ISCTE IUL
Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias da Informação

Categorização e Classificação de Notícias de *Big Data* em Tecnologias segundo o Quadrante Mágico de *Gartner*

João Miguel Ferreira Canito

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Informática e Gestão

Orientador:

Doutor Sérgio Moro, Professor Auxiliar,

ISCTE-IUL

Coorientador:

Doutor Paulo Rita, Professor Catedrático,

ISCTE *Business School*

Setembro, 2017

“Believe you can and you're halfway there.”

Theodore Roosevelt

Agradecimentos

Se estamos a ler esta página significa que alcancei o fim desta etapa. Foram bastantes horas de dedicação, muito sacrifício e esforço, mas se consegui deve-se a algumas pessoas que me acompanharam neste percurso e merecem um agradecimento para a posterioridade.

Gostaria de agradecer ao meu orientador, coorientador e outro Professor, respetivamente. Obrigado Professor Sérgio Moro por toda a sua disponibilidade, ajuda e preciosos conselhos, sem si tinha sido mais difícil. Obrigado Professor Paulo Rita pelo seu incentivo e sugestões durante todo o trabalho. Obrigado Professor Pedro Ramos pela sua recetividade e apoio.

Agradeço também aos meus amigos e entre queridos pelo apoio e compreensão, por não poder estar presente em muitos convívios, por ter recusado muitos convites e pela minha maior ausência durante este período.

Por fim, gostaria de agradecer especialmente há minha família, mãe, pai e avó. Obrigado pelo constante encorajamento, sem a vossa valiosa ajuda não teria conseguido. Vocês são a essência da pessoa que sou hoje. Obrigado por tudo. A vocês dedico este trabalho.

Resumo

O desenvolvimento das tecnologias nos últimos anos levou a um aumento contínuo de dados e sua acumulação a uma velocidade incalculável. Todos estes fatores acima mencionados levaram à banalização de um novo conceito: *Big Data*.

Neste estudo foram extraídas 11 505 notícias sobre *Big Data* do *Google News* e foram aplicadas técnicas de *Text Mining* de forma a obter conhecimento relevante e uma categorização noticiosa, através de *Latent Dirichlet Allocation*. São abordadas as Tecnologias *Big Data* relativamente aos Quadrantes de *Gartner* de forma a perceber o tipo de Tecnologias em que as empresas de um Quadrante específico investem. Desta forma, este estudo tem uma contribuição interessante para a literatura, pois fornece resultados concretos sobre o comportamento do mercado, provenientes de dados factuais.

Este estudo comprova a força das empresas integrantes do Quadrante de Gartner *leaders*, revelando que estas são cada vez mais líderes de mercado, apresentando uma solução muito completa e diversificada de Tecnologias *Big Data*. É também demonstrado que as empresas que integram o Quadrante de *Gartner challengers* não demonstram entendimento sobre a direção em que o mercado se desloca e que uma empresa que pertença ao Quadrante de *Gartner visionaries*, caso aposte fortemente na Tecnologia *Big Data stream analytics* terá a sua posição alterada no Quadrante de *Gartner*, aproximando-se cada vez mais do Quadrante *leaders* e, ao mesmo tempo, do Quadrante *niche players*.

Palavras-Chave: *Big Data News, Gartner Magic Quadrant, Big Data Technologies; Text Mining, Topic Model.*

Abstract

The development of technologies in recent years has led to a continuous increase in data, and its accumulation at an incalculable speed. All these factors mentioned above have led to the trivialization of a new concept: Big Data.

In this study 11505 Google News Big Data news were extracted and Text Mining techniques were applied to obtain relevant knowledge and a news categorization through the *Latent Dirichlet Allocation* algorithm. Big Data Technologies are approached relatively to the Gartner Quadrants in order to perceive the type of Technologies wherein companies of a specific Quadrant invest. Thus, this study has an interesting contribution to the literature, since it provides concrete results on the market behavior, coming from factual data.

This study proves the strength of the Gartner leaders quadrant, revealing that they are increasingly market leaders, presenting a very complete and diverse Big Data Technology solution. It is also demonstrated that the companies in the challengers Gartner Quadrant do not demonstrate understanding of the direction the market is moving and that a company belonging to the visionaries Gartner Quadrant betting strong on Big Data stream analytics technology will have is position modified in the Gartner Quadrant, increasingly approaching the Leaders Quadrant and, at the same time, the niche players Quadrant.

Keywords: Big Data News, Gartner Magic Quadrant, Big Data Technologies; Text Mining, Topic Model.

Índice

Índice de Figuras

Índice de Tabelas

Termos e Abreviaturas

BI = *Business Intelligence*

ML = *Machine Learning*

LDA = *Latent Dirichlet Allocation*

QG = Quadrante de *Gartner*

QSG = Quadrantes de *Gartner*

TBD = Tecnologia Big Data

TSBD = Tecnologias Big Data

TM= *Text Mining*

1. Introdução

1.1) Surgimento de *Big Data*

Desde o início do século XXI, muitas foram as mudanças tecnológicas que ocorreram na indústria da tecnologia e informação, tais como: *social networking*, *cloud computing* e *Internet of Things.* O desenvolvimento destas tecnologias disruptivas levou a um aumento contínuo de dados e sua acumulação a uma velocidade incalculável. Todos estes fatores acima mencionados levaram à banalização de um novo conceito: *Big Data* (Meng, 2013). Esta área de investigação emergiu num diverso espectro de inovações tecnológicas e oportunidades permitidas pela revolução da informação. As expectativas de que o *Big Data* conduzirá a sociedade atual a uma nova e cativante era de inovações são elevadas (Goes, 2014).

A expressão *Big Data* tem um significado diferente dependendo do seu interlocutor. As definições sobre o termo evoluíram rapidamente, o que gerou alguma confusão. Segundo um questionário *online* realizado em 2012 pela empresa Harris Interactive, em nome da SAP (SAP, 2012), podemos constatar na Figura 1 os diferentes tipos de significados de *Big Data* para os vários executivos em causa. Algumas definições focam-se em tentar identificar as suas características, enquanto outras tendem a considerar as suas potencialidades. O tamanho é a primeira característica evidenciada quando pensamos na questão "o que é *Big Data?"* porém, outras características do termo emergiram mais recentemente (Gandomi & Haider, 2015).

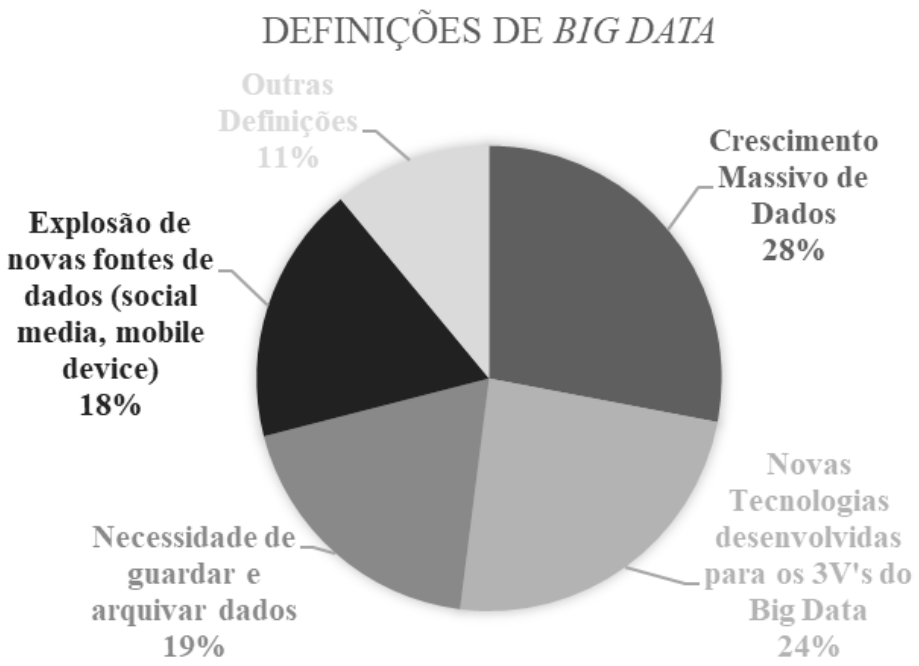

Figura 1 *- Definições de Big Data baseadas num questionário global de executivos.*

Fonte: Adaptado de Gandomi & Haider (2015).

1.2) *Buzzwords* relacionadas

Há vários outros conceitos que se relacionam com *Big Data*, incluindo diversas *buzzwords*, a maioria antecessores ao termo, como as que se seguem:

- BI é um termo abrangente que inclui infraestruturas e ferramentas, aplicações e melhores práticas que permitem o acesso e análise de informações para otimizar decisões e desempenho (Gartner IT Glossary, n.d.).
- *Business Analytics* são um conjunto de soluções usadas para construir modelos de análise e simulações para criar cenários, compreender as realidades e prever os estados futuros. *Business Analytics* inclui *Data Mining*, *Predictive Analytics*, análises aplicadas e estatísticas, e é entregue como uma aplicação adequada para um utilizador do negócio (Gartner IT Glossary, n.d.).
- *Data Mining* é o processo de descobrir correlações significativas, padrões e tendências, através de grandes quantidades de dados armazenados em repositórios. O *Data Mining* utiliza tecnologias para reconhecer padrões, assim como, técnicas estatísticas e matemáticas (Gartner IT Glossary, n.d.).
- *Predictive Analytics* é um método de análise avançada que examina os dados ou conteúdo para responder à pergunta "O que vai acontecer?", ou, mais precisamente, "O que é provável que aconteça?", e é caracterizada por técnicas como a análise de regressão, previsão, estatística multivariada, teste de correspondência padrão e modelação preditiva (Gartner IT Glossary, n.d.). Outra forma de quantificar este termo é descrever qualquer abordagem de *Data Mining* com quatro atributos (Gartner IT Glossary, n.d.):

    1. Uma ênfase de previsão (em vez de descrição, classificação ou agrupamento);
    2. Uma análise rápida medida em horas ou dias (em vez de meses, como normalmente em *Data Mining*);
    3. Ênfase na relevância negocial dos conhecimentos recolhidos;
    4. Ênfase na facilidade de uso, tornando as ferramentas acessíveis a utilizadores empresariais.

Outra das razões para o sucesso de *Big Data* prende-se com o facto de as ferramentas que existem para a sua análise estarem a tornar-se cada vez mais acessíveis. A Teradata e a IBM, dois gigantes tecnológicos no que concerne ao tratamento de dados, trabalham há mais de uma década em conjunto com o intuito de ajudar as empresas a transformar dados em conhecimento valioso para o negócio, levando melhores e mais rápidas decisões a serem tomadas.

A maioria dos investigadores têm-se focado na análise de *reviews* de produtos que estão, na sua maioria, divididos em *reviews* positivas e *reviews* negativas, ajudando as pessoas no julgamento e seleção do produto (Dave *et al*., 2003). Adicionalmente, existem estudos que também abordam análises preditivas relativamente ao mercado de ações, porém, e com base nas pesquisas feitas e critérios aplicados, não existe um trabalho que se baseie na análise e extração de conhecimento de notícias de *Big Data*, num intervalo específico de tempo, sendo a fonte de notícias o *Google News*.

1.3) Contribuições esperadas

Com este trabalho e suas contribuições é pretendido que, dentro do leque de notícias de *Big Data* recolhidas, seja possível:

- Providenciar uma forma alternativa de categorização de notícias através do tipo de TBD e do QG, sumarizando em tópicos de interesse um volume elevado de notícias sobre o tema "*Big Data*" a partir de um período de tempo específico através de técnicas de ML;
- Extração de conhecimento oculto no texto formal das notícias;
- Perceber o mercado com base no posicionamento das empresas que compõem cada QG, face aos *softwares* integrantes de cada TBD;
- Alterar a forma percetual de como um tema como o *Big Data* é abordado e comunicado nas fontes noticiosas.

2. Revisão Literária

2.1) Emergência da Era *Big Data*

Recentemente, tem-se verificado uma mudança de paradigma: ao longo de décadas a Oracle, IBM e Teradata têm implementado soluções de *Data Warehouse* a milhares de empresas com uma capacidade que, na atualidade, ronda a ordem de grandeza dos *terabytes* (International Technology Group, 2011). Porém existe agora uma tendência emergente para o *Big Data* ser armazenado entre múltiplos servidores que estão preparados para lidar com dados não estruturados e facilmente serem escaláveis à medida das necessidades. Isto deve-se ao incremento no uso de tecnologias como o Hadoop, um software *open-source* resultante de um projeto da Apache Software Foundation, que permite um processamento distribuído de grandes conjuntos de dados aglomerados em *commodity servers*. Foi criado de forma a expandir-se a partir de um único servidor para milhares de máquinas e tem um grau elevado de tolerância a falhas. Assim, o Hadoop permite de forma mais eficaz que a análise ocorra no local onde os dados residem, mas requer aptidões específicas e não é uma tecnologia de fácil adoção (Joshi, 2015).

Como podemos constatar pela Figura 2, com o passar dos anos o termo *Big Data* começou cada vez mais a ser utilizado em documentos científicos. A partir do ano 2011, deu-se um aumento exponencial no uso escrito das palavras "*Big Data*". Este conceito começou a ser cada vez mais utilizado no nosso dia a dia, pois é uma expressão comummente banalizada pelos *mass media*, porém tipicamente não é feita uma descrição para que possamos contextualizar o termo. Foi sugerido que Volume, Variedade e Velocidade (ou os três V's) fossem as três dimensões de desafios na gestão de dados (Laney, 2001). Os três V's emergiram como uma estrutura comum para descrever *Big Data* (Chen *et al.*, 2012). Assim, *Big Data* representa grande-volume, elevada-velocidade e elevada-variedade de dados que necessitam de formas eficientes e inovadoras de processamento da informação para que seja extraído conteúdo de apoio à decisão e de automatização de processos (Gartner IT Glossary, n.d.).

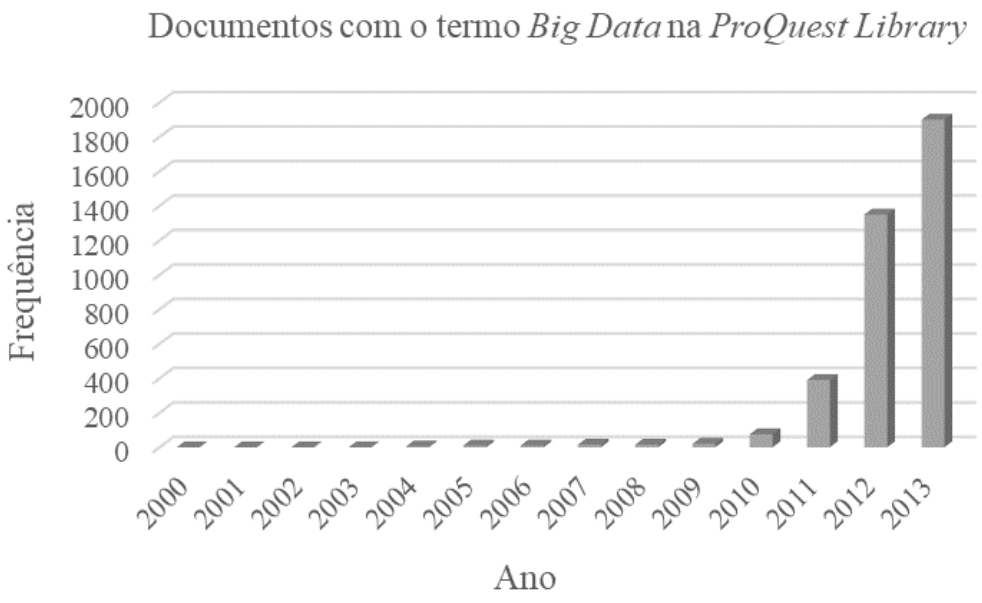

Figura 2 *- Frequências de documentos com o termo Big Data na ProQuest Research Library.*

Fonte: Adaptado de Gandomi & Haider, 2015.

Outra das definições de *Big Data* refere que é um termo que descreve grandes volumes, de alta velocidade, complexidade e variabilidade de dados que requerem técnicas e tecnologias avançadas para captura, armazenamento, distribuição, gestão e análise da informação (Gandomi & Haider, 2015). Os três V's descrevem-se da seguinte forma:

- Volume refere-se à magnitude dos dados. As definições dos volumes de *Big Data* são relativas dependendo de fatores como o tempo e o tipo de dados. O que pode ser considerado *Big Data* hoje pode não o ser no futuro, uma vez que as capacidades de armazenamento de dados irão aumentar, o que permitirá recolher maiores conjuntos de dados. Dois conjuntos de dados de tamanhos equivalentes podem requerer diferentes tecnologias de gestão de dados, baseado no seu tipo - por exemplo, dados de vídeo requerem um tratamento diferente de dados textuais. Além disto, dependendo da indústria em causa, as definições de *Big Data* variam. Assim sendo, não é correto definir um limite específico para volumes de dados a partir dos quais possamos considerar estar a lidar com *Big Data* (Gandomi & Haider, 2015)*;*
- Variedade refere-se à estrutura heterogénea de um conjunto de dados. Os avanços tecnológicos permitem as empresas usarem vários tipos dados,

sejam estruturados, semiestruturados ou não estruturados. Dados estruturados equivalem a 5% do total de dados existentes (Cukier, 2010), referem-se a dados tabulares encontrados nas tabelas ou bases de dados relacionais. Texto, imagem, áudio e vídeo são exemplos de dados não estruturados que, por vezes, não têm a organização estrutural requerida pelos algoritmos mais usuais para uma análise. *Extensible Markup Language (XML),* uma linguagem textual para trocar dados na *Web,* é um exemplo de dados semiestruturados. Os documentos *XML* contêm *tags* de dados definidos pelos utilizadores que os definem como legíveis para uma máquina (Gandomi & Haider, 2015);

- Velocidade refere-se à taxa a que os dados são gerados e à velocidade a que devem ser analisados e postos em prática. A proliferação de dispositivos digitais, como os *smartphones* e sensores, originaram uma enorme quantidade de dados que necessitam de análises em tempo real e de planeamento baseado em evidências. A cadeia Wal-Mart, por exemplo, processa mais de um milhão de transações por hora (Cukier, 2010). Os dados provenientes de aplicações de dispositivos móveis produzem informações que podem ser utilizados para, em tempo real, fazer ofertas personalizadas a clientes. Estes dados fornecem informações de clientes como a localização geoespacial, dados demográficos e padrões de compra que podem ser analisados em tempo real e, ao mesmo tempo, criarem valor ao cliente (Gandomi & Haider, 2015).

O conceito descritivo dos três V's evoluiu e outras dimensões de *Big Data* foram surgindo, tais como:

- A veracidade, introduzida como o quarto V no conceito de *Big Data* pela IBM, está relacionada com a precisão, fiabilidade e qualidade dos dados. Um dos maiores problemas que o processo de decisão pode enfrentar é a falta de credibilidade dos dados, que pode conduzir a rumos de menor confiança às empresas;
- Variabilidade e complexidade foram introduzidas como dimensões adicionais de *Big Data* pela SAS. A variabilidade refere-se ao ritmo a que a informação é gerada, não sendo constante. Existem períodos de menor geração de dados e outros períodos onde se registam picos de informação. Complexidade, por sua

vez, está relacionada com o facto de grandes volumes de dados serem gerados por uma grande variedade de diferentes fontes, levando assim à necessidade de se preparar a estrutura para fazer uma correta interligação entre os dados recebidos das diferentes origens.

- A Oracle introduziu o Valor como sendo um atributo de *Big Data*. Segundo a mesma empresa, o termo *Big Data* caracteriza-se como sendo valores de baixa densidade, ou seja, os dados recebidos na sua forma original têm pouco valor comparados ao seu volume; porém, pode ser retirado um grande valor através da análise de grandes volumes de informação.

O *Big Data* modificou o modo como utilizamos e exploramos os recursos de TI, originando novas tecnologias que tiram partido dos benefícios deste novo paradigma tecnológico. O seu impacto tem-se propagado por todas as vertentes, inclusivamente nas várias indústrias dos nossos dias. Beneficiando da propagação da internet e da crescente mudança dos recursos de TI, as organizações têm procurado nos seus respetivos negócios munir os seus processos de decisão de ferramentas que lhes permitam extrair o melhor conhecimento dos dados que contêm nas bases de dados. Assim, verifica-se uma modificação no processamento e frequências de análise dos dados, isto porque se antes as organizações procediam à análise dos seus dados de forma periódica, hoje procuram cada vez mais fazê-lo de um modo metódico e em tempo real. Uma das grandes questões para as empresas atuais é como irão armazenar e tratar a grande quantidade de dados que são gerados diariamente nas suas atividades (Chen *et al.,* 2012).

2.2) Tecnologias de *Big Data*

As redes sociais causaram um impacto tremendo na nossa sociedade, resultando num aumento do número de interações entre os seus utilizadores. A maioria dessa interação assume a forma de dados não estruturados. Assim sendo, os sistemas que outrora tinham a *performance* adequada, incluindo modelos relacionais, rapidamente deixaram de ter um tempo de resposta aceitável, dando lugar à crescente adoção de tecnologia *NoSQL*. Este conceito implica o manuseamento de sistemas que distribuem bases de dados não relacionais, permitindo processamento paralelo e soluções mais facilmente escaláveis, concebidas para armazenarem grandes quantidades de dados não estruturados, como vídeo, imagens, texto, e, assim sendo, não foram concebidas para funcionar em tabelas nem requerem mecanismos SQL para manusear os dados. (Moniruzzaman & Hossain, 2013).

Dentro deste tipo de sistemas é possível coabitarem o modelo relacional e o não relacional, porém o processamento de informação pelo modelo não relacional é muito mais rápido e flexível, evitando o uso de muitos requisitos técnicos existentes nas bases de dados relacionais. Ainda assim, e apesar de todos os fatores enumerados anteriormente, esta solução de gestão de dados apresenta alguns constrangimentos relativamente à complexidade, consistência e confiança (Leavitt, 2010).

2.2.1) Hadoop

Hadoop é um projeto *open-source* da Apache Software Foundation escrito em Java que permite a existência de processamento distribuído de grandes conjuntos de dados entre *commodity clusters*. As suas duas principais componentes são o *HDFS* (*Hadoop Distributed File System*) e o *MapReduce*, onde ambas as partes estão relacionadas uma com a outra (White, 2009).

O *HDFS* é um sistema de arquivos distribuídos concebido para ser executado sobre os sistemas de arquivos locais dos nós dos clusters e armazenar arquivos extremamente grandes que permitem ter acesso a *streaming* de dados. Este sistema é altamente tolerante a falhas e é facilmente escalável desde um simples servidor para

várias máquinas, sendo que cada uma delas oferece armazenamento e poder de computação local. Consiste em dois tipos de nós: o nó denominado de "mestre" e vários nós de dados chamados "escravos". Os ficheiros de sistema registam atributos, como o tempo, modificações, permissões e espaço em disco. O conteúdo do ficheiro está separado em grandes blocos e, cada um dos blocos do ficheiro, é independentemente replicado pelos nós para existir redundância e periodicamente serem enviados relatórios de todos os blocos existentes para o nó mestre (Shvachko *et al*., 2010).

Por sua vez, o *MapReduce* é um modelo de programação simplificado para processar grandes números de *datasets* para aplicações focadas em dados. Este modelo foi desenvolvido com base no *GFS* (*Google File System*) (Ghemawat *et al*., 2003) e foi adotado através da implementação *open-source* Hadoop. O *MapReduce* permite ainda a programadores inexperientes desenvolverem programas paralelos e criarem um programa capaz de utilizar computadores na *cloud*. Na maioria dos casos, os programadores apenas são obrigados a especificar duas funcionalidades: *a map function* (*mapper*) e a *reduce function* (*reducer*) normalmente utilizada em programação funcional.

O *mapper* considera o par chave/valor como entrada e gera uma chave intermediária. O *reducer* junta todos os pares associados à mesma chave (intermediária) e gera um *output*. A *map function* é aplicada a cada par de input (chave1, valor1), onde o domínio de entrada é diferente da lista de pares de *outputs* gerada (chave2, valor2). A lista (chave2, valor2) é então agrupada pela chave. Depois de agrupada, a lista (chave2, valor2) é dividida entre várias listas [chave2, lista (valor2)] e a *reduce function* é aplicada a cada [chave2, lista (valor2)] para gerar o resultado final pretendido (chave3, valor3) (Dean, 2008). Assim sendo, o sistema de armazenamento de dados não está fisicamente separado do sistema de processamento (Hashem *et al*., 2015). Além desta *framework*, existem outros projetos *open-source* Apache relacionados com o Hadoop, como o Hive, Hbase, Mathout, Pig, Zookeeper, Spark (Ekanayake *et al*., 2010).

O Hive, por exemplo, é uma ponte de conexão similar à base de dados (BD) SQL que permite às aplicações executarem consultas relativamente aos conjuntos de dados do Hadoop. Atingiu um nível de abstração elevado relativamente ao Hadoop, onde as consultas podem ser facilmente processadas, qualquer que seja o conjunto de dados que esteja no *cluster* em questão (Zatari, 2015).

O Pig é também uma forma de conexão do Hadoop a utilizadores de negócio, sendo muito semelhante ao Hive, descrito anteriormente. A linguagem utilizada no Pig permite que as *queries* ultrapassem os *datasets* e sejam facilmente executadas, não utilizando linguagem SQL. É de acrescentar que é uma BD relacional *open-source* (Zatari, 2015).

Outro projeto relacionado com o Hadoop é a WibiData que surgiu como resultado de uma fusão de *web analysis* com Hadoop, sendo baseada em *H Base*, uma das camadas de topo da BD do Hadoop. Com esta ferramenta, os *websites* conseguem tratar de uma forma mais apropriada, analisando e trabalhando os dados informativos dos seus clientes, facilitando a comunicação e o processamento de dados no *client side*. Por exemplo, fornecer sugestões, dados personalizados para clientes e opções aos utilizadores (Zatari, 2015).

2.2.2) *Big Data* na *Cloud*

O volume dos dados está em constante aumento e, como tal, as técnicas de armazenamento têm de acompanhar, pois é necessário existir fiabilidade, eficácia e organização. O principal desenvolvimento nesta área está relacionado com o armazenamento de dados na *cloud* (Zatari, 2015).

Com toda a envolvente em torno do Big Data, este foi-se tornando popular. Existem muitas tecnologias para tratar este tipo de dados, sendo que cada empresa normalmente utiliza a tecnologia que mais a favorece. A maioria destas tecnologias são *open-source*. A arquitetura da c*loud* é *server based* e começa a ser algo cada vez mais recorrente nos dias de hoje, pois os servidores dos utilizadores não têm capacidade suficiente de armazenamento para manter os seus próprios *Data Warehouses* de *Big Data* (Zatari, 2015). Assim, já existem vários fornecedores que permitem o armazenamento na *cloud*, onde a maioria já oferece aplicações, como o Hadoop, hospedadas no *server side*, e também o processamento a ser feito no *cloud side* para que o computador do utilizador não tenha de esperar pelos longos tempos de processamento. Desta forma, tudo o que os utilizadores têm de fazer é enviar a *query online* e ver o seu resultado, sendo que os *datasets* do Hadoop são geridos de acordo com a utilização feita pelos utilizadores (Zatari, 2015).

2.3) Fontes de *Big Data*

Em todas as indústrias de qualquer parte do mundo, deve ser questionado se realmente estão a retirar todo o valor das enormes quantidades de informação que já existem dentro das suas organizações. As novas tecnologias estão a recolher mais dados do que nunca, porém muitas organizações ainda se encontram à procura de uma solução para obter valor dos seus dados e competir no mercado (LaValle *et al*., 2011).

Esta nova gama de tecnologias possibilita a análise de dados não utilizados pelas organizações. A maioria dos dados não pertencem à própria empresa, mas aos seus clientes. De acordo com estudos de mercado, 85% desses dados têm impacto no negócio e operações das empresas (Marçalo, 2014). Muitas das vezes, as oportunidades de negócio provêm do facto das empresas terem informação preciosa, mas não saberem como a usar para obter valor (Blanchard & O'Sullivan, 2015).

2.3.1) *Social Media*

As redes sociais na *Internet* tornaram-se extremamente populares e começaram gradualmente a mudar a forma como vivemos e trabalhamos. Muitas empresas analisaram o potencial de exploração comercial nestas plataformas. Embora as atividades comerciais nas redes sociais representem um aumento de produtividade para as empresas, existe quem defenda que tais atividades são um desperdício de tempo e armadilhas de segurança (Turban *et al*., 2011).

Desde o início dos anos 2000 que começaram a ser recolhidos dados da *Internet* para análise, com o intuito de desenvolver oportunidades de negócio. Os sistemas de *Web* 1.0, baseados em *HTTP*, caracterizados por mecanismos de pesquisa na *web* como o Google e a Yahoo!, e empresas de comércio eletrónico como a Amazon e o eBay, permitem que as organizações apresentem o seu negócio *online* e interajam diretamente com os seus clientes. Além das informações de produtos tradicionais e do conteúdo de negócios *online*, detalhes como o *IP*-específico do utilizador e todos os *logs* de interação recolhidos através de *cookies* e *logs* do servidor, tornaram-se numa mais valia para entender o comportamento dos clientes e, desta forma, identificar novas oportunidades de negócio. *Web intelligence, web analytics* e conteúdo gerado pelo

utilizador, recolhido através de sistemas *Web* 2.0 *based social* e *crowd-sourcing* (Doan *et al*., 2011; O'Reilly, 2005), inauguraram uma nova e cativante era centrada na análise de texto e *web analytics* para conteúdos não estruturados na *web*. Uma imensa quantidade de informações da empresa, indústria, produto e do cliente pode ser obtida através da *web* sendo organizada e visualizada através de várias técnicas de *text* e *web mining*. (Chen *et al.*, 2012). *Web site designing*, otimizações de disposição de produtos, análises das transações de clientes, análise da estrutura de mercado e recomendação de produtos, podem ser atingidas através de *web analytics* (Chen *et al.*, 2012).

As várias aplicações desenvolvidas depois de 2004, criaram uma excessiva abundância de conteúdos gerados pelos utilizadores dos vários *online social medias* como fóruns, grupos *online*, *blogs* na *web*, *social networking sites*, *social multimedia sites* (para fotografias e vídeos), mundos virtuais e jogos sociais (O'Reilly, 2005). Além de capturar conversas de celebridades, referências a eventos do dia a dia e sentimentos sociopolíticos expressos nestes *media*, as aplicações *Web 2.0* podem eficientemente reunir um grande volume de comentários e opiniões de diversos tipos de população cliente para diferentes tipos de negócio (Chen *et al*., 2012).

Muitos investigadores de Marketing acreditam que os *social media analytics* são uma oportunidade única para as empresas tratarem o mercado como "uma conversa" entre empresas e clientes, em vez do tradicional *business-to-consumer, one-way "marketing"* (Lusch *et al*., 2010). Para que tudo isto seja possível, é necessária a integração de técnicas consolidadas e escaláveis de TM, como a extração da informação, identificação dos tópicos, *opinion mining*, *question-answering*, *web mining*, *social network analysis* e análises espaciais-temporais (Chen *et al*., 2012). Com a exceção dos recursos de pesquisa, não é considerada nenhuma técnica avançada de *text analytics* para conteúdo não estruturado nas 13 capacidades de *Gartner* BI *platfforms*, porém no *Gartner* BI *Hype Cycle*, existem outras técnicas de *text analytics*, como serviços semânticos de informação, *natural language question answering* e *content*/*text analytics* (Bitterer, 2011).

2.3.2) *Internet of Things (IoT)*

Como mencionado no artigo *The Economist* (2011), o número de telemóveis e *tablets* (cerca de 480 milhões de unidades) superou o número de computadores (cerca de 380 milhões de unidades) pela primeira vez em 2011. O mesmo artigo referenciado anteriormente, projetou que o número de dispositivos móveis conectados chegaria a 10 biliões em 2020 (*The Economist*, 2011).

Dispositivos móveis como o *iPad*, *iPhone* e outros *smartphones*, e todo o seu universo de *downloads* aplicacionais, estão a transformar as diferentes facetas da sociedade, desde a educação aos serviços de saúde ou desde o entretenimento aos governos. Dispositivos baseados em sensores via *Internet* equipados com *RFID* (*Radio-Frequency IDentification*), códigos de barras, e etiquetas de rádio, são notícias interessantes para aplicações inovadoras (Chen *et al.*, 2012). As capacidades desses dispositivos móveis que normalmente se encontram conectados à *Internet*, o facto de suportarem um forte sinal móvel, uma localização precisa e operações e transações relevantes para o contexto, continuará a oferecer resultados, desafios e oportunidades únicas ao longo da década 2010-2020 (Chen *et al.*, 2012).

Embora a chegada da era *Web 3.0* (*mobile* e *sensor based*) pareça estar cada vez mais próxima, existem ainda técnicas *mobile analytics* de recolha, processamento e visualização de dados que, face à dinâmica estrutural dos *datasets*, ainda são desconhecidas. A década de 2010-2020 promete resultados interessantes a nível de pesquisa e desenvolvimento do tema (Chen *et al.*, 2012).

2.4) *Big Data Analytics*

Ter apenas volumes de dados grandes não representa nenhum ativo para as organizações. Todo o seu potencial é desbloqueado apenas quando alavancado para conduzir à tomada de decisão, baseada em evidências factuais. Para tal, as organizações necessitam de processos eficientes para transformar grandes volumes de, rápidos e diversos, dados em informação com valor significativo. O processo geral de extração de conhecimento de *Big Data* pode ser repartido em cinco fases (Labrinidis & Jagadish, 2012). Estas cinco fases originam dois subprocessos: *Data Management* e *Analytics.*

Data Management envolve processos e tecnologias de apoio, para adquirir, armazenar dados e preparar a sua devolução para o subprocesso *Analytics* que, por sua vez, baseia-se em técnicas utilizadas para analisar e adquirir inteligência do *Big Data*. Desta forma, *Big Data Analytics* pode ser visto como um subprocesso no processo global de *insight extraction* do *Big Data* (Gandomi & Haider, 2015). Os *datasets* de *Big Data* podem ser de diferentes tipos como de texto, áudio e vídeo, como é descrito de seguida.

2.4.1) *Text Analytics*

Uma parte significativa do conteúdo não estruturado recolhido por uma organização é em formato textual, desde comunicações via *e-mail* e documentos corporativos, a *web pages* e conteúdo de redes sociais. A análise textual tem as suas raízes na recuperação de informação e linguística computacional (Chen *et al.*, 2012).

Na recuperação de informação, a representação dos documentos e o processamento de *queries* são os fundamentos para o desenvolvimento do modelo de espaço vetorial, modelo de recuperação booleano e modelo de recuperação probabilística que, por sua vez, tornou-se a base para as modernas bibliotecas digitais, ferramentas de pesquisa e sistemas de pesquisa corporativos (Salton, 1997).

Na linguística computacional, as técnicas de *statistical natural language processing* (*NLP*) para a aquisição lexical, desambiguação do sentido da palavra, *part-of-the-speech-tagging* (*POST*) e gramáticas probabilísticas livres contextualmente, foram importantes para a representação textual. Em adição ao referido anteriormente, os

modelos de utilizadores e *feedback* relevante são também importantes instrumentos para a melhoria da *performance* (Chen *et al.,* 2012).

2.4.2) *Analytics* Aplicada a Outros Tipos de Dados

Além de *Text Analytics*, existe *Analytics* aplicada outros tipos de dados, como dados de áudio e de vídeo.

Dessa forma, *Audio Analytics* trata-se da análise e extração de informação de dados de áudio não-estruturados. Quando aplicado a linguagem oral humana, a análise do áudio é também tratada como a análise da fala. Nos dias de hoje, os centros de atendimento ao cliente e os centros de saúde são as áreas onde se tem aplicado esta técnica. Os centros de atendimento a clientes utilizam *Audio Analytics* para os milhares ou até mesmo milhões de horas de chamadas gravadas. Esta técnica ajuda a melhorar a experiência do cliente, avalia a *performance* do operador, ajuda a identificar potenciais problemas nos produtos ou serviços e ficamos a conhecer melhor o comportamento do nosso cliente (Gandomi & Haider, 2015). Os sistemas de *Audio Analytics* podem ser desenhados de forma a analisar uma chamada em tempo real, formular recomendações de *cross/up-selling* baseadas nas interações, passadas e presentes, do cliente e providenciarem *feedback* aos operadores em direto. Alem disto, os *call centres* automatizados usam plataformas de *Interactive Voice Response (IVR)* para identificarem e tratarem situações em que os clientes se sintam frustrados (Gandomi & Haider, 2015).

Relativamente ao *Video Analytics* ou *vídeo content analysis* (*VCA*), envolve uma grande diversidade de técnicas de monitorização, análise e extração de informação minuciosa de fluxos de vídeo (Gandomi & Haider, 2015). Várias foram as técnicas que já foram desenvolvidas para processar vídeos em tempo real e vídeos pré-gravados. A crescente prevalência de câmaras *closed-circuit television* (*CCTV*) alineado à crescente popularidade de *websites* de *video sharing*, são os principais contribuintes para o crescimento das análises de vídeo computacionais. Porém, um desafio fundamental é o tamanho absoluto dos dados de vídeo. Para contextualizar em termos de tamanho, o equivalente a um segundo de vídeo em alta definição são 2000 páginas de texto (Manyika *et al*., 2011).

As tecnologias de *Big Data* transformam este desafio em oportunidade, pois podem ser utilizadas para extrair o conhecimento de milhares de horas de vídeo. Como resultado, as tecnologias de *Big Data* são o terceiro fator que mais contribuiu para o desenvolvimento da análise de vídeo (Gandomi & Haider, 2015).

A principal aplicação da análise de vídeo nos anos mais recentes tem sido nos sistemas automatizados de segurança e vigilância. As análises de vídeo podem efetuar eficientemente funções de segurança como a deteção de infrações em zonas restritas, identificar objetos indesejados que tenham sido removidos ou deixados numa zona específica, detetar atividades suspeitas e adulterações das câmaras. Os dados gerados pelas câmaras *CCTV* em pontos de venda podem ser extraídos para sistemas de BI, com a finalidade de reportar ao *Marketing* e às Operações. Os algoritmos podem inclusivamente recolher informação demográfica sobre os clientes, como a idade, sexo e etnia. Da mesma forma, é possível contar o número de clientes, medir o tempo que cada cliente fica numa determinada loja, detetar padrões de movimento, medir o seu tempo de permanência numa certa área e monitorizar filas de espera em tempo real. De toda esta informação, podem ser retiradas informações valiosas para o negócio, correlacionando as variáveis anteriormente citadas com a demografia dos consumidores, com o intuito de conduzir a decisões para a colocação de um produto, o preço e promoções (Gandomi & Haider, 2015).

Outra aplicação potencial para a análise de vídeo trata-se do estudo do comportamento dos grupos, ou seja, quando por exemplo uma família vai às compras, apenas um membro interage com a loja, sendo que a informação recolhida pelos sistemas tradicionais apenas remete aquele consumidor, faltando assim dados sobre os padrões de compra dos restantes familiares. Desta forma, a análise de vídeo pode auxiliar, fornecendo informações sobre o tamanho, a demografia e o comportamento individual de compra de cada um dos restantes compositores do grupo (Gandomi & Haider, 2015).

2.4.3) *Social Media Analytics*

As análises de *Social Media* referem-se à análise de dados estruturados e não estruturados de canais de *Social Media*, conceito que abrange uma variedade de plataformas *online* que permite aos seus utilizadores criar e trocar conteúdo. Os meios de comunicação social podem ser categorizados dentro dos seguintes tipos: Redes sociais, como por exemplo, o *Facebook* e o *LinkedIn*, blogues tais como o *Blogger* e o *WordPress*, *microblogs* como o *Twitter* e o *Tumblr*, notícias sociais (*Digg* e *Reddit*), *social bookmarking* (*Delicious* e *StumbleUpon*), partilha de *media* (*Instagram* e *YouTube*), *wikis* (*Wikipedia* e *Wikihow*), *websites* de perguntas e respostas (*Yahoo! Anshwers* e *Ask.com*) e sites de *reviews* (*Yelp* e *TripAdvisor*) (Barbier & Liu, 2011).

A análise dos meios de comunicação social nasceu após o surgimento da *Web* 2.0, no início de 2000. A característica chave desta análise é centrar-se nos dados. Esta pesquisa abrange vários campos, como a psicologia, sociologia, antropologia, informática, matemática, física e economia. Nos anos mais recentes, o *Marketing* tem sido a aplicação primária dos meios de comunicação social. Isto pode ser atribuído à vasta e crescente adoção dos *social media* pelos consumidores de todo o mundo (He *et al*., 2013).

Os conteúdos gerados pelos utilizadores como, por exemplo, sentimentos, imagens e vídeos, as relações e interações entre entidade como pessoas, organizações e produtos, são as duas fontes de informação dos meios de comunicação social. Baseando-se nesta categorização, os meios de comunicação social podem ser classificados em dois grupos distintos, *Content-based analytics* e *Structure-based analytics.*

O *Content-based analytics* foca-se nos *posts* dos utilizadores nas redes sociais, como *feedback* de clientes, *reviews* de produtos, imagens e vídeos. Todo este conteúdo presente nos *social media* são volumosos, não-estruturados e dinâmicos. Como referido anteriormente, o texto, áudio, e análises analíticas podem ser aplicados para obter conhecimento sobre estes tipos de dados. Além disso, as tecnologias de *Big Data* podem ser aplicadas para resolver os desafios de processamento de dados (Gandomi & Haider, 2015).

Relativamente ao *Structure-based analytics*, este concentra-se em sintetizar os atributos estruturais das redes sociais e extrair inteligência das relações entre as entidades participadoras. A estrutura de uma rede social é modelada através de um conjunto de nós e arestas, que representam as entidades participantes. São revistos dois tipos de gráficos de redes, conhecidos como gráficos sociais e gráficos de atividade (Heidemann *et al*., 2012). Nos gráficos sociais, uma aresta existente entre um par de nós significa a existência de um *link* entre as respetivas entidades (uma amizade, por exemplo). Nas redes de atividade, uma aresta significa uma interação entre qualquer par de nós. As interações envolvem trocas de informação, como por exemplo, comentários e *likes*. Os gráficos de atividade são preferíveis aos gráficos sociais, pois uma relação ativa é mais interessante de analisar do que uma mera conexão (Gandomi & Haider, 2015).

Foram várias as técnicas que emergiram recentemente para extrair informação de redes sociais estruturadas, entre as quais se destacam a *Community detection*, *Social influence* e *Link prediction*. A *Community detection*, também conhecida como descoberta comunitária, extrai comunidades implícitas dentro de uma rede. Para redes sociais *online*, uma comunidade refere-se a uma sub-rede de utilizadores que interagem mais frequentemente entre eles do que com os restantes utilizadores da rede. Esta técnica ajuda a sumarizar grandes redes, facilitando a descoberta de padrões comportamentais existentes e prevendo propriedades emergentes da rede. É similar a *clustering* (Aggarwal, 2011), uma técnica de *data mining* usada para dividir um conjunto de dados em subconjuntos disjuntos com base na semelhança de pontos dos dados. A deteção comunitária encontrou várias áreas de aplicação, incluindo o marketing e a *World Wide Web* (Parthasarathy *et al*., 2011), permitindo assim às empresas desenvolverem sistemas de recomendação de produtos mais eficientes.

Por sua vez, a *Social influence* refere-se a técnicas que dão preponderância à modelação e avaliação da influência de utilizadores e conexões numa rede social. O comportamento de um utilizador numa rede social é influenciado pelos outros. Assim sendo, é necessário avaliar a influência dos participantes, quantificar a força das conexões e descobrir padrões de difusão de influência numa rede. Um aspeto a salientar da análise de influência é quantificar a importância dos nós de uma rede. Para esse propósito, desenvolveram-se várias medidas, como por exemplo: *degree centrality*, *betweenness centrality*, *closeness centrality* e *eigenvector centrality*. Outras medidas avaliam a força das conexões representadas pelas arestas ou modelam a influência nas redes sociais, como o *Linear Threshold Model* (*LTM*) e o *Independent Cascade Model* (*ICM*) (Sun & Tang, 2011).

Relativamente à *Link prediction*, esta aborda especificamente o problema de prever futuras ligações entre nós existentes na rede subjacente. Normalmente a estrutura das redes sociais não é estática e cresce continuamente através da criação de novos nós e arestas. Assim sendo, um objetivo natural é perceber e prever a dinâmica da rede. As técnicas de *link prediction* prevêem a ocorrência de interação, colaboração ou influência entre entidades de uma rede num específico intervalo de tempo (Liben-Nowell & Kleinberg, 2007). No contexto dos *social media*, a aplicação principal da *link prediction* é a recomendação e desenvolvimento de sistemas de recomendação, como no *Facebook* – "Pessoas que possas conhecer", *Youtube* – "Recomendados para ti", entre outras plataformas que utilizam sistemas de recomendação semelhantes.

2.4.4) *Predictive Analytics*

Predictive Analytics compreende uma variedade de técnicas que pressupõem resultados futuros baseados em dados históricos e do presente. Esta análise pode ser aplicada em quase todas as áreas – desde a previsão de falha de motores de um jato com base num fluxo de dados de milhares de sensores, à previsão dos próximos passos dos consumidores baseando-se nas suas compras, nas datas das compras e até mesmo no que por eles é dito nas redes sociais. O método procura descobrir padrões e capturar relações entre os dados, estando dividida em dois grupos.

Algumas técnicas, como a *moving average*, tentam descobrir padrões históricos nos resultados e extrapolá-los para o futuro. Outras, como a regressão linear, procuram capturar interdependências entre os resultados e suas variáveis explicativas para, de seguida, as explorarem para previsões. Com base na metodologia subjacente, as técnicas também podem ser categorizadas em dois grupos: técnicas de regressão e técnicas de ML. Outra classificação é baseada no tipo de resultados: técnicas como a regressão linear apontam para variáveis de resultado contínuo (exemplo: preço de venda das casas) enquanto outras como as *random forests* são aplicadas a variáveis de resultado discretas (exemplo: crédito bancário). As técnicas de análise preditiva baseiam-se principalmente em métodos estatísticos (Gandomi & Haider, 2015).

Vários fatores determinam o desenvolvimento de métodos estatísticos para *Big Data*. Em primeiro lugar, os métodos estatísticos convencionais estão enraizados em significâncias estatísticas: é recolhida uma pequena amostra da população e o resultado é comparado com a hipótese de analisar a significância de uma relação particular. A conclusão é então generalizada para toda a população. Por outro lado, as amostras de *Big Data* são massivas e representam a maioria, senão toda a população. Como resultado, a noção de relevância estatística não é tão relevante para *Big Data*. Em termos de eficiência computacional, muitos métodos convencionais para pequenas amostras não escalam até *Big Data* (Fan *et al*., 2014).

2.5) *Text Mining*

Segundo a *Gartner*, TM incide no processo de extrair informações de coleções de dados textuais e utilizá-las para objetivos de negócios (Gartner IT Glossary, n.d.), enquanto *Text Analytics* refere-se ao processo de retirar informações de fontes de texto. É utilizado para vários propósitos, tais como: resumir (tentar encontrar o conteúdo-chave em um corpo textual ou num único documento), análise de sentimentos (qual é a natureza do comentário sobre um assunto), investigações (quais são os casos particulares de uma questão específica) e classificação (que assunto ou que partes de conteúdo-chave o texto fala) (Gartner IT Glossary, n.d.). Assim, TM refere-se a técnicas que extraem informação sobre dados textuais (Gandomi & Haider, 2015). *Feeds* de notícias de uma rede social, *e-mails*, blogues, fóruns *online*, resposta a questionários e notícias são alguns exemplos de dados textuais. A análise de texto envolve análise estatística, linguística computacional e ML (Gandomi & Haider, 2015).

As ferramentas de TM permitem às organizações converter grandes volumes de texto gerado por humanos num simples resumo, utilizando para isso ferramentas qualitativas de análise textual onde são contabilizadas o número de ocorrências de determinado texto com uma determinada relevância que, após aplicação de métodos quantitativos para se extrair conhecimento, servem como instrumento numa tomada de decisões baseada em evidências. Assim, um dos exemplos da sua utilização é na previsão do mercado de ações, baseando-se em informações extraídas de notícias do foro financeiro (Oliveira *et al*., 2016).

Uma das técnicas é denominada de *Information Extraction* (IE) e extrai dados estruturados de texto não estruturado (Chung, 2014). Por exemplo, os algoritmos de *IE* conseguem extrair o nome do medicamento, a sua dosagem e a periodicidade da toma através da prescrição médica. Duas subaéreas nas *IE* são o reconhecimento de entidades (*ER*) e a relação extraída (*RE*) (Gandomi & Haider, 2015).

Através de *RE*, é possível identificar nomes em texto e classificá-los de acordo com as categorias predefinas anteriormente como pessoa, data, localização e organização. Além disso, encontra e extrai relações de semântica entre as entidades no texto, como por exemplo pessoas, organizações, remédios e genes. Na seguinte frase: "O Steve Jobs foi cofundador da Apple Inc. em 1976", um sistema de *RE* consegue

extrair as relações como fundador de [Steve Jobs, Apple Inc.] ou fundada em [Apple Inc., 1976] (Gandomi & Haider, 2015).

Outra técnica considerada é a de sumarização. Através das *summarization techniques* produz-se automaticamente uma sumarização sucinta de um ou vários documentos. O resultado final transmite a informação chave do texto original. Aplicações para esta técnica podem variar muito de contexto, desde artigos científicos, anúncios, *e-mails*, *blogs*. Por norma, a sumarização segue duas abordagens: a extrativa e abstrata. Na sumarização extrativa, um resumo é criado a partir do texto original (normalmente a partir das frases). O resultado é assim um subconjunto do documento original, onde a criação de um resumo, segundo esta abordagem, envolve destacar os principais fatores de um texto e agrupá-los. A importância de unidades textuais passa pela análise da sua localização e frequência textual. Estas técnicas não requerem uma compreensão textual.

Por outro lado, a sumarização abstrata envolve técnicas de extração de informação semântica do texto. O resumo pode conter unidades textuais que não necessitam de estar presentes no texto original. Para converter o texto de origem e gerar o resumo, esta abordagem incorpora técnicas avançadas de *NLP* (*Natural Language Processing*). Como resultado, este tipo de sistema tende a gerar resumos mais coerentes do que os sistemas extrativos, porém os sistemas extrativos são mais fáceis de implementar (Hahn & Mani, 2000), especialmente quando estamos a falar em *Big Data*.

Por sua vez, as técnicas de *Question Answering* (*QA*) dão-nos respostas às perguntas feitas em *natural language*. A Siri e a Watson, da Apple e IBM respetivamente, são exemplos de sistemas de *QA*. Estes sistemas já foram implementados na saúde, finanças, marketing e educação. Assim como a sumarização abstrata, os sistemas de *QA* utilizam técnicas *NLP*. As técnicas de *QA* são classificadas em três categorias: *information retrieval* (*IR*) *approach*, *knowledge-based approach* e a *hybrid approach* (Gandomi & Haider 2015). Os sistemas baseados em *IR QA* normalmente dividem-se em três subcomponentes (Gandomi & Haider, 2015):

1. *Question processing* – utilizado para determinar detalhes, como o tipo de questão, o foco da questão e o tipo da resposta, que são utilizados para fazer uma *query*;

2. *Document processing* – usado para devolver informação relevante pré-escrita de um conjunto de documentos existentes, utilizando a *query* formulada no processamento da questão;
3. *Answer processing* – serve para extrair respostas candidatas do resultado final da componente anterior, para as classificar e retornar o candidato com o melhor *rank* como *output* do sistema de *QA*.

Os sistemas *QA* baseados em *Knowledge* geram uma semântica descritiva para a questão, que é então utilizada para consultar fontes estruturadas. Estes sistemas são particularmente úteis para domínios restritos, como o turismo, medicina e transportes, onde não existem grandes volumes de documentos pré-escritos. Esses domínios carecem de redundância, o que é necessário para sistemas *QA* baseados em *IR*. A Siri, da Apple, é um exemplo de sistema *QA* que explora a abordagem baseada em conhecimento (Gandomi & Haider, 2015).

Nos *Hybrid QA systems*, como o Watson da IBM, enquanto a questão é semanticamente analisada, a resposta candidata é gerada utilizando métodos *IR* (Gandomi & Haider, 2015).

Outras técnicas a considerar são as de *Sentiment analysis* (*opinion mining*), que analisam texto de opinião, onde existe pareceres de pessoas sobre entidades, produtos, organizações, indivíduos e eventos. Os negócios começam a, cada vez mais, capturar informação sobre os sentimentos dos seus clientes, o que levou à proliferação da análise de sentimentos (Liu, 2012). As áreas onde mais se aplica a análise de sentimentos são no marketing, finanças e ciências políticas e sociais. Estas técnicas são ainda divididas em três subgrupos, nomeadamente *document-level*, *sentence-level* e *aspect-based* (Gandomi & Haider, 2015).

As técnicas de *document-level* determinam se todo o documento expressa um sentimento negativo ou positivo. Assume-se que o documento contém sentimentos sobre uma entidade concreta. Enquanto certas técnicas categorizam um documento em duas classes, positiva e negativa, outras incorporam mais classes de sentimentos (como o sistema *five-star* da Amazon) (Feldman, 2013).

Por sua vez, as *Sentence-level techniques* tentam determinar a polaridade de um único sentimento de uma entidade conhecida numa única frase. Estas técnicas necessitam de primeiro distinguir frases subjetivas de objetivas. Consequentemente, as

técnicas de *sentence-level* tendem a ser mais complexas comparativamente às técnicas citadas anteriormente (*document-level*) (Gandomi & Haider, 2015).

Para finalizar, as técnicas de *aspect-based* reconhecem todos os sentimentos dentro de um documento e identificam os aspetos da entidade à qual cada sentimento diz respeito. Normalmente *reviews* de produtos contêm opiniões sobre diferentes aspetos (ou características) de um produto. Utilizando este tipo de técnicas, o vendedor pode obter informação valiosa sobre diferentes atributos do produto que provavelmente seriam esquecidas se o sentimento fosse apenas classificado em termos de polaridade (Gandomi & Haider, 2015).

2.6) *Mining News*

Big Data tornou-se uma questão importante para um grande número de áreas de pesquisa, como *data mining*, ML, inteligência computacional, *information fusion*, a *web* semântica e redes sociais. O surgimento de diferentes estruturas de dados como o Apache Hadoop e, mais recentemente, o Spark para o processamento massivo de dados, baseia-se no paradigma *MapReduce*, permitindo a utilização eficiente de métodos de *data mining* e algoritmos de ML em diferentes domínios (Bello-Orgaz *et al*., 2016).

A combinação de TSBD e algoritmos tradicionais de ML gerou novos e interessantes desafios em outras áreas, como nas redes sociais. Estes novos desafios são focados principalmente em problemas como processamento de dados, armazenamento de dados, representação de dados e como os dados podem ser usados para *pattern mining*, analisando comportamentos de utilizadores e na visualização e procura de dados (Bello-Orgaz *et al*., 2016).

Assim, e de forma a encontrar o *research gap*, foram analisados cerca de quarenta artigos científicos, resultantes da filtragem a partir de 2013, no agregador *Google Scholar*, com diferentes combinações entre os seguintes termos de pesquisa: "*Text Mining*" "*Google News*" "*Big Data*". De entre todos, decidiu-se optar pelos artigos mais próximos deste trabalho, escolhendo os onze mais representativos. Para tal, foi criada a Tabela 1 (ver próximas páginas), com os critérios "Autores", "Fonte", "Objetivo" e "Método".

A análise da Fonte, Objetivo e Método deste trabalho encontra-se na última linha da Tabela 1 para efeitos de comparação direta. Para um melhor destaque, os conceitos Fonte, Objetivo e Método(s) similar(es) aos abordados neste estudo, foram delineados a **negrito**.

A título conclusivo, podemos verificar através da Tabela 1 a existência de apenas dois trabalhos com a mesma Fonte do que a utilizada neste estudo ("Google News"). Existem outros trabalhos com parte do seu Objetivo semelhante, e vários estudos com Método(s) equivalente(s) no tratamento dos dados. Porém não é verificado, com base nos critérios enumerados anteriormente, um trabalho científico com a combinação de uma Fonte, Objetivo e Método(s) semelhante(s) às que este trabalho se propõe.

__Tabela 1__ - Comparação entre artigos e este trabalho face à Fonte, Objetivo e Método.

Artigo	Autores	Fonte	Objetivo	Método
Topic bias in the social media - The representation of political topics in Obama's 'Open for Questions'-campaign in comparison to traditional media and the blogosphere.	(Riedel & Send, 2009).	Keywords de pesquisa no **Google News** baseadas nas perguntas do website "Open for Questions".	Mostrar se existe uma correlação entre os relatórios nos tópicos políticos em diferentes fontes de comunicação e um tópico específico sem entrar numa profundidade qualitativa do conteúdo de cada fonte.	-
Text mining of news-headlines for FOREX market prediction - A Multi-layer Dimension Reduction Algorithm with semantics and sentiment.	(Nassirtoussi *et al*., 2015).	MarketWatch.com ou outros através de Really Simple Syndication (RSS).	Exploração de um novo caso de estudo específico: "Short term FOREX prediction based on **news**-headlines".	**ML** - Support Vector Machine (SVM) e **Natural Language Processing (NLP)**.
Building emotional dictionary for sentiment analysis of online news.	(Rao *et al*., 2014).	Sina e SemEval.	Propor um algoritmo eficiente e estratégias para construir automaticamente um **dicionário de palavras** emotivas para a deteção de palavras de emoção social.	-
Stock market sentiment lexicon acquisition using microblogging data and statistical measures.	(Oliveira *et al*., 2016).	StockTwits e Twitter REST API.	Apresentar uma nova e rápida maneira de criação de léxicos do mercado de ações.	Stanford Core**NLP**.

Predicting consumer sentiments from online text.	(Bai, 2011).	**Google News**, Reuters, This is Money, The Motley Fool e Infonic Ltd.	Propor um modelo de pesquisa heurística baseado no Modelo de Markov que é capaz de capturar as dependências entre palavras e fornecer um vocabulário adequado para a finalidade de extrair sentimentos.	**ML.**
Large scale opinion mining for social, news and blog data.	(Godbole *et al*., 2007).	Palo Pro.	Propor uma plataforma de "opinion mining" para resposta em tempo real (**categorização** e classificação **do conteúdo recolhido entre o domínio de notícias**).	**ML e Natural Language Processing (NLP).**
The impact of social and conventional media on firm equity value: A sentiment analysis approach.	(Yu *et al*., 2013).	COMPUSTAT e o Center of Research in Security Prices (CRSP).	Investigar o efeito dos *social* e *conventional medias*, a sua importância e a sua inter-relação no desempenho de curto prazo do mercado de ações.	**ML e Natural Language Processing (NLP).**
Mining Online Text Data for Sentiment and News Impact Analysis.	(Wei, 2013).	Datasets públicos como o "Movie Review Data". Para algumas tarefas e análises, extração e categorização manual.	Desenvolver abordagens para **analisar notícias** de negócios **online** e reviews de clientes online, **extraindo conhecimento desse texto**.	**ML.**
A social-media-based approach to predicting stock comovement.	(Liu *et al*., 2015).	Bolsa de Valores: NYSE e NASDAQ.	Um modelo inovador para identificar grupos de ações homogéneos e prever o movimento de ações em relação às métricas de social media de uma empresa específica.	-
Detecting tension in online communities with computational Twitter analysis.	(Burnap *et al*., 2015).	Website COSMOS.	Investiga a possibilidade de previsão de picos no serviço de tensão social da polícia do Reino Unido através dos social media.	Implementação de uma Tension Analysis Engine e de uma abordagem de **ML**.

Social big data - Recent achievements and new challenges	(Bello-Orgaz *et al*., 2016).	-	Revisão das novas metodologias desenvolvidas para o tratamento de dados e de informações dos social media das novas aplicações e ferramentas que atualmente aparecem sobre a alçada das redes sociais, social media e paradigmas **Big Data**.	Revisão Literária.
Categorização e Classificação de Notícias de Big Data em Tecnologias segundo o Quadrante Mágico de *Gartner*	A abordagem proposta.	**Google News**	Providenciar uma forma alternativa de categorização de notícias através do tipo de TBD e do QG, sumarizando em tópicos de interesse um volume elevado de notícias sobre o tema "*Big Data*", a partir de um período de tempo específico através de técnicas de **ML** e extrair conhecimento oculto do texto formal das notícias.	**ML** e **Natural Language Processing (NLP).**

3. Metodologia

3.1) Recolha de Dados

Para a recolha de notícias, foi escolhido um agregador de notícias, de forma a ser possível, partindo de um único sítio, ter todas as notícias sobre um determinado tópico. Assim, escolheu-se como agregador de notícias o *Google News*.

Foi definido que a recolha de notícias seria entre o ano de 2013 e 2016, na língua Inglesa, deveria conter a expressão "*big data*" no título da notícia (*case-insensitive,* ou seja, a expressão "*Big Data*" ou "*big Data*", por exemplo, aplicam-se) e as notícias mais relevantes apareceriam em primeiro lugar. Posto isto, podemos observar na Figura 3 um exemplo de uma pesquisa no agregador de notícias *Google News*, com os critérios de busca destacados a cinzento. Esta pesquisa foi filtrada para que apareça no título de todas as notícias a expressão "Big Data" (*case-insensitive*), apenas notícias entre o período definido, neste caso entre 01/01/2016 e 31/01/2016, e apresentadas pela sua relevância.

Inicialmente, foi idealizado que a recolha de notícias seria feita anualmente, porém o *Google News* apresenta sempre o mesmo número máximo de páginas de notícias para diferentes universos que atinjam o máximo de páginas, ou seja: imaginando que em 1 página de notícias existem sempre 10 notícias e existe um total de 200 notícias é equivalente a 20 páginas de notícias apresentadas pelo *Google News*. Porém, diferentes resultados serão mostrados caso se altere a data. Se a nossa data for compreendida entre 01/01/2016 e 31/12/2016, o ano completo de 2016, com os critérios definidos, temos 500 resultados por exemplo. Da mesma forma, o resultado será semelhante quanto ao número de páginas se definirmos a data situada entre 01/01/2016 e 31/01/2016, o mês de janeiro completo do ano de 2016, com os mesmos critérios, onde emergem 300 resultados, por exemplo. Em espaços temporais bastantes diferentes temos um conjunto de resultados muito semelhante, relativamente ao número de páginas. Desta forma, percebe-se a importância do critério relevância e existe um entendimento real de como os resultados se adaptam à nossa pesquisa. Assim, tratando-se de *web scraping* de um dos *websites* mais complicados de o fazer, foi definida a estratégia de extrair mensalmente as notícias, de forma a obter um número maior de notícias.

Outra situação identificada foi a língua das notícias que, apesar de existir uma filtragem no *Google News* para apenas apresentar notícias redigidas na língua Inglesa, nas páginas finais de resultados existiam notícias de línguas diferentes. Na secção Extração Script Python, este tema será novamente abordado e será explicado com maior detalhe.

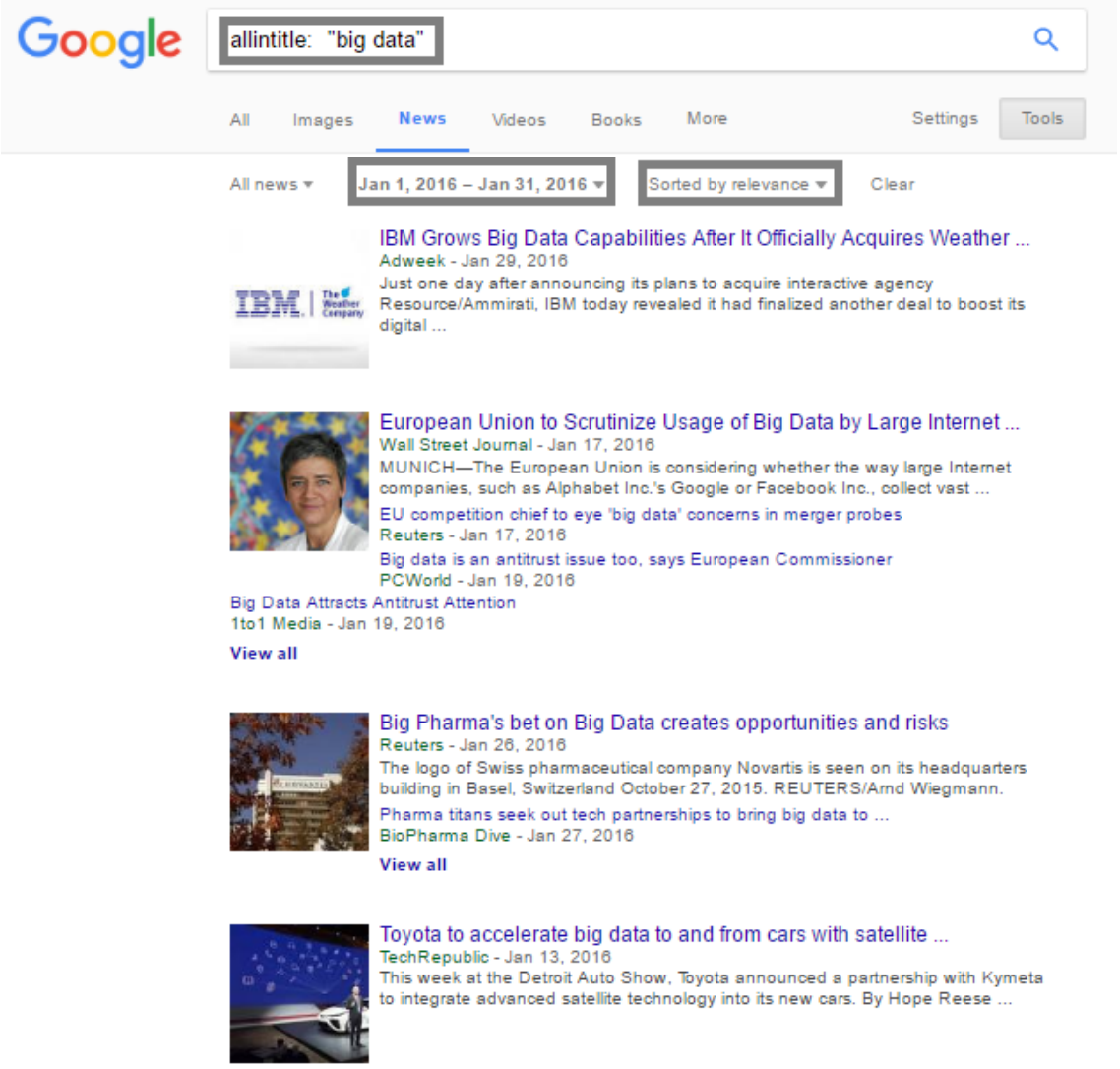

Figura 3 *- Pesquisa no Google News com os critérios de filtragem adequados.*

O processo de recolha de dados foi bastante complexo em termos de implementação, muito derivado ao agregador de notícias em questão, porém foi possível extrair os dados necessários. Por sua vez, esta extração divide-se em duas fases distintas: Extração Parcial da Notícia e Extração *Script Python*.

3.1.1) Extração Parcial da Notícia

Na extração parcial da notícia, foi utilizado um *software* denominado de *Octoparse*, uma ferramenta *client-side* escrita em *.NET*, que serve para extração de dados de *websites*. Foi montado um ciclo nesta ferramenta com o intuito de se conseguir extrair parte da informação pretendida da notícia, como se constata na Figura 4.

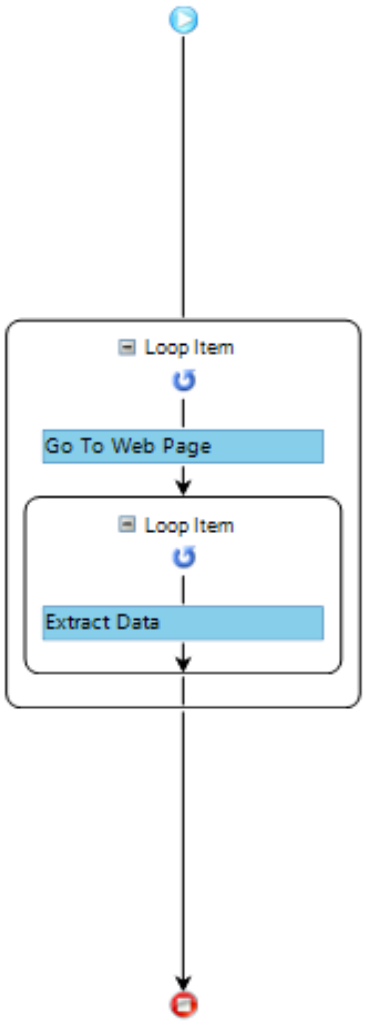

Figura 4 *- Ciclo criado no Octoparse para extrair a informação pretendida da notícia.*

É introduzida uma lista de *URLs*, resultante da pesquisa realizada, no primeiro *Loop Item* do ciclo desenvolvido no *software Octoparse*. Esta lista é composta pelo *URL* da respetiva página de resultados e é extraída manualmente, página a página, para um ficheiro do tipo *.txt*. Se existirem 11 páginas de resultados, haverá uma lista de

URLs composta por 11 endereços. Um exemplo de uma lista de *URLs* encontra-se na Tabela 2.

***Tabela 2** - Lista de URLs de todas as páginas de resultados da pesquisa realizada.*

Número da Página	URL da Página
1	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**10**&sa=N&dpr=1
2	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**20**&sa=N&dpr=1
3	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**30**&sa=N&dpr=1
4	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**40**&sa=N&dpr=1
5	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**50**&sa=N&dpr=1
6	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**60**&sa=N&dpr=1
7	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**70**&sa=N&dpr=1
8	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**80**&sa=N&dpr=1
9	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**90**&sa=N&dpr=1
10	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**100**&sa=N&dpr=1
11	https://www.google.pt/search?q=allintitle:++"big+data"&hl=en&as_drrb=b&authuser=0&tbas=0&biw=1600&bih=763&noj=1&tbs=cdr:1,cd_min:1/1/2016,cd_max:1/31/2016&tbm=nws&ei=Q62fWKXkGIaVaLXmo5AK&start=**110**&sa=N&dpr=1

De seguida itera-se a lista anterior e cada iteração corresponde a um endereço *URL* no *Go To Web Page* do ciclo da Figura 4. Estar numa página específica, é equiparável ao segundo *Loop Item* do mesmo ciclo, onde cada uma das iterações equivale a uma das notícias destacadas da respetiva página, como é visível na Figura 5.

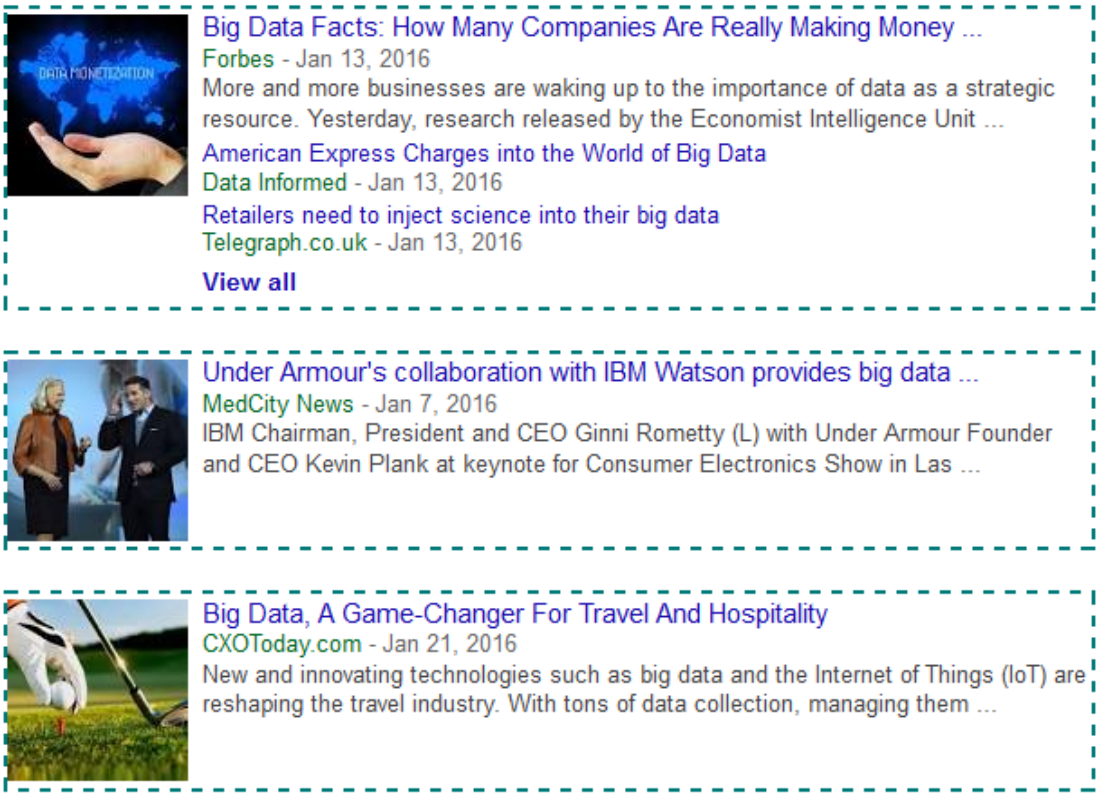

Figura 5 *- Notícias de uma página específica da pesquisa aplicada.*

Assim, chega-se à última etapa do ciclo, o *Extract Data*. Nesta fase, está-se num valor específico da iteração do segundo *Loop Item*, correspondendo a uma notícia em particular. Dessa notícia irá ser extraída a respetiva informação para os campos URL_Notícia, URL_Página, Data_Extração, Título_Notícia, Fonte_Notícia, Data_Notícia e Sumário_Notícia, que integram um ficheiro do tipo *.csv*. Na Figura 6 pode-se observar um exemplo dos campos recolhidos a tracejado.

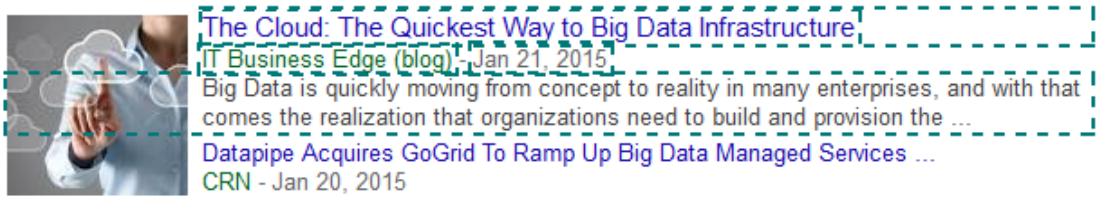

Figura 6 *- Campos a serem extraídos da Notícia.*

Um grande desafio desta extração foi o facto de ter de ser feito *web scraping* de um dos *websites* mais desafiantes para extrair informação, devido a todos os mecanismos *anti-bot* que têm para bloquear extração de informação de uma forma não manual, como por exemplo a utilização de *CAPTCHAS* (um teste cognitivo, utilizado como ferramenta *anti-spam*) e bloqueamento do *Ip Address*.

Após todas as extrações de todos os meses do intervalo de anos em questão terem sido realizadas, foi gerado um ficheiro *.csv* para cada um dos meses de um determinado ano. Foi então criado um novo ficheiro *.csv* denominado de *merged_lines.csv*, onde todas as linhas de resultados de todos os meses de todos os anos são juntas. Estando a ser abordado notícias deste tópico, achou-se bem trabalhar com uma amostra de resultados *Big Data*, resultando o ficheiro *merged_lines.csv* com cerca de 11505 linhas após todas as filtragens e limpezas que serão abordadas no próximo segmento.

3.1.2) Extração *Script Python*

Nesta secção aborda-se como foi obtido o texto das notícias. Através do *software Octoparse* apenas se conseguiu extrair os campos URL_Notícia, URL_Página, Data_Extração, Título_Notícia, Fonte_Notícia, Data_Notícia e Sumário_Notícia, tal como referido anteriormente. Porém, além destes campos, no *merged_lines.csv* foram acrescentadas duas colunas (provisórias) denominadas de **normal** e ***goose***, onde se encontrarão os textos das notícias que irão ser analisadas.

Era necessário, após o resultado de uma pesquisa de notícias, como se observa na Figura 3, entrar em cada uma dessas páginas no *website* noticioso onde a notícia foi efetivamente publicada e extrair todo o texto dessa notícia. Assim, surgiu a necessidade de se desenvolver um mecanismo automatizado de forma a resolver esta questão. Desta forma, foi desenvolvido, de raiz, um *script* em *Python*, que através do valor existente no campo URL_Notícia do ficheiro *merged_lines.csv*, vai à página correspondente e extrai todo o texto da notícia. Para chegar a tal resultado, foi necessária alguma pesquisa na *web* para perceber se existia algo que fizesse sentido utilizar de forma a auxiliar este processo.

Foram utilizadas duas bibliotecas diferentes: a normal que funciona baseado num documento *HTML*, onde é retirado o corpo do texto principal e a *goose* que, além de extrair o corpo principal do texto, também extrai a imagem principal (caso exista), assim como vídeos embebidos na notícia. Como é de esperar, o texto extraído pela biblioteca denominada de normal é introduzido na coluna normal do ficheiro *merged_lines.csv* e o texto extraído pela biblioteca *goose* é inserido na coluna *goose* do mesmo ficheiro. Desta forma, conseguiu-se retirar todo o texto das notícias, porém foi necessário existir um trabalho de limpeza do ficheiro que, em bruto, contava com cerca de 15 mil linhas.

Foram feitas várias iterações neste processo de extração, desde a filtragem pela língua Inglesa, apesar de no *Google News* esse filtro já ter sido feito, existiam notícias de outras línguas, como referido anteriormente. Assim, foi realizada esta nova filtragem no *Python*, sobre os campos de cariz textual do ficheiro *merged_lines.csv*, utilizando o package *langdetect*, com o intuito de perceber se a língua era a Inglesa.

De seguida, utilizaram-se vários critérios de filtragem para excluir notícias que não correspondiam ao *output* pretendido. Desta forma, caso o par normal/goose tivesse uma das seguintes combinações – branco/branco, *forbidden*/branco, branco/*forbidden*, *error*/branco ou branco/error – a notícia era eliminada do *dataset*. Após esta filtragem ainda existiam notícias que não interessavam, como as que apresentavam mensagens de erro *standart* de cada *website*. De forma a conseguir reunir todas estas incidências, foram removidos os casos em que o texto da notícia não tivesse mais de cem caracteres. Existiu ainda necessidade de apagar algumas notícias manualmente, que correspondiam a critérios bem definidos e conseguiram passar em todo o processo de filtragem. Finda a extensa filtragem realizada, resultaram as notícias com uma estrutura correta.

Foi necessário transformar as colunas provisórias, normal e *goose*, numa nova coluna denominada de Texto_Notícia. Nesta coluna foram integradas todas as notícias da coluna normal, onde a *goose* tinha um dos seguintes valores: branco, *forbidden* ou *error*. O mesmo processo foi feito nos casos onde a coluna normal apresentava notícias com o valor branco, *forbidden* ou *error*. Nestas ocorrências utilizou-se as notícias existentes na coluna *goose*. Nos acontecimentos em que existiam notícias em ambas as colunas, normal e *goose*, optou-se por escolher a notícia da coluna *goose*, pois era mais completa do que a notícia da coluna normal, apesar de a diferença ser muito pouco significante.

As notícias recolhidas divergiram nas suas fontes, como apresentado na Tabela 3. Nesta tabela podemos observar o *Top* 50 de Fontes Noticiosas de *Big Data*, o número de notícias por cada fonte e a influência de cada fonte relativamente ao número total de notícias.

Tabela 3 *- Top 50 de Fontes de Notícias Big Data.*

Fonte	Nr. Notícias Fonte	% Notícias Fonte
Forbes	678	5,89%
TechTarget	249	2,16%
Smart Data Collective	230	2,00%
VentureBeat	220	1,91%
PR Newswire (press release)	197	1,71%
InformationWeek	179	1,56%
Health IT Analytics	162	1,41%
TechRepublic	146	1,27%
Business Wire (press release)	134	1,16%
ZDNet	128	1,11%
insideBIGDATA	124	1,08%
Information Age	111	0,96%
Datanami	108	0,94%
TechCrunch	97	0,84%
Huffington Post	94	0,82%
V3.co.uk	94	0,82%
Wall Street Journal (blog)	92	0,80%
DZone News	91	0,79%
ITProPortal	86	0,75%
CIO	83	0,72%
ComputerWeekly.com	74	0,64%
GigaOM	74	0,64%
PR Web (press release)	71	0,62%
Fortune	71	0,62%
Data Informed	70	0,61%
Phys.Org	69	0,60%
Network World	64	0,56%
BusinessBecause	63	0,55%
InfoWorld	62	0,54%
Wired	62	0,54%
SiliconANGLE (blog)	59	0,51%
Bloomberg	58	0,50%
eWeek	56	0,49%
Marketwired (press release)	55	0,48%
CRN	53	0,46%
Computerworld	53	0,46%
CMSWire	53	0,46%
RCR Wireless News	52	0,45%
IT Business Edge (blog)	52	0,45%
MedCity News	50	0,43%
Entrepreneur	50	0,43%
Wall Street Journal	47	0,41%
GCN.com	46	0,40%
Healthcare IT News	46	0,40%
BetaNews	44	0,38%
ADT Magazine	43	0,37%
Techworld.com	43	0,37%
The Guardian	43	0,37%
Data Center Knowledge	43	0,37%
Computer Business Review	42	0,37%

Neste caso, a análise foca-se principalmente no posicionamento das TSBD no Quadrante Mágico de *Gartner* segundo um *input*, o *dataset* de notícias. Este constitui o elemento principal de onde os padrões escondidos de conhecimento textual são extraídos. O procedimento é também alimentado com um dicionário de léxicos estabelecido com termos relevantes para a definição das TSBD e dos QSG. Assim sendo, o processo de limpeza e conversão de dados utilizam o léxico contido neste dicionário, de forma a reduzirem o texto das notícias em conjuntos de termos relevantes.

O léxico constituído por ambos os termos de TSBD e dos QSG foi compilado como um único *input*, de forma a ser possível obter-se uma relação entre os dois diferentes domínios analisados. O principal *output* do procedimento de TM é o *document term matrix*, como podemos observar na Figura 7. Esta matriz tem duas dimensões: as notícias de *Big Data* e cada um dos termos considerados. Cada uma das suas células contém a frequência com que cada termo ocorre em cada uma das notícias. Para análise, existe uma tabela de frequências, onde são contadas o número de ocorrências de cada termo, e uma *word cloud*, que nos apresenta uma forma mais fácil de interpretação dessas mesmas ocorrências, como veremos na secção Análise e Exploração dos Dados.

Finalmente, para obter os tópicos mais relevantes, a *document term matrix* serve como *input* para o LDA *topic modeling*. O output final do LDA é uma matriz tridimensional que engloba termos, notícias e tópicos. Assim sendo, é possível obter-se para cada tópico uma relação com os termos do dicionário, através de uma distribuição β. É ainda possível observar-se, para cada notícia, que tópico melhor a exprime.

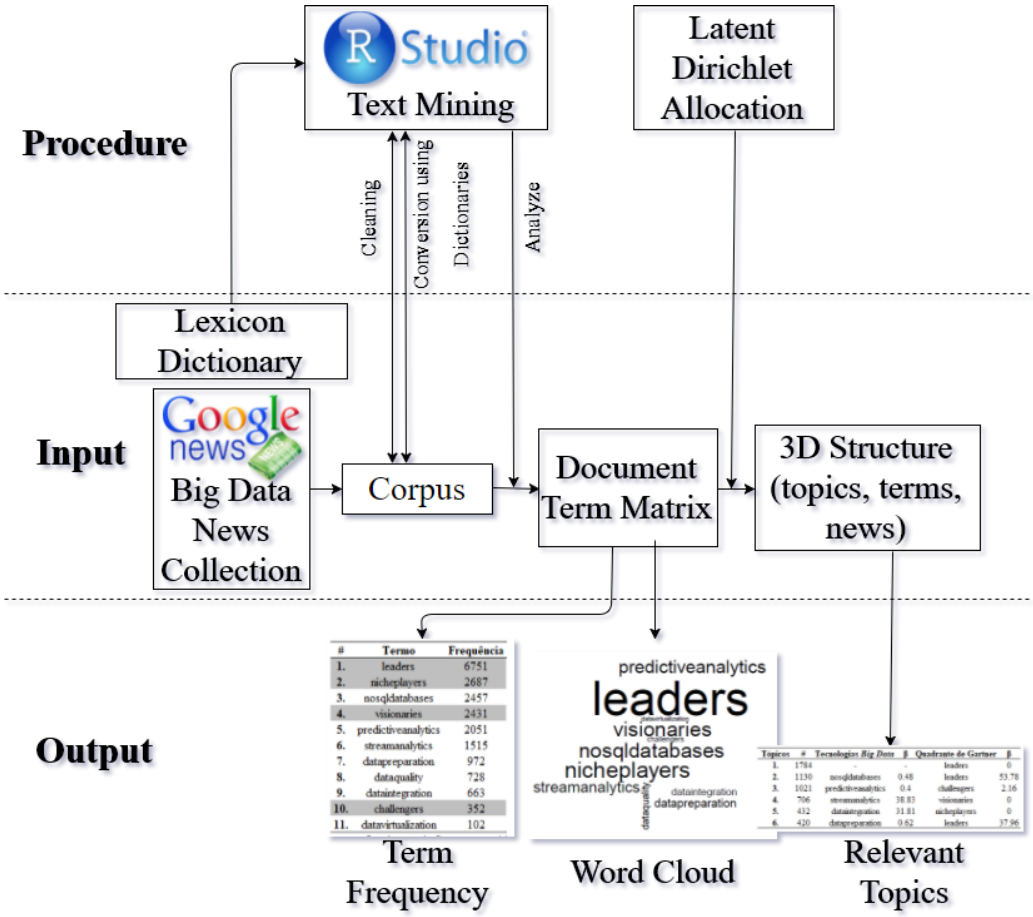

***Figura 7** - Abordagem criada para extração de informação textual não estruturada do dataset.*

Fonte: Adaptado de Calheiros *et al.*, 2017.

Considerando que o objetivo é analisar a relação entre as TSBD e os QSG, apenas os QSG e TSBD mais relevantes serão abordadas futuramente, sendo que foi delimitado um β máximo de 55 a partir do qual se considera que a relação face ao tópico é demasiado ténue para ser relevante.

3.3) *Text Mining* para Análise das Notícias

Esta técnica foi utilizada para avaliar cerca de 11505 notícias de *Big Data*, recolhidas de diferentes fontes noticiosas, tal como referenciado anteriormente.

Deve-se incorporar um dicionário que englobe os termos e conceitos mais comuns das TSBD e um dicionário que englobe as empresas presentes no QG, em vez de deixar o algoritmo TM pesquisar, agrupar e contar palavras sem qualquer tipo de critério, considerando que se pretende uma análise focada.

Visto que este estudo se centra em duas áreas distintas, TSBD e empresas de *Big Data*, é necessário construir um dicionário que atue como base de conhecimento para associar palavras-chaves a conceitos específicos. Assim, dois dicionários distintos foram implementados, um no domínio das principais TSBD baseado na Tabela 4 e outro englobando as principais empresas de *Big Data*, segundo a Tabela 5, cada um contendo uma lista de termos compostos por uma ou mais palavras (*n-grams*). Por se considerarem *n-grams* (Soper & Turel, 2012), este procedimento pode incorporar algum contexto através da combinação de algumas palavras.

Os termos reduzidos apresentados na Tabela 4 e na Tabela 5, não podem conter nenhum tipo de espaço ou caracter especial, como um *underscore* (_), existindo a necessidade de serem compostos por uma só palavra, mesmo que derive de um conjunto de palavras. A título de exemplo, o conjunto de palavras *predictive analytics* transformar-se-á no termo reduzido *predictiveanalytics*.

No processo de construção de um dicionário, as tecnologias específicas de uma empresa foram alteradas, de modo ao dicionário das tecnologias não conter nenhum tipo de empresa, por exemplo a tecnologia “sas data integration studio” foi convertida para “data integration studio”. Todas as palavras foram reduzidas para *lowercase* de forma a facilitar a comparação do termo pelo procedimento de TM.

Tabela 4 - Dicionário de Equivalentes de TSBD.

Termo Reduzido	Termo semelhante ou do mesmo domínio*
predictive analytics	predictiveanalytics,predictive analytics,abm,actian analytics platform,advancedminer,alpine chorus,alteryx analytics,anaconda,angoss predictive analytics,cmsr data miner suite,datarobot,datarpm,dmway,emcien,feature labs,fico model central,gmdh shell,good data,graphlab create,hp haven predictive analytics,information builders webfocus plataform,lavastorm analytics engine,mathematica,matlab,azure machine learning,minitab,data mining odm,portrait predictive analytics,predixion insight,qiware,rapid insight veera,rapid miner,salford systems spm,infiniteinsight,sap predictive analytics,skytree,statistica,tibco spotfire,timi suite,vanguard business analytics suite,viscovery software suite,xlminer
nosql databases	nosqldatabases,nosql databases,hadoop/hbase,cassandra,hypertable,accumulo,simpledb,cloud data,hpcc,flink,splice,mongodb,elastic search,couchbase server,couchdb,rethinkdb,ravendb,marklogic server,clusterpoint server,nedb,terrastore,jasdb,raptordb,djondb,edb,amisa server,densodb,sisodb,sdb,unqlite,thrudb,dynamodb,azure table storage,riak,redis,aerospike,foundationdb,leveldb,berkeley db,oracle nosql database,geniedb,bangdb,scalaris,tokyo cabnit/tyrant,voldemort,dynomite,memcachedb,c-treeace database,kitarodb,hamsterdb,stsdb,tarantool,quasardb,raptordb,activespaces db,nessdb,hyperdex,lmdb,lightning memory mapped database,pickledb,light cloud,hibari,genome,neo4j,infinitegraph,dex,titan,infogrid,hypergraphdb, trinity,allegrograph,white database,virtuoso,vertxdb,flockdb,brightstardb
stream analytics	streamanalytics,stream analytics,apache flink,spark streaming,apache samza,apache storm,streams,software ag's apama streaming analytics,azure stram analytics,data torrent,streamanalytix,sqlstream blaze,event stream processor,stream analytics,tibco's event analytics,striim,informatica,wso2 complex event processor,event stream processing,cisco connected streaming analytics
data virtualization	datavirtualization,data virtualization,actifio sky,cisco data virtualization,datacurrent,denodo,smartcloud data virtualization,informatica data virtualization,data service integrator,red hat jboss data virtualization,federation server,stone bond enterprise enabler
data integration	dataintegration,data integration,actian dataconnect,actian pervasive data integrator,analyza,azuqua platform,bedrock data,businessobjects data integrator,businessobjects data services,ca live api creator,centerprise data integrator,clear analytics,connectall,datacloud,datafuse,dataloader.io,datarush,datasphere, dell boomi,denodo,ediconnect,elastic.io integration platform,infosphere information server,infor cloverleaf integration suite,informatica enterprise data integration,informatica master data management,infosphere data event publisher,infosphere replication

	server,invantive data access point,iway datamigrator,iway enterprise information management suite,iway integration suite,iway parallel service manager,netweaver process integration,data integrator,data service integrator,goldengate,warehouse builder,palantir gotham,phocas business intelligence,pipemonk,powercenter,powerexchange connectors,redpoint convergent marketing platform,redpoint data management,reportminer,reprezen api studio,data integration studio,data management platform,dataflux,enterprise data integration server,federation server,sql server integration services,ssis data flow components,stone bond enterprise enabler,sybase replication server,syncfrog,synscort dmexpress,talend data integration,talend open studio,task factory,unifi,unit4 consolidation,universal adapter framework
data preparation	datapreparation,data preparation,platfora,paxata,datawatch,power query for excel,tamr platform,alterx,clearstory data,lavastorm,teradata loom,spss,looker,informatica rev,sap lumira,trifacta,waterline,datameer,advanced miner,big data analyzer,pentaho 5,dell toad data point,dataworks,enterprise miner,progress easyl,omniscope,infactum
data quality	dataquality,data quality,acquiro leadmatch,clear analytics,data quality batch suite,data quality explorer,data quality manager,data quality monitor,data quality real-time services,datactics data quality suite,datafuse,finscan,health language enterprise terminology platform,hiquality data improver,hiquality identify,hiquality name worldwide,hiquality suite,i/lytics data profiler,i/lytics enterprise data quality suite,infosphere information analyzer,infosphere qualitystage,informatica data explorer,informatica data quality,neteffect,enterprise data quality,reachforce,redpoint data management,redpoint data management for hadoop,business objects data quality management,data insight,data services,dataflux,spectrum enterprise ondemand,tibco clarity,trillium software system,ts discovery,ts insight,validata

*** Todos os termos estão em letra minúscula, separados por vírgulas e sem espaços**

Tabela 5 *- Dicionário de Equivalentes do Quadrante Mágico de Gartner.*

Termo Reduzido	Termo semelhante ou do mesmo domínio*
leaders	leaders,ibm,sas,rapidminer,knime
visionaries	visionaries,microsoft,h2o ai,dataiku,domino data lab,alpine data
nicheplayers	nicheplayers,niche players,fico,sap,teradata
challengers	challengers,mathworks,quest,alteryx,angoss

*** Todos os termos estão em letra minúscula, separados por vírgulas e sem espaços**

A norma utilizada para escolher as principais TSBD foi baseada em diferentes critérios, como o valor acrescentado de negócio ajustado para a incerteza (fundamentado pelo seu potencial impacto, *feedback* e evidencias das implementações feitas e reputação de mercado), a trajetória futura e a fase em que a tecnologia se encontra. Desta forma, apenas se consideraram as tecnologias com um o valor acrescentado de negócio ajustado para a incerteza com um valor médio ou alto, as tecnologias com uma trajetória futura com sucesso significante, representada pela curva superior na Figura 8 e as tecnologias que se encontram numa fase de sobrevivência ou crescimento (Forbes, 2016). Do leque de opções resultante, foram escolhidas as TSBD que se encontram presentes no Dicionário de Equivalentes de TSBD, destacando-se com um retângulo circundante, como se visualiza na Figura 8.

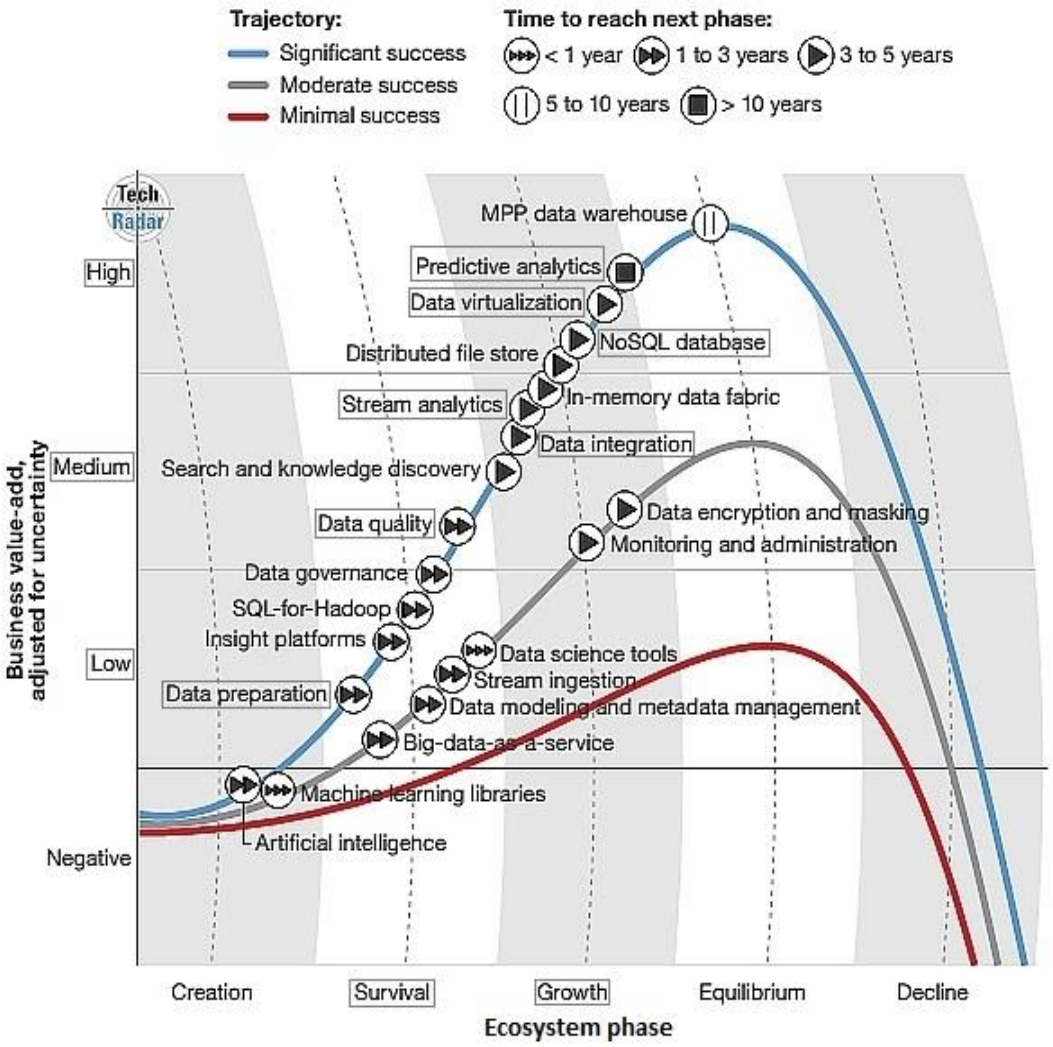

Figura 8 *- TSBD segundo o seu valor para o Negócio e Ciclo de Vida.*

Fonte: Forbes, 2016.

Por sua vez, para definir o dicionário de Empresas de *Big Data* foi utilizado o Quadrante Mágico de *Gartner*, que pode ser definido como um resultado de uma pesquisa num mercado específico, oferecendo uma ampla visão gráfica das posições dos concorrentes desse mesmo mercado. Desta forma, o Quadrante Mágico de *Gartner* divide-se em quatro partes, tal como o seu nome indica (Gartner Methodologies, n.d.).

Existe o quadrante dos *Leaders*, onde se situam as empresas que executam bem a sua atual visão e estão bem posicionadas para o futuro; o quadrante dos *Visionaries*, composto pelas empresas que entendem para onde o mercado está a caminhar ou têm uma visão para alterar a regras do mercado, mas ainda não executam bem; o quadrante dos *Niche Players*, englobado pelas empresas que se concentram com sucesso num determinado segmento de mercado ou estão desfocadas e não inovam ou superam os outros concorrentes; o quadrante dos *Challengers*, constituído pelas empresas que executam bem hoje ou podem dominar um grande segmento de mercado, mas não demonstram um entendimento sobre a direção em que o mercado se desloca (Gartner Methodologies, n.d.).

Figura 9 *- Quadrante Mágico de Gartner 2017 para Advanced Analytics Platforms.*

Fonte: Gartner, 2016.

3.4) Ferramentas de Text Mining

Para realizar o procedimento de TM, a ferramenta estatística utilizada foi o R, um *software open-source*. Existe uma grande comunidade CRAN (Comprehensive R Archive Network – https://cran.r-project.org/), que contribui com *packages* de fácil instalação.

Por sua vez, foi utilizado o *package* tm, tendo sido desenvolvido especificamente para as funções de análise textual (Meyer *et al*., 2008). Este pacote fornece funções para converter dados não estruturados em dados estruturados, reduzindo a dimensionalidade dos dados, mantendo informações relevantes e analisando conjuntamente dados quantitativos e qualitativos (Calheiros *et al*., 2017).

Para construir os tópicos que agrupam as notícias, foi utilizado o *package* "*topicmodels*". Como *input* são recebidas as estruturas de dados resultantes do *package* tm de forma a serem fornecidas as estruturas básicas para o *package* "*topicmodels*" (Hornik & Grun, 2011), onde irá ser implementado o algoritmo LDA (Blei *et al*., 2003).

O algoritmo LDA é um processo de modelação *Bayesian* hierárquico de três níveis, que agrupa um conjunto de itens em tópicos definidos por palavras ou termos, onde cada dos termos identificados caracterizará um tópico (Blei, 2012). Este algoritmo é implementado e pode ser computado com apenas dois parâmetros, o número de tópicos e a *document term matrix* criada para o TM (Calheiros *et al*., 2017).

Este modelo permite analisar a relevância relativa de cada termo usando o valor de distribuição β, que caracteriza a relação entre o tópico e o termo especificado. Um β próximo de zero representa uma relação mais forte entre o termo e o seu tópico correspondente.

3.5) Classificação de Tópicos

Nesta secção, os tópicos são classificados através do modelo LDA, tal como descrito anteriormente. Visto que ambos os dicionários estão fundidos, isto pode significar que um tópico pode ser melhor caracterizado por um dos termos relacionados a uma única categoria. No entanto, esta técnica fornece conhecimentos interessantes sobre as relações entre categorias (Moro *et al*., 2015).

Neste estudo apenas o tópico mais provável para as notícias, segundo o algoritmo LDA, é considerado. Para proceder à sua computação, o número ideal de tópicos é um parâmetro necessário, que pode ser ajustado para resultados ideais (Yi & Allan, 2009).

Foi então utilizado um *package* denominado de "*ldatuning*", onde é possível, segundo algumas métricas, definir um intervalo onde o número de tópicos é ideal. Como resultado obtemos dois gráficos: um deles onde é possível retirar o valor mínimo do intervalo e o outro onde é possível retirar o valor máximo do intervalo, como é possível observar na Figura 10.

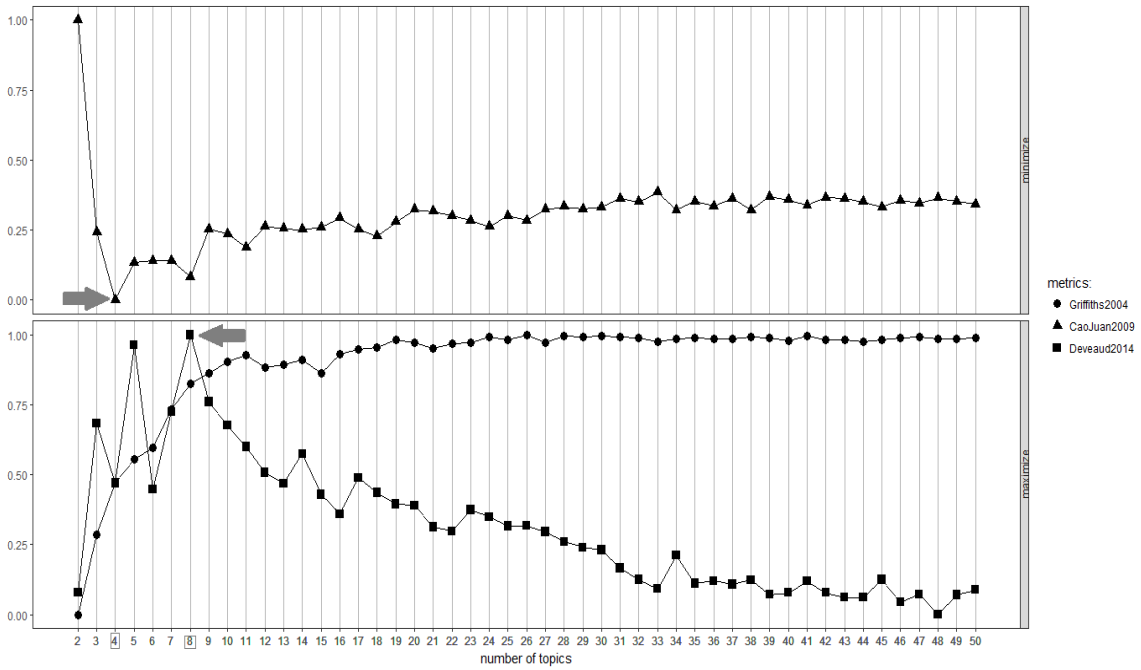

Figura 10 *- Número ideal de Tópicos segundo métricas utilizadas.*

Foi definido que a computação ocorresse até ao limite máximo de cinquenta tópicos, um a um, utilizando os dois *cores* da máquina onde foi realizada. Este tipo de procedimento demora algum tempo a ser efetuado e quanto maior o número de tópicos e menor as capacidades da máquina onde corre, maior duração terá.

Como se observa na Figura 10, delimitado a cinzento por duas setas, o número ideal apresentado no intervalo de tópicos é entre quatro a oito tópicos, tendo sido feitas experiências entre as várias possibilidades, concluindo-se que seis era a escolha perfeita para o número de tópicos, pois o *output* gerado é o que melhor expressa, em valores de β, os diferentes tópicos e termos que os caracterizam, como poderemos averiguar na secção subsequente. As métricas utilizadas basearam-se em estudos feitos sobre o algoritmo LDA e como escolher o número ideal de tópicos, pois é algo muito importante, porém bastante difícil.

Segundo o trabalho realizado por Cao *et al*. (2009), através do estudo entre a melhor estrutura de tópicos e a distância entre os mesmos no LDA, é proposto um método de seleção adaptativa do melhor modelo do algoritmo LDA, baseando-se na densidade. Propõe ainda, conforme as experiências sugerem, que a metodologia proposta atinja a *performance* expectável sem ajustar o número de tópicos manualmente.

Por sua vez, Deveaud *et al*. (2014) utilizou o algoritmo LDA para exibir tópicos altamente específicos relacionados com o *feedback* de documentos altamente relevantes. Os seus descobrimentos demonstram que a abordagem proposta define minuciosamente e eficazmente conceitos latentes.

Por último, Griffiths & Steyvers (2004) desenvolveram um método que descobre um conjunto de tópicos expressados por documentos, fornecendo medidas quantitativas que podem ser utilizadas para identificar o conteúdo desses documentos e expressar a semelhança entre os mesmos. O método é também utilizado para demonstrar como os tópicos utilizados podem ser usados para obter conhecimento em alguma área científica.

4. Resultados e Discussão

Os resultados são apresentados em duas secções distintas: na primeira, os resultados serão analisados com base nas frequências dos termos definidos nos dicionários para o nosso *dataset* global de notícias, composto por cerca de 11505 notícias de *Big Data*, através de técnicas de TM. Estes resultados irão ser apresentados utilizando uma tabela de frequências e uma *word cloud*, que utiliza uma fonte maior, consoante a relevância do termo. Após esta análise, os resultados serão apresentados através dos tópicos gerados pelo algoritmo LDA, que agrupa notícias semelhantes em tópicos. Na segunda secção, serão demonstrados exemplos de notícias por tópico de forma a ilustrar a tendência.

4.1) Análise e Exploração dos Dados

Os resultados obtidos neste procedimento de TM são apresentados na Tabela 5, exibindo o número de ocorrências para cada um dos termos de acordo com as equivalências dos dicionários representados na Tabela 4 e Tabela 5. Os termos destacados a cinzento representam os QSG.

É notório ao olharmos para a Tabela 6, que os *leaders* têm a sua posição consolidada no mercado, sendo de longe a maior diferença numérica comparativamente aos termos de TSBD e QSG, ocorrendo cerca de 6751 vezes num total de 11505 notícias. O procedimento de TM detetou também que num universo de onze termos, três dos QSG encontram-se nas posições iniciais, concluindo-se que, na maioria das notícias de *Big Data*, é possível encontrar o nome das principais empresas presentes no atual mercado e, por sua vez, no Quadrante Mágico de *Gartner*.

Tabela 6 *- Frequências por TBD e Quadrante de Gartner.*

#	Termo	Frequência
1.	leaders	6751
2.	nicheplayers	2687
3.	nosqldatabases	2457
4.	visionaries	2431
5.	predictiveanalytics	2051
6.	streamanalytics	1515
7.	datapreparation	972
8.	dataquality	728
9.	dataintegration	663
10.	challengers	352
11.	datavirtualization	102

Nota: Os termos relativos aos QSG estão identificados a cinzento para uma melhor identificação.

A Figura 11 mostra-nos a *word cloud* para os termos dos QSG, assim como dos termos de TSBD, proporcionando uma interpretação visual dos resultados.

Os resultados visíveis na Figura 11 contabilizam um total de onze termos, demonstrando que o termo *leaders* é, inequivocamente, o de maior destaque, segundo as notícias escrupulosamente esmiuçadas. Assim, é possível evidenciar a forte presença das empresas líderes de mercado nas notícias, realçando também a relevância dos termos *nicheplayers*, *nosqldatabases*, *visionaries* e *predictiveanalytics*.

Figura 11 *- Word Cloud das TSBD e Quadrantes de Gartner.*

De seguida, irá ser abordada uma análise mais interessante para o estudo, pois será possível perceber a relação entre os QSG e as TSBD consideradas, sendo possível avaliar a força da relação. Através do uso do algoritmo LDA foi possível gerar tópicos agregadores de notícias, sendo o resultado final demonstrado na Tabela 7.

Tabela 7 - Agrupamento por Tópicos segundo o algoritmo LDA.

Tópicos	#	Tecnologias *Big Data*	β	Quadrante de *Gartner*	β
1.	1784	-	-	leaders	0,00
2.	1130	nosqldatabases	0,48	leaders	53,78
3.	1021	predictiveanalytics	0,40	challengers	2,16
4.	706	streamanalytics	38,83	visionaries	0,00
5.	432	dataintegration	31,81	nicheplayers	0,00
6.	420	datapreparation	0,62	leaders	37,96

Representa o número de notícias por tópico; β corresponde à correlação entre o tópico e o termo.

Cada tópico é representado por uma linha horizontal da Tabela 7, sendo na coluna Tópicos enumerados cada um dos tópicos; na coluna #, tal como referenciado na legenda da tabela anterior, estão representados os números da notícia associados a cada tópico; na coluna Tecnologias *Big Data* podemos observar o atributo mais relevante das TSBD para o tópico em questão; na coluna Quadrante de *Gartner* estão representados os atributos mais significantes dos QSG para o tópico em questão; e na coluna β apresentam-se os valores de correlação do termo anterior, sendo que quanto mais próximo de zero este valor se encontrar, mais forte é a relação;

Para cada tópico existe um termo de TSBD dominante com um valor β que o aproxima de um determinado termo do QG. Assim, dado que os dois termos mais relevantes são mostrados para cada tópico, um relativo às TSBD outro referente ao QG, podemos analisar cada um dos tópicos relativamente ao QG expresso pelo atributo das TSBD.

Numa primeira análise, existem algumas características interessantes relativas aos tópicos como, por exemplo, o facto de existirem TSBD diferentes para cada um dos tópicos e cada uma delas representar um QG, com a exceção do Quadrante *leaders* que é representado por mais do que uma TBD. Confirma-se assim a análise feita anteriormente, onde se verificava uma grande predominância e importância do QG *leaders* neste estudo. As empresas que compõem este quadrante são as líderes de mercado, oferecendo uma panóplia de tecnologias integradas na sua solução e são as que mais vezes aparecem nas notícias analisadas.

Tratando-se de TSBD e trabalhando com volumes de dados dessa mesma escala, foi estipulado para esta análise, como referido nas secções precedentes, que o β tem um valor máximo de 55.

Da mesma forma como se sucedeu com o trabalho de Moro *et al*. (2015), com uma diferença entre o termo dominante com o valor 0,03 e o segundo termo mais influente com o valor 4,35, aqui também existe uma diferença, sendo a maior compreendida entre 0,48 e 53,78. Estes valores enaltecem a fraca correlação descoberta no estudo de Moro *et al*. (2015).

Para o tópico número 1, com um valor de 1784 notícias identificadas, não existe TBD nenhuma com um β inferior a 55, ou seja, este tópico claramente não se identifica com nenhuma tecnologia específica, o que revela um pouco a ideia de os *leaders* não comercializarem um produto, mas sim uma solução global, tal como se pode comprovar na Figura 12. Por outro lado, a relação demonstrada pelo β relativamente ao QG é muito forte, tendo esta o valor de 0. É assim possível afirmar que este tópico se relaciona perfeitamente com os *leaders*.

O tópico número 2 encontra-se com 1130 notícias reconhecidas, sendo que o β tem o valor de 0,48, identificando-se assim, claramente, este tópico com a TBD *nosqldatabases*. O valor do β relativamente à relação entre o QG e o tópico, que já vimos que se relaciona com a tecnologia *nosqldatabases*, é de 53,78, o que significa que este tópico pouco se relaciona com os *leaders*, levando à conclusão expressa anteriormente: os *leaders* não se focam numa tecnologia específica.

Por sua vez, o tópico número 3 apresenta-se com 1021 notícias representativas, tendo um β de 0,4 relativamente à TBD *predictiveanalytics*, podendo afirmar-se que este tópico é muito relacionado com esta tecnologia. Além disso, a relação entre o QG e o tópico tem um valor de β igual a 2,16, o que tendo em conta o panorama da Tabela 7, pode-se afirmar que este tópico é relacionado com este Quadrante. Assim sendo, conclui-se que a TBD *predictiveanalytics* e o QG *challengers* se encontram relacionados entre si, demonstrando que os *challengers* não refletem um entendimento sobre o rumo em que o mercado se desloca (Gartner Methodologies, n.d.) e como tal investem numa tecnologia que representa um nicho de mercado, como é demonstrado posteriormente, na Figura 12.

Para o tópico número 4, existem cerca de 706 notícias. Este tópico não se identifica muito com a TBD *streamanalytics* com um valor de β igual a 38,83. É possível afirmar-se que este tópico é perfeitamente representado pelo QG *visionaries*, em que o β tem o valor de 0.

Relativamente ao tópico 5, existem 432 notícias identificadas para este tópico. A TBD *dataintegration* tem um β com o valor 31,81, não sendo assim possível dizer que este tópico é representado por esta tecnologia.

Por fim, o tópico 6 apresenta-se com 420 notícias. Pode afirmar-se que tem uma forte relação com a TBD *datapreparation*, com um valor de β de 0,62. Relativamente ao valor de β para o QG, este é de 37,96, o que se reflete na conclusão que os *leaders* não apresentam uma relação significativa com este tópico (e com a TBD que representa o tópico), sendo que, uma vez mais, os *leaders* não se cingem a uma única Tecnologia, como será analisado de seguida.

A Figura 12, que se apresenta de seguida, representa uma interpretação visual dos resultados apresentados na Tabela 7. Apresenta as relações entre as TSBD e os QSG. É também de realçar a apresentação das restantes ligações não identificadas na Tabela 7, mas interessantes para a análise e justificações anteriormente feitas.

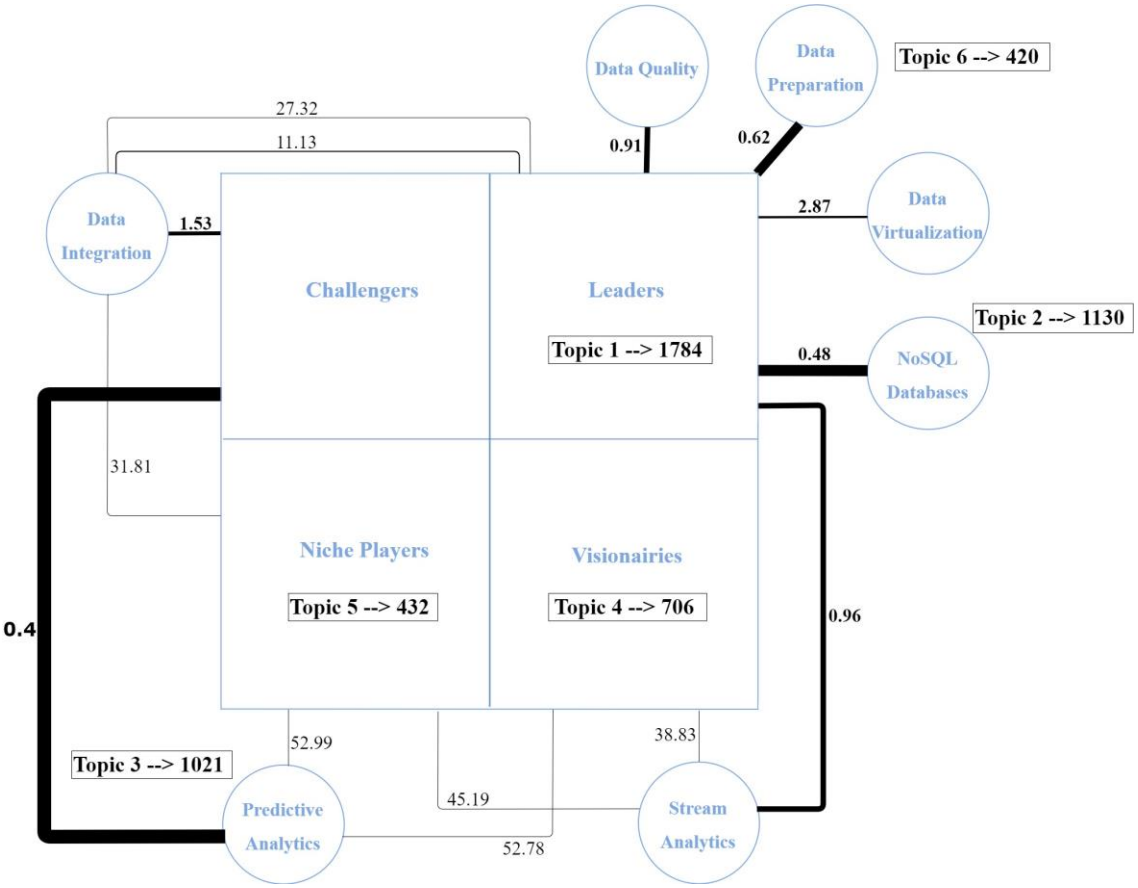

Figura 12 *- Relação entre as TSBD e os Quadrantes de Gartner.*

Através da Figura 12, é possível comprovar a ideia anteriormente demonstrada: os *leaders* comandam de tal forma o mercado, que não se focam num determinado produto ou numa tecnologia específica, o que é oferecido é uma solução que envolve uma panóplia de tecnologias, como se pode comprovar pela forte relação entre os *leaders* e as tecnologias *stream analytics*, *nosql databases*, *data virtualization*, *data preparation* e *data quality*.

No quadrante dos *leaders*, como foi abordado na secção *Text Mining* para Análise das Notícias, situam-se as empresas que executam bem a sua atual visão e estão bem posicionadas para o futuro, enquanto no quadrante *challengers* encontram-se as empresas que executam bem hoje ou podem dominar um grande segmento de mercado, mas não demonstram um entendimento sobre a direção em que o mercado se desloca (Gartner Methodologies, n.d.). Seguindo esta abordagem, é possível afirmar que as tecnologias que estejam diretamente relacionadas com os *leaders* implicam sucesso atual e futuro, pois as empresas ganham reconhecimento e prestígio no mercado devido

ao produto que oferecem aos seus clientes. Assim, uma conclusão interessante que é possível retirar através da Figura 1, baseia-se no facto de não existir uma ligação entre os *leaders* e a tecnologia *predictive analytics*, porém existe uma forte relação entre esta mesma tecnologia e os *challengers*, indicando que a esta tecnologia não será uma aposta futura com valor de mercado ou tratar-se-á de um nicho de mercado.

O mesmo se comprova com a tecnologia *data integration*, que apresenta fracas relações com ambos os tópicos onde o QG predominante é o *leaders* e, por sua vez, apresenta uma relação significativamente forte com o QG *challengers*.

É ainda de destacar que os *visionaries* são definidos como um conjunto de empresas que sabem para onde o mercado está a caminhar, mas ainda não executam bem (Gartner Methodologies, n.d.). A sua fraca relação com a TBD *streamanalytics* representa isso mesmo, pois sendo esta uma TBD muito relacionada com o QG *leaders*, como observado na Figura 12, implica que será uma aposta futura com valor (Gartner Methodologies, n.d.) e que os *visionaries* ao apostarem na TBD *streamanalytics* demonstram astúcia referente ao futuro, porém a aposta feita ainda não é suficiente.

4.2 - LDA e Validação dos Tópicos

Na secção anterior, foi aplicado o algoritmo de LDA com a finalidade de agrupar notícias por tópicos, caracterizados por termos, como apresentado na Tabela 7. Porém, sendo esta uma abordagem automática, contém algumas limitações, como o agrupamento de documentos ser totalmente dependente da técnica utilizada para criar os *clusters*, que por sua vez é baseada na identificação do termo (Thomas *et al*., 2011).

O problema principal resume-se a certos termos terem um significado diferente baseados no contexto em que estão inseridos, sendo a deteção de subjetividade uma tarefa ainda bastante difícil de endereçar. Nesta secção, aborda-se este tema identificando uma notícia representativa de cada tópico, como é mostrado nas próximas páginas. Esta abordagem refletiu-se numa minuciosa análise manual das notícias, a fim de confirmar as hipóteses sugeridas pelos tópicos encontrados, numa perspetiva semelhante à proposta realizada por Moro *et al*. (2015), como é possível observar na Tabela 8.

Tabela 8 *– Exemplo de uma Notícia Representativa para cada Tópico.*

Número do Tópico	Texto da Notícia	Tecnologia *Big Data*	Frequência Tecnologia *Big Data*	Quadrante de Gartner	Frequência Quadrante de Gartner
1.	John Kelly III, senior vice president and director of IBM Research, addresses those gathered during an event at RPI on Wednesday, Jan. 30, 2013, in Troy, NY. The event was held for IBM and RPI to announce that IBM will provide a modified version of the IBM Watson system to the college for use by students and faculty. (Paul Buckowski / Times Union) ... "The docs are blown away by what Watson can do" sifting through vast amounts of medical information in a very short time, Kelly said. Named after IBM founder Thomas J. Watson, the computer is capable of analyzing unstructured data such as language and images. While Watson will be used as a research tool, it also will be expanding its capabilities while at RPI. And university officials expect it'll be a lure when it comes to recruiting researchers, faculty and students to the Troy campus. "(T)he experience of working on Watson will give our students an advantage as they compete for the best jobs in Big Data, analytics and cognitive computing," said RPI President Shirley Ann Jackson. RPI already has an IBM supercomputer housed at the Rensselaer Technology Park, and Jackson and Kelly said the computers, with their different strengths, could work together. The supercomputer, for example, can perform calculations that Watson can't, said Kelly. Watson's time at RPI was made possible by a three-year Shared University Research Award from IBM Research.	-	-	leaders	42

2.	Forrester examined the following 15 vendors: Aerospike, Amazon Web Services (AWS), Basho Technologies, Couchbase, DataStax, Google, IBM, MarkLogic, MapR Technologies, MongoDB, Microsoft, Neo Technology, Oracle, OrientDB and Redis Labs. Among those, MongoDB was a standout. Although Forrester didn't provide a strict numerical ranking of the offerings, MongoDB was positioned near the top of both chart axes in the Wave graphic and led the numerical scores in the Deployment category, trailing only Couchbase (4.48 to 4.40) in the Current Offering weighting. Forrester also started its top 9 evaluations with MongoDB. "MongoDB remains the most popular NoSQL database," the report said. "MongoDB is an open source NoSQL document-oriented database optimized for natively storing, processing and accessing documents and other types of data sets.	nosqldatabases	18	leaders	2
3.	The 1 million-member National Association of Realtors launched a new predictive analytics department today with the rehire of industry veteran Todd Carpenter... Today, Carpenter begins his duties as NAR's managing director of data analytics in NAR's new predictive analytics group, which will work closely with the Center for Realtor Technology, NAR's research division, and other NAR groups to create analytics from NAR's treasure trove of real estate-related data and other sources as needed... Predictive analytics "is something that's kind of brand-new to the industry," Carpenter told Inman News. And as someone who has spent much of his career helping the industry understand complicated technological problems, this "sounded like a great new job," he said. Carpenter's new role is similar to his old role at NAR in that, as the trade group's social media manager, he helped NAR's leadership and its members understand a particular concept — social media.	predictiveanalytics	10	challengers	10

4.	News Azure Data Lake Service for Big Data Analyses Now Available Microsoft this week released Azure Data Lake as a generally available (GA), production-ready service, backed by Microsoft's 99.9 percent service-level agreement. Azure Data Lake is a service for "Big Data" massively parallel types of analyses, with the ability to tap into pools of structured and unstructured data without limits. The service has been at the preview stage since November of 2015, according to a Microsoft Channel 9 presentation, so it's taken one year to arrive fully baked. Microsoft is marketing the Azure Data Lake service as enabling "Big Cognition." The idea is glean insights from multiple inputs of various data types. It's about "joining all the extracted cognitive data with other types of data, so you can do some really powerful analytics with it," according to a Microsoft announcement. Azure Data Lake Components Azure Data Lake is composed of three Azure services, according to the presentation. It has HDInsight, which is Microsoft's Hadoop-based Big Data service. Another component is the new Data Lake Store (GA this week), a repository for structured and unstructured data that can scale to meet developer needs.	streamanalytics	2	visionaries	20
5.	SAP provides the SAP Data services to help migrate on premise data to any size of cloud service. The cloud services supported include Microsoft Azure, Amazon S3, and Google Cloud Storage. Included in the update is the ability to read data from HPE Vertica as well as process Microsoft Outlook data files. This allows business to include data in their analytics from messages, calendar events and archived items. Helping to analyse your big data The enhancements to SAP Agile Data Preparation allow users to prepare data for later analysis. By monitoring data quality, defining data domains and setting up scorecards, users can ensure that the data used in the final analysis is optimised. The SAP Information Steward provides users with information about the quality of the data being used.	dataintegration	3	nicheplayers	22

6.	Datameer is hitting a critical stage of growth as modern BI platforms are becoming fundamental for both new and established companies," said Executive Chairman of Datameer, Mark Burton. "Christian's international track record for building and scaling sales and service organizations makes him uniquely qualified to capitalize on the success Datameer has already seen. We're excited to have him on board and look forward to his leadership." About DatameerDatameer makes big data analytics simple. Datameer gives users a unified, self-service environment to integrate, prepare, analyze, visualize, and operationalize big data. Hundreds of customers, including CIOs, CMOs, CTOs, doctors, scientists, law enforcement officials, and even Olympic athletes all rely on Datameer to help them get from raw data to insight faster than ever. Datameer combines Hadoop's unlimited storage and compute power with a common spreadsheet interface and powerful functionality, quickly transforming businesses into agile, data-driven organizations.	datapreparation	17	leaders	2

Como foi possível observar anteriormente, a coluna Número do Tópico é constituída pela numeração de cada um dos tópicos identificados na Tabela 7; de seguida, na coluna Texto da Notícia, podemos ver cada uma das notícias representativas do seu tópico; segue-se a coluna Tecnologia *Big* Data e Frequência Tecnologia *Big* Data, onde podemos atentar qual é a tecnologia mais representativa na notícia, assim como a sua frequência, respetivamente; por sua vez, encontra-se a coluna Quadrante de *Gartner*, onde é possível examinar qual o quadrante mais significativo da notícia; por fim, deparamo-nos com a coluna Frequência Quadrante de *Gartner*, onde se apresenta a contagem do Quadrante de *Gartner*, identificado na coluna anterior, da notícia.

Todos os termos expressos nas colunas Tecnologia *Big* Data e Quadrante de *Gartner*, dizem respeito ao dicionário abordado na Tabela 4 e na Tabela 5. Além disso, na coluna Texto da Notícia, encontram-se as palavras das TSBD e dos QSG e não os seus termos. Exemplificando: Na notícia do tópico número 1 encontra-se no seu texto a empresa "*IBM*", num total de 42 vezes. Porém, e devido a todo o processo de TM definido na secção da Metodologia, esta empresa é convertida para o seu respetivo termo, sendo o resultado final a contagem do termo "*leaders*" em 42 ocasiões.

De forma a existir critério na escolha de notícias, foi desenvolvida a Tabela 9, constituída por um β e por uma Frequência. Tanto o β como a Frequência dizem respeito quer às TSBD como aos QSG.

O β varia de acordo com os valores da Tabela 7 e foram definidos intervalos de forma a conseguir categorizar as notícias. Por exemplo, uma notícia que possui um determinado valor de β, apenas pode ter uma frequência de TBD e QG entre o valor mínimo e o valor máximo do intervalo correspondente. Por sua vez, as Frequências variam dependendo do valor de β.

Tabela 9 *- Critério de escolha das Notícias para cada Tópico.*

Critérios de Escolha das Notícias de *Big Data*				
β	0	[0,1;2,5]	[2,6; 54]	[55, +∞ [
Frequências	[20, +∞ [	[10,19]	[1, 3]	0

Feita a explicação dos critérios de seleção das notícias, será feita uma análise tópico a tópico, com o intuito de comprovar a análise desenvolvida anteriormente. As notícias foram cortadas, sendo apenas uma parte das mesmas mantidas, de forma a não ocuparem demasiado espaço e, consequentemente, simplificar a tarefa do leitor.

No tópico número 1, é possível observar uma notícia onde não existe qualquer TBD associada, comprovando-se assim o mapeamento do algoritmo LDA realizado para o tópico número 1 da Tabela 7. Testemunha-se também a forte ligação entre o tópico número 1 e o QG *leaders*, sendo a frequência total do termo *leaders* no texto de 42 vezes.

O tópico número 2 é predominado pela TBD *nosqldatabases*, existindo uma relação íntegra entre este tópico e a tecnologia referida. Na notícia destacada, o termo *nosqldatabases* é referenciado 18 vezes. Por sua vez, o QG *leaders* ocorre em 2 situações na notícia, comprovando-se a sua fraca relação com o tópico.

Relativamente ao tópico número 3, pode-se observar o único caso onde existe uma relação significativa entre o tópico, a TBD e o QG. A tecnologia *predictiveanalytics* repete-se 10 vezes no texto da notícia, assim como o quadrante *challengers*.

A análise sobre o tópico número 4 e 5 é, de certa forma, semelhante. Ambos os tópicos apresentam uma relação perfeita relativamente ao QG, porém uma relação fraca e distante relativamente à TBD. No tópico número 4 a tecnologia *streamanalytics* repete-se 2 vezes, enquanto o quadrante *visionaries* é referenciado em 20 repetições. O tópico 5 apresenta-se com uma contagem de 3 repetições da tecnologia e de 22 vezes do quadrante *nicheplayers*, comprovando-se assim a investigação referenciada anteriormente.

Por fim, encontra-se o tópico número 6, onde é possível denotar a forte relação entre o tópico e a TBD *datapreparation*, com uma repetição do termo em 17 vezes, enquanto o QG *leaders* apresenta-se com 2 repetências, testemunhando a fraca relação com o tópico.

Como foi possível comprovar através das notícias destacadas, a análise feita com a informação das mesmas, vai ao encontro do que foi destacado nesta secção em concordância com a Tabela 7.

5. Conclusões

O TM tem sido utilizado em bastantes áreas, porém raramente são exploradas todas as suas capacidades. Neste trabalho, foram aplicadas técnicas de TM a cerca de 11505 notícias de *Big Data*, assim como a identificação das relações inerentes, utilizando a modelação LDA entre dois diferentes domínios: TSBD e os QSG. A singularidade deste estudo apresenta-se desde o método de recolha das notícias sobre um tema não explorado à utilização de dados não estruturados para entender o relacionamento entre as TSBD e o QG onde se integram, podendo concluir-se que determinada TBD é mais utilizada por um quadrante específico de *Gartner*. Este estudo tem uma contribuição interessante para a literatura, pois fornece resultados concretos sobre o comportamento do mercado. Resultados esses provenientes de dados factuais, sem corpo sentimental integrante, visto uma notícia ser meramente um relato de factos e não se exprimir de forma alguma o sentimento do seu redator.

Assim, é possível afirmar que os resultados obtidos revelam o que é falado internacionalmente sobre *Big Data* nas mais variadas fontes de informação, podendo através da análise deste material em bruto, perceber o posicionamento das empresas que compõem cada QG, face aos *softwares* integrantes de cada TBD. Além disso, este estudo transforma-se numa vantagem competitiva para uma empresa de qualquer QG, pois consegue facilmente perceber o seu foco de mercado e o foco de mercado das empresas que constituem os outros quadrantes. Ou seja, é possível uma empresa perceber no que tem de apostar para conseguir chegar ao quadrante pretendido. De uma forma prática e sintética, este estudo enaltece a força das empresas no quadrante dos *leaders*, revelando que estas empresas são cada vez mais líderes de mercado, pois apresentam uma solução de tal forma completa e diversificada, sendo difícil para as outras empresas de diferentes quadrantes igualar um leque de escolhas tão variado. É também possível concluir a não existência de uma ligação entre os *leaders* e a tecnologia *predictive analytics*, porém existir uma forte relação entre esta mesma tecnologia e os *challengers*, comprovando que as empresas que compõem o QG *challengers* não demonstram entendimento sobre a direção em que o mercado se desloca. O mesmo acontece com a tecnologia *data integration*, onde existe uma ténue relação entre os *leaders* e a tecnologia, porém há uma ligação muito forte entre a tecnologia e os *challengers*. Por sua vez, os *visionaries* ao apostarem na TBD *stream*

analytics, bastante relacionada com os *leaders*, demonstram audácia relativamente ao futuro, contudo a aposta feita não é suficiente para alterar as regras do mercado. Uma empresa pertencente ao QG *visionaries* que aposte fortemente na TBD *streamanalytics* verá a sua posição alterada no QG, aproximando-se cada vez mais do Quadrante *leaders* e, ao mesmo tempo, do Quadrante *niche players*.

Uma das limitações deste estudo, que se remate para um futuro aprofundamento, prende-se ao facto desta análise ser feita por um termo Tecnológico (um conjunto de tecnologias representa o termo *data preparation*, por exemplo) face a um QG (um aglomerado de empresas são exibidas por um termo específico, como por exemplo os *leaders*). Pretende-se futuramente estudar e perceber este tipo de relações caso a caso, ou seja, perceber qual a relação para cada uma das tecnologias componentes do termo *data preparation* relativamente a cada uma das empresas que constitui o termo *leaders*. Para isso, considera-se a implementação de um sistema, composto pela parte automática e manual, com o intuído da algoritmia resolver a maioria dos casos (parte automática). Todavia, quando existe dificuldade em assimilar uma determinada notícia a um tópico devido a uma penosa semântica ou a muita subjetividade, pode ser realizada uma escolha manual prudente (parte manual).

6. Referências Bibliográficas

Aggarwal, C. C. (2011). An introduction to social network data analytics. *Social network data analytics* (pp. 1-15). Springer US.

Barbier, G., & Liu, H. (2011). Data mining in social media. In *Social network data analytics* (pp. 327-352). Springer US.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, 28, 45-59.

Bitterer, A. (2011). Hype Cycle for Business Intelligence. *Gartner, Inc., Stamford, CT*.

Blanchard, R., & O'Sullivan, K. (2015). Big data risk and opportunity: having an action plan to address both can add tremendous value to the organization. *Internal Auditor*, *72*(5), 65-67.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling. *Journal of Hospitality Marketing & Management*, *In press*.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7), 1775-1781.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.

Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*, *34*(2), 272-284.

Cukier K., The Economist, Data, data everywhere: A special report on managing information, 2010, February 25, Retirado de http://www.economist.com/node/15557443

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international conference on World Wide Web* (pp. 519-528). ACM.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1), 61-84.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, *54*(4), 86-96.

Ekanayake, J., Li, H., Zhang, B., Gunarathne, T., Bae, S. H., Qiu, J., & Fox, G. (2010, June). Twister: a runtime for iterative mapreduce. In *Proceedings of the 19th ACM*

International Symposium on High Performance Distributed Computing (pp. 810-818). ACM.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2), 293–314.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

Forbes (2016). Retirado de https://www.forbes.com/sites/gilpress/2016/03/14/top-10-hot-big-data-technologies/?utm_content=bufferc6343&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer#6003d51d65d7

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Gartner IT Glossary (n.d.). Retirado de http://www.gartner.com/it-glossary

Gartner Methodologies (n.d.). Retirado de http://www.gartner.com/technology/research/methodologies/research_mq.jsp

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *ACM SIGOPS operating systems review* (Vol. 37, No. 5, pp. 29-43). ACM.

Goes, P. B., (2014). Editor's Comments: Big Data and IS Reseacrh. *MIS Quarterly*, Vol. 38 No., 3 pp. iii-viii.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. *Computer*, *33*(11), 29-36.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98-115.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464-472.

Heidemann, J., Klier, M., & Probst, F. (2012). Online social networks: A survey of a global phenomenon. *Computer Networks*, *56*(18), 3866-3878.

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13), 1-30.

International Technology Group (2011). Retirado de ftp://public.dhe.ibm.com/software/data/sw-library/infosphere/analyst-reports/ITG-ISAS-Exadata-Teradata.pdf

Joshi, P. (2015). Analyzing Big Data Tools and Deployment Platforms. *Int J Multi Approach Studies*, *2*, 45-56.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, *5*(12), 2032-2033.

Laney, D. (2001, February 6). 3-D data management: Controlling data volume, velocityand variety. Application Delivery Strategies by META Group Inc. Retirado de http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. MIT sloan management review, 52(2), 21.

Leavitt, N. (2010). Will NoSQL databases live up to their promise?. *Computer*, *43*(2), 12-14.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, *58*(7), 1019-1031.

Lusch, R., Liu, Y., & Chen, Y. (2010). The phase transition of markets and organizations: The new intelligence and entrepreneurial frontier. *IEEE Intelligent Systems, IEEE Intelligent Systems*, *25*(1), 71-75.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Retirado de http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Marçalo, C. (2014). "O impacto do Big Data nas organizações". *Semana Informática*.

Meng, X. F., & Ci, X. (2013) Big Data Management: Concepts, Techniques and Challenges. *Journal of Computer Research and Development 50*(1), pp 146–169.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of statistical software*, 25(5), 1-54.

Moniruzzaman, A. B. M., & Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, 42(3), 1314-1324.

O'Reilly, T. 2005. "What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software," September 30. Retirado de http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85, 62-73.

Parthasarathy, S., Ruan, Y., & Satuluri, V. (2011). Community discovery in social networks: Applications, methods and emerging trends. *Social network data analytics* (pp. 79-113). Springer US.

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, *33*(2), 193-207.

SAP. (2012, June 26). Small and midsize companies look to make big gains with “big data,” de acordo com uma votação recente conduzida em nome da SAP. Retirado de http://global.sap.com/corporate-en/news.epx?PressID=19188

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)* (pp. 1-10). IEEE.

Soper, D. S., & Turel, O. (2012). An n-gram analysis of Communications 2000--2010. *Communications of the ACM*, 55(5), 81-87.

Sun, J., & Tang, J. (2011). A survey of models and algorithms for social influence analysis. *Social network data analytics* (pp. 177-214). Springer US.

The Economist. 2011. “Beyond the PC,” Special Report on Personal Technology, October 8. Retirado de http://www.economist.com/sites/default/files/special-reports-pdfs/20111008_personal_technology.pdf

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1), 1-14.

Turban, E., Bolloju, N. e Liang, T. (2011). “Enterprise Social Networking: Opportunities, Adoption, and Risk Mitigation”. *Journal of Organizational Computing and Electronic Commerce* 21(3): 202-220.

White, T. (2009). Hadoop: The Definitive Guide O'Reilly. *Scbastopol, California*.

Yi, X., & Allan, J. (2009, April). A Comparative Study of Utilizing Topic Models for Information Retrieval. *In ECIR* (Vol. 9, pp. 29-41).

Zatari, T (2015). Big Data Technologies. *International Journal of Scientific & Engineering Research*.