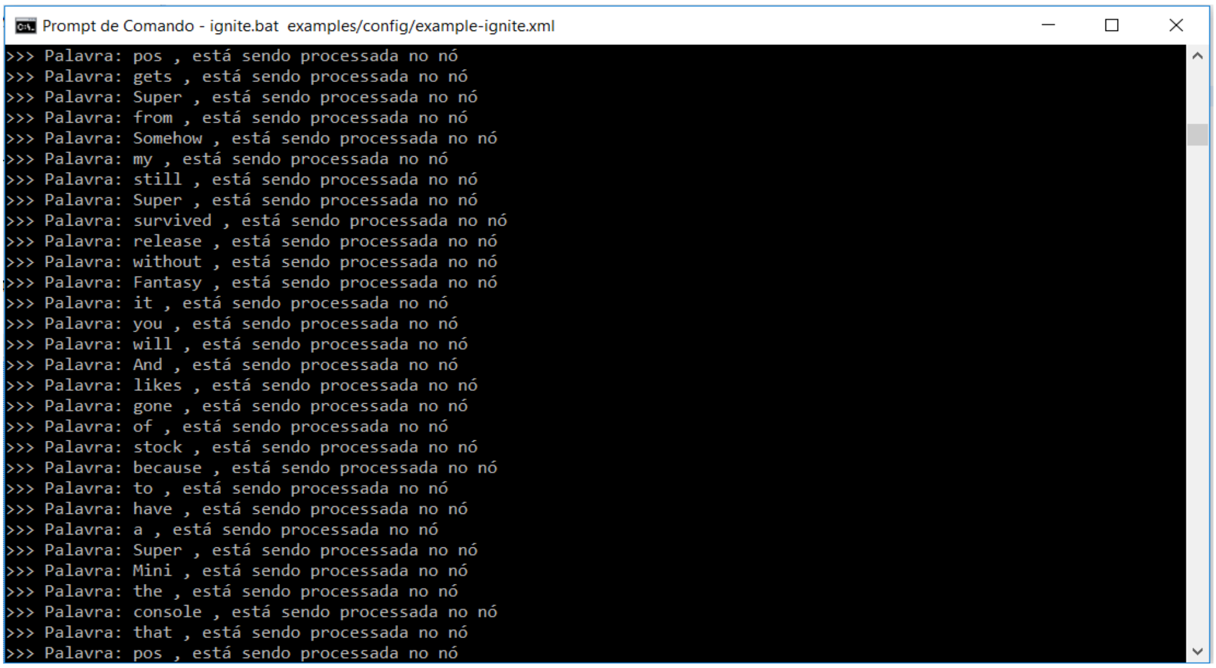


Fonte: COSTA (2017)

O nó mestre então inicia a tarefa *map*, quebrando o *tweet* em palavras individuais (*tokens*) e cria um *job* filho para cada uma dessa palavras e os envia para os nós trabalhadores. Então cada nó irá verificar se a palavra consta na lista de *stop words*. Se sim o nó retorna o valor 0, senão busca a palavra no dicionário e retorna o valor associado a ela. A Figura 20 mostra a saída de processamento de um dos nós.

Uma vez feitas as comparações, cada nó do Apache Ignite retorna uma estrutura *Map* contendo o resultado do processamento do *job*. Esse *Map* é encaminhado para outro nó que vai realizar a etapa de *reduce*. Somente depois que todas as palavras do *tweet* forem processadas é que se realiza a tarefa *reduce*.

Figura 20 – Saída de processamento em um nó



Fonte: COSTA (2017)

3.2.3.2 Classificação

A etapa *reduce* tem como objetivo calcular o sentimento total do *tweet*. Ela recebe como parâmetro o agrupamento dos resultados do processamento realizado nos nós. A equação aplicada (Equação 1) é baseada em uma das metodologias apresentas no trabalho de Kolchyna et al. (2015) para cálculo de polaridade.

$$Score_{AVG} = \frac{1}{m}\sum_{i=1}^{m} Wi \qquad (1)$$

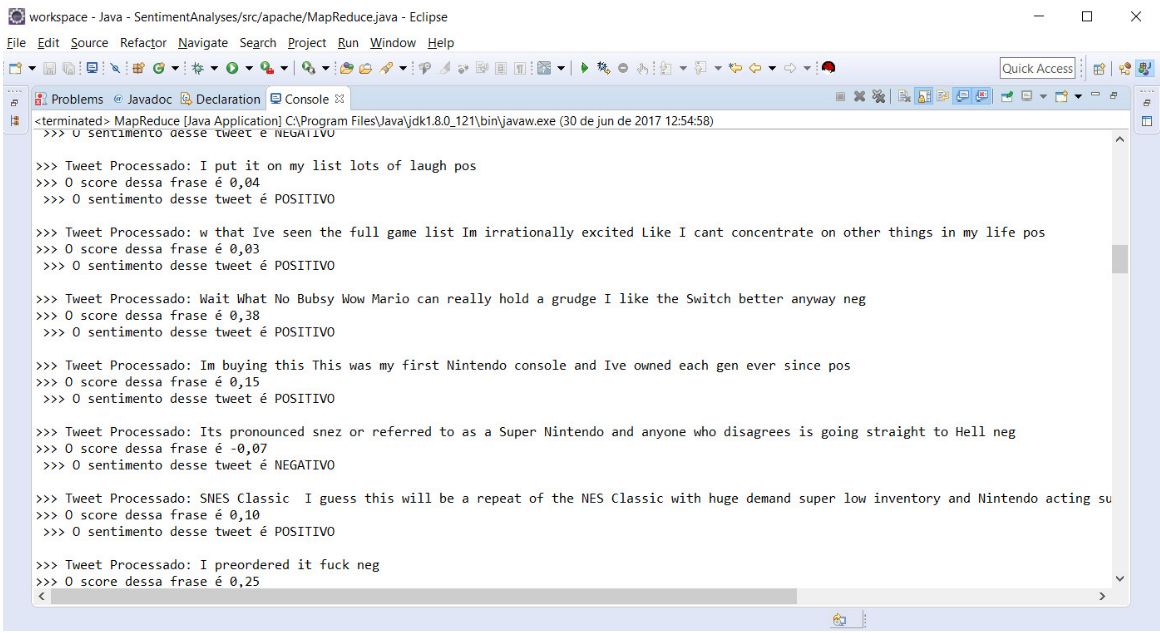
O cálculo da polaridade é dado pela divisão do somatório das pontuações de todas as palavras do *tweet* pelo número de palavras relevantes. Em outras palavras, a soma será dividida apenas pelo número de palavras que contém pontuação diferente de 0. O cálculo destacado abaixo (Equação 2) esclarece que, das quatro palavras que compõe o *tweet* abaixo, apenas "*love*" possui pontuação, "*I*" e "*you*" são *stop words* e Nintendo não possui significado, o que retorna 0.

$$\frac{\text{Nitendo }[0] + \text{I}[0] + \text{love}[0.98] + \text{You}[0]}{1} = 0{,}98 \qquad (2)$$

A tarefa *reduce* é apenas responsável pelo cálculo da pontuação total do *tweet,* portanto a classificação como positiva e negativa é feita pelo nó mestre. O intervalo de classificação considerado por Kolchyna et al. (2015) é entre -1 e 1, sendo que quanto mais próximo de 1 mais positivo e quanto mais próximo de -1 mais negativo. O valor 0 é considerado neutro. O resultado da análise é impresso pelo nó mestre na saída do console do Eclipse, como mostrado na Figura 21. A avaliação dos resultados e da acurácia do método, assim como as conclusões que podem ser inferidas da análise, são discutidas no tópico seguinte.

Assim que o processamento termina, o nó mestre é encerrado e sai da topologia de nós. Os nós trabalhadores não são encerrados e continuam instanciados. Caso a aplicação venha a ser iniciada novamente eles são identificados pelo novo nó mestre.

Figura 21 – Resultado da análise de sentimentos no console do Eclipse

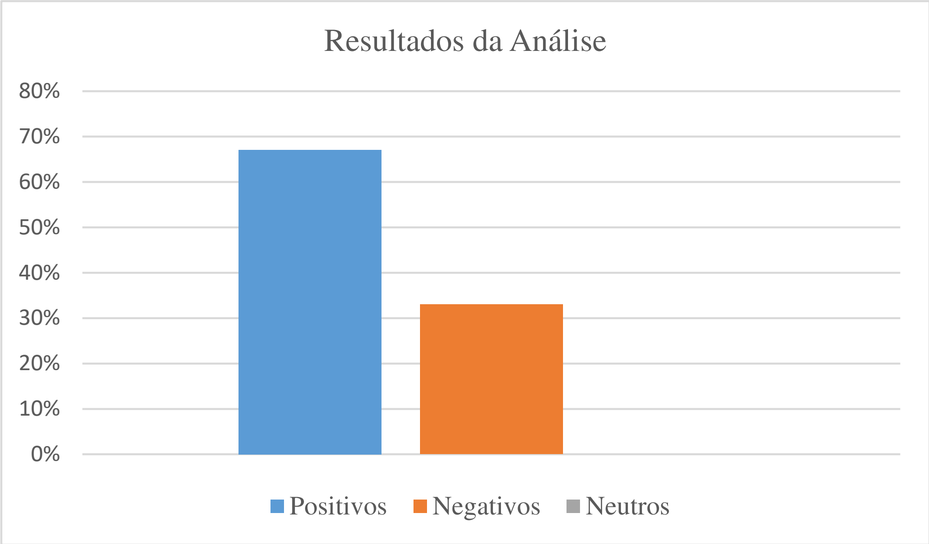


Fonte: COSTA (2017)

3.2.4 Avaliação e Discussão dos Resultados

Nesta etapa é analisado os resultados da análise de sentimentos proposta sobre o anúncio do relançamento do console Super Nintendo. No intuito de compreender os resultados obtidos, são feitas análises individuais de alguns *tweets*. Além de ser levado em consideração o histórico da empresa Nintendo e do console. Para assim poder discorrer sobre a relação dos resultados encontrados com a situação atual da empresa.

Figura 22 – Gráfico de resultados



Fonte: COSTA (2017)

Entre os 94 *tweets* presentes na base, 62 foram classificados manualmente como positivos e 32 como negativos. Os resultados produzidos pela aplicação estão representados em forma de gráfico na Figura 22. Como pode ser observado, a aplicação classificou 63 (67,02%) dos *tweets* como positivo, 31 (32,98%) como negativo e nenhum como neutro. Entretanto, ao comparar com a classificação manual, da quantidade classificada como positivo, apenas 43 eram realmente positivos e 16 negativos.

De modo geral a aplicação classificou corretamente, entre positivos e negativos, 62 (67,39%) dos *tweets*. O número de falsos negativos (*tweets* positivos classificados como negativo) foi de 15 (15,95%). Já a porcentagem de *tweets* classificados falsos positivos (*tweets* negativos classificados como positivos) foi 17 (18,08%). Na Figura 23 estão representados esses números em forma de matriz de confusão.

Figura 23 – Matriz de Confusão

	Positivos	Negativos
Negativos	47	15
Positivos	17	15

Fonte: COSTA (2017)

Na intenção de compreender a classificação feita pela aplicação serão analisados dois exemplos de *tweets* classificados corretamente e dois erroneamente. Na Figura 24, temos um caso onde o *tweet* foi classificado corretamente. Percebe-se que que o dono da opinião (subtópico 2.3.1) não usou gírias e se expressou de forma clara, ou seja, sua opinião é regular e explícita (subtópico 2.3.2). Esse tipo de opinião facilita a identificação do sentimento, o que levou a classificação correta.

Figura 24 – *Tweet* positivo

```
>>> Tweet Processado: Great news because the Super Nintendo is coming back just in time for Christmas pos
>>> O score dessa frase é 0,16
 >>> O sentimento desse tweet é POSITIVO
```

Fonte: COSTA (2017)

Da mesma forma que ocorreu com a classificação do *tweet* positivo da Figura 24, o dono da opinião do *tweet* da Figura 25 foi objetivo e direto. Apesar de estar classificado corretamente, esse *tweet* poderia ter sua negatividade intensificada. O dicionário léxico, apesar de seu vasto número de *synsets*, não consegue lidar com conjugações e plural de palavras. Ao procurar a palavra "*dumbasses*" (plural) o dicionário retorna 0, mesmo tendo a palavra "*dumbass*" (singular). Dito isso deixou-se de contabilizar a pontuação de 0.375.

Figura 25 – *Tweet* negativo

```
>>> Tweet Processado: Whoops guess Im not getting one cause nintendo is a bunch of dumbasses neg
>>> O score dessa frase é -0,17
 >>> O sentimento desse tweet é NEGATIVO
```

Fonte: COSTA (2017)

Observando a Figura 26, temos um *tweet* que é positivo, mas foi classificado pela aplicação como negativo. A razão que levou à má classificação foi que a expressão "*shut up and take my money*", que é uma gíria e que significa que pessoa deseja muito algo independente do preço. As palavras isoladas não carregaram o mesmo sentido. Logo, o dicionário não dá o peso que deveria para a expressão.

Figura 26 – *Tweet* falso negativo

```
>>> Tweet Processado: SHUT UP AND TAKE MY MONEY pos
>>> O score dessa frase é -0,03
 >>> O sentimento desse tweet é NEGATIVO
```

Fonte: COSTA (2017)

Como discutido no subtópico 2.3.3, sentimentos implícitos são difíceis de serem identificados. No *tweet* da Figura 27 está um exemplo disso. Nenhuma das palavras do texto é negativa, mas mesmo assim, ao classificar manualmente, é possível notar o sentimento de frustação do dono da opinião do *tweet* em relação a Nintendo. Todavia a aplicação não consegue inferir a mesma coisa.

Figura 27 – *Tweet* falso positivo

```
>>> Tweet Processado: produces enough to keep up with demand this time neg
>>> O score dessa frase é 0,01
 >>> O sentimento desse tweet é POSITIVO
```

Fonte: COSTA (2017)

Ao se tratar de *tweets* negativos a análise de sentimentos torna-se mais trabalhosa. Na aplicação desenvolvida o problema da negação (subtópico 2.3.3), em inglês *negation handling,* não é tratado. Segundo Kolchyna et al. (2015), um dos métodos para tratar negações, é inverter a polaridade das palavras que se encontram antes de uma palavra de negação. Na aplicação desenvolvida o nó mestre envia as palavras fora de ordem para os nós trabalhadores e esses não guardam histórico das palavras processadas por ele, impossibilitando inverter a polaridade. Contudo, mesmo sem o tratamento das negações, a aplicação conseguiu identificar *tweets* negativos. Além do mais não influencia o objetivo do trabalho que é demonstrar a aplicação do *framework* Apache Ignite juntamente com o modelo de programação MapReduce na tarefa de análise de sentimentos.

A baixa acurácia é devido a técnica escolhida para a metodologia proposta. Análises léxicas, quando feitas em textos onde há muita linguagem informal, tendem a não serem muito eficazes. No Twitter, os usuários usam muitas gírias, cometem muitos erros ortográficos, além de se expressarem através de *emoticons*. A metodologia proposta é uma análise léxica simples que poderia, se combinada com o uso de um dicionário léxico de gírias e *emoticons*, vir a apresentar resultados mais próximos do da análise manual.

Levando em consideração a análise de sentimentos feita, o anúncio do relançamento do Super Nintendo foi, no geral, bem recebido pelos usuários do Twitter. A Nintendo é uma das empresas pioneiras em jogos eletrônicos, se consagrou no mercado nas décadas de 80 e 90, com lançamento do Super Nintendo e de jogos como *Legends of Zelda* e, principalmente, *Super Mario World*, que já teve mais de 20 milhões de cópias vendidas no mundo todo (WIKIPEDIA, 2017).

Com histórico e altos e baixos a Nintendo teve uma década difícil. Segundo Bogost (2016) há dez anos, a empresa estava no auge do sucesso comercial dos seus produtos Wii e DS. Os preços das suas ações estavam altos até 2007, mas os controles remotos do Wii se mostraram ser uma tendência de curta duração. Em 2012, a Nintendo lançou uma nova versão do Wii, o Wii U, mas o console recebeu muitas críticas e desapontou os consumidores, ocasionando o uma queda de 80% do valor da empresa ao final de 2012. Entre 2012 e 2016 o cenário não mudou e não conseguiram melhorar. O lançamento Pokémon Go em julho de 2016 deu algum impulso, mas em meados de novembro do mesmo ano o jogo já havia perdido mais de 60% de seus usuários. Ao relançarem o Super Nintendo, eles apostam em reviver nos usuários sentimentos como nostalgia e felicidade, para assim se reerguerem no mercado. Dessa forma, a análise de sentimentos pode ser uma grande aliada, pois tendo o *feedback* rápido do cliente é possível atacar os erros rapidamente e evitar novos fracassos.

3.2.5 Comparação com Trabalhos Relacionados

Dos trabalhos relacionados apresentados no Capítulo 1, os métodos apresentados por Kolchyna et al. (2015), Moreira et al. (2016) e Oliveira (2013) utilizam a mesma técnica e nível de análise de sentimentos que adotados na metodologia proposta. Os trabalhos realizados por Kolchyna et al. (2015) e Oliveira (2013) inclusive utilizam *tweets* para a classificação de sentimento em nível de sentença. Ambos apresentaram baixa acurácia na classificação dos *tweets*. Entretanto nenhum deles aplicaram o processamento dos dados em paralelo, sendo necessária a adaptação dos métodos para funcionar em um ambiente distribuído.

A análise de sentimentos conduzida por Moreira et al. (2016) utiliza a técnica de dicionário léxico para análise de sentimentos de *posts* do Facebook. O autor utiliza três diferentes dicionários léxicos para comparação dos resultados com análise manual. Uma amostra de textos foi coletada da base e, assim como foi feita na metodologia proposta neste trabalho, foram classificados a mão. Apesar de se tratar de fontes de dados diferentes, as conclusões quanto as limitações da análise de sentimentos coincidiram em alguns pontos: limitação dos dicionários na classificação palavra e limitação ao identificar intensidade. Logo, apenas o uso do dicionário léxico não é suficiente para uma boa análise de sentimentos. Ele deve ser combinado com um dicionário que contenha regionalismos, gírias e abreviações e que estejam sendo constantemente atualizado. Além de adaptar o dicionário léxico a variações de uma mesma palavra.

O método proposto neste trabalho alcançou as mesmas conclusões que o trabalho de Sousa (2014), quanto a adaptação do modelo de programação MapReduce e a utilização de um *framework* para processamento paralelo na tarefa de análise de sentimentos. As metodologias se divergem quanto ao nível de análise proposto e fontes de dados diferentes. Independentemente, comprovou-se que o processamento em paralelo e distribuído dos dados são uma nova abordagem para análise de sentimentos. A vantagem deste trabalho é que ele apresentou uma aplicação adaptável tanto ao hardware quanto a plataforma. O baixo custo de investimento e manutenção, permite instituições de todo porte aplicar a metodologia na avaliação e estudo de melhorias de produtos ou serviços.

Neste capítulo foi detalhado todas as etapas, características e limitações da aplicação desenvolvida para o estudo de caso. Assim com os resultados obtidos e o significado deles para a empresa Nintendo.

No próximo capítulo são feitas as conclusões acerca da monografia desenvolvida, levantando as limitações encontradas e sugestões para trabalhos futuros.

4 CONCLUSÃO

Este trabalho teve como objetivo o desenvolvimento de uma metodologia para análise de sentimentos em nível de sentença a partir de dados extraídos do Twitter. Um ponto a destacar deste para outros trabalhos da área, foi o uso do *framework* Apache Ignite e da incorporação do modelo de programação MapReduce na tarefa de análise de sentimentos. No intuito de exemplificar seu funcionamento, foi desenvolvida uma aplicação que utilizou o método baseado em léxico para determinar a polaridade de *tweets* sobre o lançamento do Super Nintendo.

Durante o desenvolvimento deste trabalho, foi visto que a análise de sentimentos de *tweets* não é uma tarefa simples. Além de apresentar os desafios da análise de sentimentos discutidos no subtópico 2.3.3, erros ortográficos dos usuários, ambiguidades das frases assim como limitações do método léxico, se mostraram empecilhos para uma classificação mais precisa. Por consequência, a análise realizada apresentou acurácia baixa. Todavia, o método empregado de análise cumpriu seu papel de demonstrar a o uso do *framework* Apache Ignite.

A utilização de uma ferramenta que permita a utilização de uma estrutura existente implica em adaptá-la a diferentes ambientes e/ou estrutura de *hardwares*. Toda a aplicação desenvolvida, graças ao *framework* Apache Ignite, permite que o processamento seja realizado de forma paralela e distribuída em diferentes plataformas, utilizando o *hardware* disponível, sem prejudicar a realização da análise de sentimentos.

Apesar das dificuldades encontradas, pode-se concluir que a metodologia se mostrou eficaz, pois ao incorporar essas ferramentas na análise de sentimentos foi visto que ela garante o processamento de grande volume de dados de uma forma simples e distribuída. Além de mostrar que é possível utilizar o Twitter como fonte de dados, e extrair dele informações relevantes para o processo de tomada de decisões de empresas e organizações, até mesmo em tempo real, utilizado algoritmos e técnicas de *Big Data*.

Em trabalhos futuros, sugere-se resolver, principalmente, o problema da negação em aplicações baseadas no MapReduce. Expandir a metodologia para processamento de informações de múltiplas redes sociais. Além de aplicar a metodologia proposta em *tweets* em português, traduzindo e ampliando o dicionário léxico.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMASHRAE, M.; MONETT, D.; PASCHKE, A. **Emotion Level Sentiment Analysis: The Affective Opinion Evaluation.** 2016. Disponível em: <http://ceur-ws.org/Vol-1613/paper_1.pdf>. Acessado em: 29 mai. 2017.

APACHE HADOOP. **Documentação**. 2017. Disponível em: <http://hadoop.apache.org/core>. Acessado em: 12 de mai. 2017.

APACHE IGNITE. **Documentação**. 2017. Disponível em: <https://ignite.apache.org/>. Acessado em: 08 de mai. 2017

ARAUJO, G. et al. **Análise de sentimentos sobre temas de saúde em mídia social**. Journal of Health Informatics (JHI). São Paulo, v. 4, n. 3, p. 95-99. Jul.- Set. 2012.

BECKER, K.; TUMITAN, D. **Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios.** In: LECTURES OF THE 28TH BRAZILIAN SYMPOSIUM ON DATABASES. 1ed. Pernambuco: CIN - UFPE, 2013, v. , p. 27-52.

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. **Métodos para Análise de Sentimentos em mídias sociais**. 2015. Disponível em: <http://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>. Acessado em: 09 jun. 2017.

BOGOST, I. **Nintendo's Sad Struggle for Survival**. 2016. Disponível em: <https://www.theatlantic.com/technology/archive/2016/12/super-marios-sorrow/511187/>. Acessado em: 30 jun. 2017.

BONGIRWAR, V. A Survey on Sentence Level Sentiment Analysis**. International Journal Of Computer Science Trends And Technology (IJCTT**). India, v. 3, n. 3. p. 110-113. Mai.-Jun. 2015.

CAVANILLAS, J.; CURRY, E.; WOLFGANG, W. **New Horizons for a Data-Driven Economy:** A Roadmap for Usage and Exploitation of *Big Data* in Europe. Heidelberg: Springer Nature, 2016.

CHEN, M. et al. **Big Data: Related Technologies, Challenges and Future Prospects**. Briefs in Computer Science. Springer International Publishing. 2014.

CISCO VISUAL NETWORKING INDEX. **The Zettabyte Era: Trends and Analysis**. 2017. Disponível em: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html>. Acessado em: 20 de mar. 2017.

CURRY, E. et al. **Public deliverable of the EU-Project BIG**. 2012. Disponível em: <https://big-project.eu/sites/default/files/BIG_D2_2_2.pdf >. Acessado em:06 jun. 2017.

DEAN, J.; GHEMAWAT, S. **MapReduce: Simplified Data Processing on Large Clusters**. 2004. Disponível em: <http://research.google.com/archive/mapreduce-osdi04.pdf>. Acessado em 06 mai. 2017.

DEMCHENKO, Y. et al. **Addressing Big Data Issues in Scientific Data Infrastructure**. 2013. Disponível em: <https://tnc2013.terena.org/includes/tnc2013/documents/bigdata-nren.pdf >. Acessado em: 16 mai. 2017.

DERN, D. **GridGain In-Memory Data Fabric Becomes Apache Ignite**. 2015. Disponível em: <https://www.linux.com/news/gridgain-memory-data-fabric-becomes-apache-ignite>. Acessado em: 10 Abr. 2017.

DICIONÁRIO do Aurélio. 2017. Disponível em: <https://dicionariodoaurelio.com/>. Acessado em: 08 jun. 2017.

DUBEY, V.; GUPTA, D. **Sentiment analysis using Singular Value Decomposition**. International Journal of Current Engineering and Technology. India, v. 6, n. 4. Ago. 2016.

FARRA et al. **Sentence-level and Document-level Sentiment Mining for Arabic Texts**. IEEE International Conference on Data Mining Workshops. 2010.

FERILLI et al. **Towards Sentiment and Emotion Analysis of User Feedback for Digital Libraries**. 2016. Disponível em: <https://www.researchgate.net/profile/Esposito_Floriana/publication/293482613_Towards_Sentiment_and_Emotion_Analysis_of_User_Feedback_for_Digital_Libraries/links/5774e96408ae1b18a7df9af0/Towards-Sentiment-and-Emotion-Analysis -of-User-Feedback-for-Digital-Libraries.pdf >. Acessado em: 04 abr. 2017.

FRANÇA, T. C.; OLIVEIRA, J. Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. In: III BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 3, 2014, Brasília. **Anais do Congresso da Sociedade Brasileira de Computação**. Porto Alegre: SBC. 2014. p. 128-139.

G1. **Facebook atinge marca de 1 bilhão de usuários todos os dias**. 2016. Disponível em: <http://g1.globo.com/tecnologia/noticia/2016/04/facebook-atinge-marca-de1-bilhao-de-usuarios-todos-os-dias.html>. Acessado em: 23 de mar. 2017.

G1. **Super Nintendo será relançado em setembro com 21 jogos na memória**. 2017. Disponível em: <http://g1.globo.com/tecnologia/games/noticia/super-nintendo-sera-relancado-em-setembro-com-21-jogos-na-memoria.ghtml>. Acessado em: 29 jun. 2017.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. In: **International Journal of Information Management**. Toronto, v. 35, n. 2, p. 137-144. Abr. 2015.

GODSAY, M. **The Process of Sentiment Analysis: A Study**. International Journal of Computer Applications. v. 126, n. 7. Set. 2015.

GRAÇA NETO, A. **Sentimentalista: Um framework para análise de sentimentos baseado em processamento de linguagem natural**. 2016. 136 f. Dissertação (Mestrado em Ciência e Tecnologia da Computação) – Universidade Federal de Itajubá, Itajubá, 2016.

GRIDGAIN. **Documentação**. 2017. Disponível em: <http://www.gridgain.org>. Acessado em: 17 mai. 2017.

HU et al. **Toward Scalable Systems for Big Data Analytics: A Technology Tutorial**. IEEE Access, v. 2, p. 652-687, Mai. 2014.

IDC. Brasil. **Previsão da IDC para o mercado de TIC no Brasil em 2016 aponta crescimento de 2,6%**. 2016. Disponível em: <http://br.idclatin.com/releases/news.aspx?id=1970>. Acesso em: 25 de mai. 2017.

IDC. GANTZ, J.; REINSEL, D. **The Digital Universe In 2020: Big Data, Bigger Digital Shadows, And Biggest Growth In The Far East**. IDC iView, pp 1–16. 2012. Disponível em:<https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-020.pdf>. Acessado em: 12 jun. 2017.

IRMA. **Big Data: Concepts, Methodologies, Tools and Applications.** IGI Global. USA: 2016.

IVANOV, N. **Fire up big data processing with Apache Ignite**. 2016. Disponível em: <http://www.infoworld.com/article/3135070/data-center/fire-up-big-data-processing-with-apache-ignite.html>. Acessado em: 03 mai. 2017.

JAIN, A. **The 5 Vs of Big Data**. 2016. Disponível em: < https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/ >. Acessado em: 16 mai 2017.

JUREK, A.; MULVENNNA, M.; BI, Y. **Improved lexicon-based sentiment analysis for social media analytics**. Spring Open. Security Informatics. Dez. 2015.

KAUER, A**. Análise de Sentimentos baseada em Aspectos e Atribuição de Polaridade**. 2016. 76 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal Do Rio Grande Do Sul, Porto Alegre, 2016.

KOLCHYNA, O. et al. **Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination**. 2015. Disponível em: <https://arxiv.org/pdf/1507.00955.pdf>. Acessado em: 22 nov. 2016.

KOLKUR, S.; DANTAL, G., MAHE, R. **Study of Different Levels for Sentiment Analysis.** In: International Journal of Current Engineering and Technology, v.5, n.2. Abr. 2015.

LEE, K. **The Definitive List of Social Media Acronyms and Abbreviations**. 2015. Disponível em: <https://blog.bufferapp.com/social-media-acronyms-abbreviations> Acesso em: 16 jun. de 2017.

LIU, B.**Sentiment analysis: Mining opinions, sentiments, and emotions.** Cambridge University Press. 2015.

LIU, B. **Sentiment Analysis and Opinion Mining.** Morgan & Claypool Publishers. 2012.

MADHOUSHI, Z.; HAMDAN, A.; ZAINUDIN, S. **Sentiment Analysis Techniques in Recent Works**. Science and Information Conference. Jul. 2015.

MANYIKA, J. et al. **Big Data: The Next Frontier for Innovation, Competition and Productivity**, McKinsey & Company, 2011. Disponível em: <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>. Acesso em: 17 de Mai. 2017.

MARR, B. **Big Data: The 5 Vs Everyone Must Know**. Disponível em: < https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>. Acessado em: 16 Mai. 2017.

MEDHAT, W.; HASSAN, A; KORASHY, H. **Sentiment analysis algorithms and applications: A survey**. Ain Shams Engineering Journal. 2014.

MILLER, H.; MOKER, P. **From Data to Decisions: A Value Chain for Big Data**. IT Professional, IEEE, v. 15, n. 1, p. 57-59. Jan.-Feb.2013.

MOENS, M-F.; LI, J.; CHUA, T-S. **Mining User Generated Content**. Chapman & Hall. 2014.

MOREIRA, V. et al. Análise de Sentimentos: Comparando o uso de ferramentas e a análise humana. In: SYMPOSIUM ON INFORMATION SYSTEMS, 12., 2016, Florianópolis.**Proceedings of the XII Brazilian Symposium on Information Systems**. Florianópolis: UFSC. SBC. 2016. p. 441-448.

OLIVEIRA, F. **Análise de Sentimentos de comentários em Português Utilizando SentiWordNet**. 2013. 45 f. Trabalho de Especialização em Desenvolvimento de Sistemas para Web – Universidade Estadual de Maringá, Maringá, 2013.

PAIVA, R. **O que é e como funciona o Map Reduce usado pelo Google**. 2011. Disponível em: <http://blog.werneckpaiva.com.br/2011/08/como-funciona-o-map-reduce-usado-pelo-google/>. Acessado em: 22 Abr. 2017.

PANG et al. **Thumbs up?: Sentiment Classification Using Machine Learning Techniques**. Proceedings Of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). 2002.

PANG, B.; LEE, L. **A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts**.2004. Disponível em: < http://www.cs.cornell.edu/home/llee/papers/cutsent.pdf>. Acessado em 19 mai. 2017.

PANG, B.; LEE, L. **Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval**. v. 2. p. 1–135. 2008.

PATEL, V.; PRABHU, G.; BHOWMICK, K. **A Survey of Opinion Mining and Sentiment Analysis**. International Journal of Computer Applications, v. 131, n.1. Dez. 2015.

RICE, D.; ZHORN, C. **Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies**. 2013. Disponível em: <http://www.kenbenoit.net/pdfs/NDATAD2013/Rice-Zorn-LSE-V01.pdf >. Acessado em: 21 jun. 2017.

SAGIROGLU, S.; SINANC, D. Big data: A review. In: **International Conference on Collaboration Technologies and Systems (CTS)**. San Diego, p. 42-47. Mai. 2013.

SANTOS, W. **Análise dos Tweets sobre a Black Friday através da Mineração de Texto e Análise de Sentimentos**. 2016. 51 f. Monografia (Graduação em Sistemas de Informação) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

SCHAUWEN, S. **Machine Learning Approaches to Sentiment Analysis using the Dutch Netlog Corpus**. 2010. Disponível em: <http://www.clips.ua.ac.be/sites/default/files/ctrs-001-small.pdf>. Acessado em: 21 jun. 2017.

SCHROECK, M. et al. **Analytics: The real-world use of big data:** How innovative enterprises extract value from uncertain data. 2012. Disponível em: <https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=GBE03519USEN>. Acessado em: 25 mai. 2017.

SILVA, 2013. **Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo**. 2013. 67 f. Monografia (Graduação em Ciência da Computação) – Universidade Federal de Brasília, Brasília, 2013.

SILVA JUNIOR, S. **Aplicação de Recuperação de Informação e Análise de Sentimentos para Suporte à Pesquisas de Mercado**. 2015. 65 f. Monografia (Graduação em Sistemas da Informação) – Universidade Federal Rural de Pernambuco, Recife, 2015.

SOUSA, G. **Tweetmining: Análise De Opinião Contida Em Textos Extraídos Do Twitter**. 2012. 66 f. Monografia (Graduação em Sistemas de Informação) – Universidade Federal de Lavras, Lavras, 2012.

SOUSA, J. **Big Data: Análise de Sentimento em Dados de Pesquisa de Opinião utilizando o Framework GridGain e Processamento em Memória**. 2014. 76 f. Monografia (Graduação em Ciência da Computação) – Universidade Federal do Maranhão, São Luís, 2014.

TABOADA, M. et al. **Lexicon-Based Methods for Sentiment Analysis**. Computational Linguistics. v. 37, n. 2. 2011.

THE ECONOMIST. **The world's most valuable resource is no longer oil, but data**. 2017a. Disponível em: <http://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource>. Acessado em: 14 de mai. 2017.

THE ECONOMIST. **Data is giving rise to a new economy**. 2017b. Disponível em: <http://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy>. Acessado em: 13 mai. 2017.

TURNEY, P. **Thumbs up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews**. Proceedings of Annual Meeting of the Association For Computational Linguistics (ACL-2002). 2002.

TWITTER. **Documentação**. 2017. Disponível em: <https://dev.twitter.com/docs>. Acessado em: 28 jun. 2017.

WARD, J.; BARKER, A. **Undefined By Data: A Survey of Big Data Definitions**. 2013. 2 f. University of St Andrews, Reino Unido. 2013.

WIKIPEDIA. **Super Mario World**. 2017. Disponível em: < https://pt.wikipedia.org/wiki/Super_Mario_World>. Acessado em: 30 jun. 2017.

ZHANG et al. **Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis**. 2011. Disponível em: < http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf>. Acessado em: 26 jun. 2017.

ZIKOPOULOS et al. **Harness the Power of Big Data**: The IBM Big Data Plataform. The McGraw-Hill Companies, 2012.