Universidade
Estadual de Londrina

GUILHERME SAKAJI KIDO

EXTRAÇÃO DE TÓPICOS COM MODULARIDADE NO TWITTER

LONDRINA–PR

2015

GUILHERME SAKAJI KIDO

EXTRAÇÃO DE TÓPICOS COM MODULARIDADE NO TWITTER

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof(a). Dr(a). Sylvio Barbon Júnior

LONDRINA–PR

2015

Guilherme Sakaji Kido
Extração de Tópicos com Modularidade no Twitter/ Guilherme Sakaji Kido. – Londrina–PR, 2015-
67 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof(a). Dr(a). Sylvio Barbon Júnior

– Universidade Estadual de Londrina, 2015.

1. Extração de Tópicos. 2. Modularidade. 3. Microblog. 4. Método de Louvain. I. Prof. Dr. Sylvio Barbon Junior. II. Universidade Estadual de Londrina. III. Departamento de Computação. IV. Extração de Tópico com Modularidade no Twitter

CDU 02:141:005.7

GUILHERME SAKAJI KIDO

EXTRAÇÃO DE TÓPICOS COM MODULARIDADE NO TWITTER

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA

Prof(a). Dr(a). Sylvio Barbon Júnior
Universidade Estadual de Londrina
Orientador

Prof. Dr. Segundo Membro da Banca
Universidade/Instituição do Segundo Membro da Banca

Prof. Dr. Terceiro Membro da Banca
Universidade/Instituição do Terceiro Membro da Banca

Prof. Ms. Quarto Membro da Banca
Universidade/Instituição do Quarto Membro da Banca

Londrina–PR, 24 de novembro de 2015

Este trabalho é dedicado aos meus heróis - meus pais.

AGRADECIMENTOS

Agradeço aos meus pais pelo total apoio e incentivo nesses anos de graduação.

Ao grupo Hatsumi Taiko e Ishindaiko, ou melhor dizendo, à família Hatsumishin pelas melhores experiências de vida.

Aos professores pelos ensinamentos e influências no meu crescimento profissional e pessoal.

Ao meu orientador, pelos ensinamentos, paciência e compreensão desde os primeiros trabalhos do grupo REMID e no projeto de iniciação científica.

"Toda ação humana, quer se torne positiva ou negativa,
precisa depender de motivação."
(Dalai Lama)

KIDO, G. S.. **Extração de Tópicos com Modularidade no Twitter**. 67 p. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina–PR, 2015.

RESUMO

Devido a era da disseminação da informação em larga escala, a necessidade de encontrar informações específicas em grandes bancos de dados proporciona o marketing e a competitividade nas empresas. Redes Sociais Onlines são os meios de comunicações mais utilizadas atualmente, sendo os serviços de microblogs, como Twitter, mídias sociais que ganham destaque por fornecer valiosas informações em poucos caracteres. Com a Mineração de Texto, o presente trabalho propõe um modelo de extração de tópicos por meio do método de Louvain, aplicando em textos desestruturados e semi-estruturados do Twitter. Os resultados serão discutidos através de um especialista humano.

Palavras-chave: Extração de tópicos. Modularidade. Louvain. Twitter.

KIDO, G. S.. **Topic Extraction with Modularity on Twitter**. 67 p. Final Project (Bachelor of Science in Computer Science) – State University of Londrina, Londrina–PR, 2015.

ABSTRACT

Due to the era of information dissemination on a large scale, the need to find specific information in large databases provides the marketing and competitiveness in companies. Onlines Social Networks are the most media used now, where the microblogging services, like Twitter, are social media that are highlighted to provide valuable information in a few characters. With the Text Mining, this paper proposes a model of topics extraction by the Louvain method, applying in unstructured and semi-structured texts of Twitter. The results will be discussed using a human specialist.

Keywords: Topic Extraction. Modularity. Louvain. Twitter.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MD	Mineração de Dados
MT	Mineração de Texto
RSO	Rede Social Online
AVDF	Adaptive Distribution of Vocabulary Frequencies
KDD	Knowledge Discovery from Data
KDT	Knowledge Discovery from Text
PLN	Processamento de Linguagem Natural
TF-IDF	Term Frequency – Inverse Document Frequency

SUMÁRIO

1 INTRODUÇÃO

Atualmente o mundo encontra-se na era da explosão da informação, sendo sobrecarregado com grandes quantidades de dados e informações. Se grande parte desses dados estivessem em forma lógica e estruturada, o alcance da informação multiplicaria muitas vezes. Infelizmente 90% das informações disponíveis estão em formatos desestruturados [4]. Procurar conhecimentos em grandes quantidades de dados desestruturados é difícil e problemático. Para empresas que usufruem de informações para melhor estratégia de marketing de produtos, qualquer conteúdo adicional descoberto pode resultar em melhores planejamentos e assim atrair mais consumidores para determinado produto. Satisfazer o usuário final é uma tarefa desafiante. O objetivo é encontrar informações específicas dentro da infinidade de informações disponíveis, um termo específico dentro de um grande volume de dados existentes.

Além das empresas de marketing que necessitam analisar os dados para identificar o perfil do consumidor e por meio deste elaborarem melhores atrativos, existem outras aplicações onde a extração de informação é um importante fator para o sucesso. De acordo com Chitra et al. [5], através das informações extraídas, é possível analisar situações de clientes de uma empresa de banco, identificando padrões. O banco poderá elaborar melhores planejamentos, oferecendo incentivos individuais para cada necessidade de cliente, de modo a deixá-lo satisfeito e eliminando a possibilidade do cliente trocar de empresa, além de ser um novo atrativo. Perder clientes pode custar muito caro para a empresa. Para Soelistio et al. [6] é possível extrair sentimentos positivos e negativos a cerca de um político específico através da análise de artigos digitais. Já para Choi et al. [7], por meio da análise de artigos, é possível identificar possíveis documentos que tratam de terrorismo, que podem ser utilizados por um governo como uma medida de segurança nacional. Existem outras aplicações em que a extração de informações pode ser importante, tais como: cartão de créditos, telemarketing, medicina e segurança (identificação de hackers e de seus ataques) [8].

Para a realização das aplicações mencionadas, essa grande quantidade de dados vem sendo armazenada nos sistemas devido ao grande avanço do hardware, sem o qual avanço antes não seria possível. Além disto, novas estruturas de armazenamento foram desenvolvidas, tais como: banco de dados, Data Warehouses e Bibliotecas Virtuais [9]. Acumular informação é fácil, encontrar informação relevante em demanda pode ser difícil [10]. Nestas situações, técnicas tradicionais não foram adequadas para tratar a manipulação da imensa quantidade de dados na grande maioria de repositórios [8, 11]. A partir da década de 90, o processo de Mineração de Dados surge como solução para extração do conhecimento em grande escala.

Como o próprio já diz, a Mineração de Dados (MD) leva em consideração a extração de informações a partir de qualquer tipo de dado, seja áudio, texto e vídeo, de modo que a acessibilidade e a abundância desses dados mostram-se uma questão de importância e necessidade de encontrar padrões, relacionamentos interessantes e conhecimentos úteis em modelos compreensíveis para o ser humano. O processo completo visa a obtenção de dados em alto nível a partir de dados de baixo nível [11].

Textos são inerentemente dados desestruturados e imprecisos [10]. Para tratar deste tipo de dados, a Mineração de Texto (MT) é uma extensão da MD, referente no processo de extrair padrões interessantes e não-triviais ou conhecimento a partir de documentos de textos desestruturados de diferentes fontes [12]. Por ser um campo multidisciplinar, MT envolve recuperação e extração de informação, análise de texto, clusterização, categorização, visualização e aprendizado de máquina.

Os textos são os dados mais presentes atualmente. Recentes estudos indicam que 80% de informação de companhias estão contidas em documentos de textos [10]. Além de documentos de textos impressos, a produção de documentos digitais não para de aumentar. Pessoas escrevem artigos em websites, fóruns, redes sociais, blogs e e-mails. A acessibilidade desses meios tornou-se algo fácil, rápido e útil.

Rede Social Online (RSO) são ambientes onde as pessoas discutem e expressam pensamentos e opiniões sobre qualquer assunto [13]. Devido ao grande número de textos, as RSOs tem se mostrado extremamente valiosas para companhias de pesquisa de marketing e organizações de opinião pública de modo a encontrar inúmeras opiniões sobre certos tópicos [14]. Porém os métodos utilizados para MT geralmente são aplicados para textos tradicionais da web, como artigos e reportagens, estruturas diferentes dos textos das RSOs [15]. Neste último tipo, existe a presença de abreviações, gírias, erros gramaticais e uso de diferentes idiomas na mesma estrutura. Além disso, há presença de textos feitos por *spams* que interfere nas precisões de resultados. Devido a esses fatos, a MT passa a analisar textos desestruturados das RSO como uma nova área de pesquisa. Em Igawa et al. [14], onde o uso de técnicas de MT adaptadas para RSOs influenciou em seu trabalho de reconhecimento de contas comprometidas (sob influência de *bots*) na rede social Twitter. Como resultado do crescimento deste tipo de dados, técnicas de mineração e visualização são necessárias para análise, agrupamento e entendimento dos conteúdos.

Agrupar textos necessita conhecer seu conteúdo, no qual cada texto pode ser determinado por um ou vários tópicos, palavras-chaves que descrevem o assunto principal tratado. Existem vários métodos para encontrar essas palavras-chaves, sendo a extração de tópicos a área principal dessa atividade. Para Zeng et al. [16], tópico é considerado uma agregação de palavras e sua frequência, que pode ser extraído do documento. Sendo assim, tópico, palavra e documentos são unidades importantes no processo de modelagem de tópico.

Este trabalho têm como objetivo desenvolver uma nova metodologia para extração de tópicos levando em consideração uma solução adaptável para diversos problemas da atualidade. A RSO utilizada para os experimentos será a mídia social Twitter[1], considerada um dos maiores serviços de microblog existentes hoje. Seu conteúdo será extraído, filtrado, processado e visualizado em forma de grafos de modo a formar redes de palavras. A detecção e extração de tópicos será baseada no princípio da Modularidade que determinará comunidades nesses grafos.

A divisão deste trabalho será da seguinte estrutura:

- Capítulo 2 - Problema: referenciamento de problemas da MT em RSOs e MTO;
- Capítulo 3 - Hipótese: apresentação de uma proposta para os problemas mencionados;
- Capítulo 4 - Trabalhos Relacionados: abordagem de outras técnicas e ferramentas que tratam do assunto;
- Capítulo 5 - Fundamentação Teórica: abordagem de processos, técnicas e ferramentas utilizadas neste trabalho;
- Capítulo 6 - Modelo Proposto: apresentação em forma de fluxograma e sua complexidade;
- Capítulo 7 - Metodologia: explicação dos processos utilizados;
- Capítulo 8 - Resultado: discussão dos resultados e com outros trabalhos;
- Capítulo 9 - Conclusão: finalização do trabalho.

[1] www.twitter.com

2 PROBLEMA

RSOs, especialmente microblogs, são serviços que surgiram como um novo meio de comunicação [17] entre indivíduos e organizações. Estes serviços oferecem uma plataforma essencial para usuários compartilharem pensamentos, ideias, status e experiências. Essas informações são ricas para organizações como banco, universidades, governo e marketing, onde muitas equipes mineram e analisam esses dados, que contêm interesses, preocupações e críticas dos usuários e provêm de pontos para as organizações melhorem seus produtos e serviços [18]; na política, estes dados podem ajustar posicionamentos de políticos em respeito a análise de sentimentos de seu público alvo [19].

Extração de tópicos, como LDA [20] e pLDA [21], têm sido popularizadas em documentos de textos tradicionais, que necessitam de enorme quantidade de dados, ou seja, milhares de documentos com milhares de palavras para gerar tópicos coerentes [22]. Como muitos microblogs possuem número limitado de caracteres (ex. Twitter com limite de 140 caracteres), a grande quantidade de dados que essas técnicas tradicionais pedem aumenta a dificuldade da extração de tópicos em blogs. [17].

Textos em microblogs normalmente possuem uma linguagem mais casual e são informais, carregando menos quantidade de informações. Devido ao limite no número de caracteres, usuários publicam de maneira simplificada, utilizando a linguagem coloquial, abreviações, gírias e geralmente utilizam links, emoticons, fotos, vídeos, e entre outros [23]. Embora todas essas características determinam um texto com informações desestruturadas, elas são extremamente valiosas nos quais os métodos tradicionais não as integram em sua análise.

Este trabalho tem como estudar uma nova solução que contorne o problema da extração de tópicos em RSOs, mais especificamente em microblogs. A mídia social utilizada será o Twitter e uma hipótese para o problema será descrita no capítulo a seguir.

3 HIPÓTESE

Para o problema mencionado no capítulo anterior, o objetivo deste trabalho é a elaboração de um modelo aplicado em blogs levando em consideração as características de uma RSO. Primeiramente, a base formada por textos de blogs são desestruturadas e apresentam grande quantidade de ruídos proporcionados pela linguagem coloquial e informal, gírias, abreviações. Para que o sistema tenha resultado mais precisos em relação ao tema retratado na base, é necessário que esses ruídos sejam eliminados. Para eliminação destes ruídos, este trabalho adotará o uso da reta *Adaptive Distribution of Vocabulary Frequencies* (AVDF) [1]. Sendo uma adaptação da lei de Zipf, a AVDF verificá a linearidade dos termos em relação as suas frequências nos dados a serem analisados. Os termo cuja frequência estiver distante da reta AVDF é considerado um ruído ou uma *stopword*. O resultado desse filtragem serão palavras chaves dos *tweets* extraídos.

Os relacionamentos dessas palavras chaves serão atribuídos de acordo com suas coocorrências umas com as outras. A coocorrência de cada par de palavras chaves será o total de *tweets* em que as elas se encontram simultaneamente. O resultado deste processamento será uma lista de adjacência com um peso para cada relacionamento. A lista será os dados de entrada para a geração de um grafo, sendo seus vértices as palavras chaves extraídas após o AVDF e suas arestas, sua coocorrência.

Após a criação do grafo principal, será necessário desmembrá-lo em comunidades através de um processo de clusterização. Este trabalho usará o método do Louvain [3], no qual usa o cálculo da modularidade para determinar comunidades em grafos. A modularidade de uma comunidade mede a densidade de arestas dentro de comunidades em comparação com as arestas entre comunidades.

4 TRABALHOS RELACIONADOS

Neste capítulo será tratado trabalhos da literatura de outros autores que introduzem, mencionem e resolvam o problema citado no capítulo 2.

4.1 Mineração de texto em blogs

A mineração em blogs consiste em técnicas similares à texto e documentos na web, porém importantes distinções tornam-se desafios para o processamento de linguagem natural, como o uso de abreviações, gírias, erros gramaticais e uso de várias línguas no mesmo documento. Geralmente, por se tratar de um conteúdo mais pessoal e informal, são comuns técnicas de Mineração de Opinião e Análise de Sentimento em blogs. O objetivo da pesquisa de opiniões em RSOs é identificar as tendencias sociais emergentes com base em pontos de vista, disposições, humores, atitudes e expectativas dos grupos de partes interessadas ou ao público em geral [24]. Em [25], o trabalho consistiu em um classificador de sentimentos com textos da mídia social Twitter, no qual classificava os texto em objetivos (neutros em sentimentos) ou subjetivos (positivos ou negativos em sentimentos ou opiniões). Utilizando *tweets* como base de dados, estes eram usado como treino para o classificador de sentimentos, baseado no multinominal Naive Baiyes utilizando N-gramas. Uma das maiores aplicações da MO em blogs é na área da política. Um dos objetivos é ter uma antecipação dos impactos das medidas políticas e uma melhor visualização das consequências.

Outra área importante na mineração em blogs é em relação a segurança dos usuários. A identificação de *spams* e possivelmente *bots* são medidas que deixam o usuário de blogs mais seguros e confiantes em relação as suas informações pessoais e conteúdos expostos na rede. Em Igawa et al.[14], o reconhecimento de contas afetadas por *bots* é uma medida de prevenção para que conteúdos maliciosos não se espalhem pela rede. Uma conta legítima que foi invadida por *bots* e partir deste momento começa a publicar conteúdos maliciosos como um *spammer*, esta conta é considerada comprometida e que deve ser alertada ao dono que foi invadida. A aplicação utiliza a comparação de N-gramas dos *tweets* extraídos de cada usuário, verificando se o padrão de escrita do autor é uniforme nos testes denominados. Como conclusão, foi possível possível verificar características da escrita do autor em um conjunto de 100 palavras (entre 6 à 10 *tweets*), dependendo do quanto por *tweet* era escrito.

Em [26], o trabalho consiste em um sistema de classificação de conta no Twitter em *bots*, *cyborgs* e humanos. *Cyborgs* são considerados humanos que produzem conteúdos específicos devido ao uso de outra aplicações. Utilizando o cálculo de entropia das classes,

foi possível verificar que humanos apresentaram alta entropia devido ao comportamento indefinido, já os *bots* e *cyborgs* apresentaram baixa entropia, característico de um comportamento mais periódico. Analisando os textos, a maioria dos *bots* apresentaram grande quantidades de *spams* em seus conteúdos além da presença de urls, do qual facilitou a detecção de uma automação. A partir da análise de 500.000 usuários, o autor infere que a proporção de usuários humanos, *cyborgs* e *bots* é respectivamente 5:4:1 no Twitter.

4.2 Extração de tópicos em blogs

Os sistemas tradicionais de extração de tópicos geralmente eram usados em rádio, TV, blogs, fóruns e outras mídias como fonte de dados, através de uma série de métodos da MD [27]. Os primeiros métodos eram baseados principalmente em *vector space model* utilizando técnicas de clusterização. Com o desenvolvimento da estrutura e compreensão do texto, métodos probabilísticos e analises estatísticas, surgiram método como *Latent Semantic Analysis* (LSA) [20], *probabilistic Latent Semantic Analysis* (pLSA) [21] e *Latent Dirichlet Allocation* (LDA) [28], este muito utilizado em outra áreas da MT.

No trabalho [15], Tsai faz uma análise do método Author-Topic (AT), uma extensão do LDA e para visualização de quais tópicos eram similares com os outros, utiliza o *Isometric feature mapping* (Isomap). O AT foi aplicado na base Nielson BuzzMetrics, blogs que tratam de ameaças de segurança e relatos de incidentes de cyber crimes e vírus de computadores. Os autores tiveram sucesso em seus resultados, porém verificando a presença de ruídos devido ao rotulamento dos blogs pelos usuários, gerando uma despadronização dos dados. Uma solução para este problema seria a redução do domínio dos rótulos.

Comparando as técnicas tradicionais com modelos atuais de extração de tópicos, o trabalho [27] avalia sua metodologia baseada em *vector space model* em relação ao LDA simples. A ferramenta desenvolvida pelo autores, primeiramente, faz o pré-processados de dados, eliminando de stopwords e pontuações. Cada termo resultante do pré-processamento passa por um processo de pesagem determinado pelo algoritmo TF-IDF, retornando um valor que mede a importância do termo em relação a base toda. Por fim, o processo de clusterização utiliza-se o algoritmo *K-means*. A base utilizada neste trabalho era composta por textos do Sina Weibo, microblog popular na região da China. Comparando com o LDA, a conclusão do trabalho determinou que a metodologia desenvolvida pelo trabalho apresentou melhores desempenhos e indexações. As técnicas utilizadas pelo trabalho enriqueciam informações características do texto original.

5 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será tratado os métodos e conceitos necessários para a resolução do problema e utilizados no modelo proposto.

5.1 Mineração de Dados

Mineração de Dados é referente ao processo de extrair conhecimento e informações revelantes, como padrões, associações, mudanças, anomalias e estruturas, em grande quantidade de dados armazenados em banco de dados, depósitos de dados ou outros repositórios de informações [9]. Este processo está associado a outro termo popular conhecido como *Knowledge Discovery from Data* (KDD). O processo de KDD é mais abrangente e envolve um ciclo de etapas/tarefas para a obtenção do conhecimento final a partir de um conjunto de dados *raw* - sem nenhum processamento envolvido. A Figura 1 ilustra as principais etapas do KDD. Para obtenção do dados iniciais, é necessário um objetivo, passo inicial para a preparação da aplicação. É necessário entender e definir os objetivos da aplicação final e o ambiente em que ocorrerá.

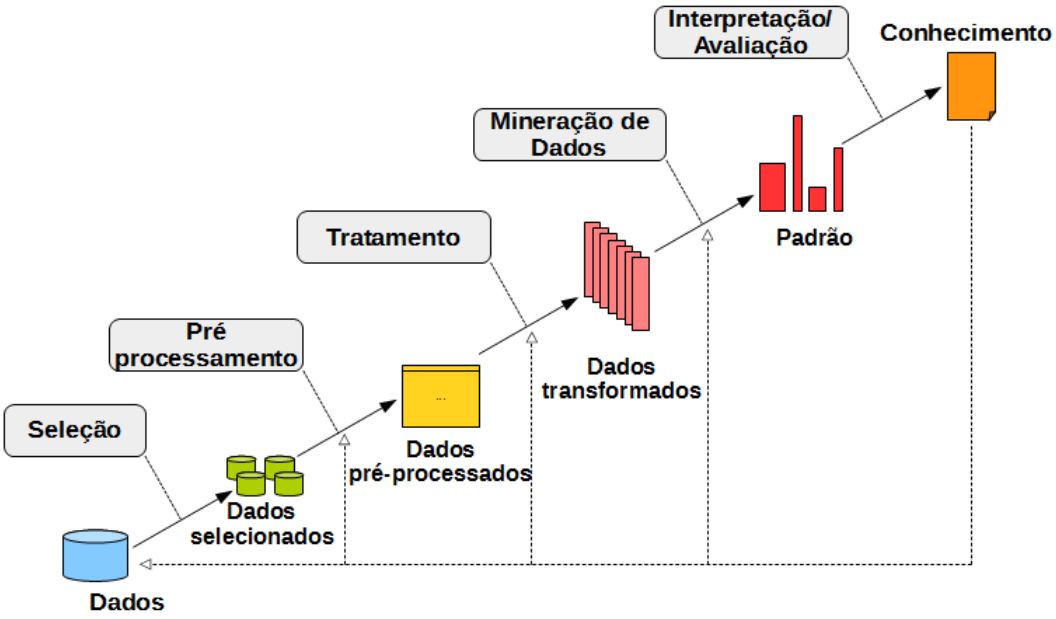


Figura 1 – Etapas do KDD.

5.1.1 Etapas

Após os objetivos serem definidos, os dados que serão usados para obtenção do conhecimento devem ser determinados. Isso inclui encontrar quais dados estão disponí-

veis, obtendo conteúdo adicionais e depois integrando todos os restantes em uma única estrutura. Este processo é importante pois a Mineração de Dados aprende e descobre a partir dos dados disponíveis. Este é a base para o modelo KDD. Se algum atributo importante faltar, toda a pesquisa pode falhar. Quanto mais atributos forem obtidos, menor é o perigo de recomeçar os experimentos.

A seguir, será explicado as principais etapas do KDD:

- **Seleção**. A partir do conjunto total de dados e informações que foram obtidas, esta etapa seleciona os atributos importantes para a análise. Deve ser verificado a importância de cada atributo em relação ao objetivo final.
- **Pré-processamento**. A confiabilidade dos dados é aprimorada. Os dados selecionados são passados por um processo de limpeza com o objetivo de eliminar ruídos - dados desapropriados em relação ao objetivo inicial escolhido.
- **Transformação**. Os dados podem ser separados em subconjuntos e/ou aplicados em métodos como redução na dimensão ou transformação do atributo.
- **Mineração de Dados**. Este processo é dividido em duas atividades:
    1. **Seleção do tipo**. De acordo com o objetivo inicial, decidir qual o tipo de Mineração de Dados usar, por exemplo, classificação, regressão ou clusterização.
    2. **Seleção e aplicação do algoritmo**. A partir do tipo escolhido, selecionar um ou mais algoritmo mais aplicar na base de dados de modo a reconhecer padrões.
- **Interpretação/Avaliação**. Análise do padrão minerado, com respeito em relação aos objetivos adotados.

5.1.2 Tarefas

As tarefas que a Mineração de Dados pode trabalhar são diversas e distintas devido aos diversos tipos de padrões em grande base de dados. São necessários diferentes tipos de métodos e técnicas para encontrar diferentes tipos de padrões. Baseado no tipo de padrão que o objetivo está procurando, as tarefas podem ser classificadas em sumarização, classificação, clusterização, associação e análises de tendências.

- **Sumarização**. É a generalização dos dados resultando em pequenos conjuntos de dados em uma visão geral, geralmente com agregação de informações. Diferentes níveis de agregações e dimensões podem revelar diferentes tipo de padrões e regularidade nos dados.

- **Classificação**. É a derivação em uma função ou modelo que determina a classe de um objeto baseado em seus atributos. Cada objeto é dado como um vetor de atributo e sua classe sendo que um conjunto desses objetos é dado como treinamento. A função ou modelo de classificação será construída a partir desse treinamento, analisando relações entre atributos e suas classes. Essa classificação pode ser usada para classificar objetos futuros e desenvolve um melhor entendimento acerca das classes da base de dados.

- **Associação**. É o descobrimento de união e conexão entre objetos denominados por uma regra de associação. Essa regra revela o relacionamento entre objetos, verificando o quão forte é a relação.

- **Clusterização**. É a identificação de classes/grupos de objetos cujas estas são desconhecidas. Baseado nos atributos, as semelhanças dentro de uma classe são maximizadas e as semelhança entre as classes são minimizadas. Uma vez que as classes são definidas, o objetivo é rotula-las, e as características em comuns dos objetos de uma classe são generalizadas para forma a descrição da classe.

- **Análise de Tendências**. É a identificação de padrões através dos comportamentos dos objetos durante um período. É construído um modelo ou uma função para simular o comportamento do objeto, que pode ser usado para prever comportamentos futuros.

5.1.3 Desafios

Comparações empíricas do desempenho em diferentes aplicações e suas variações tem mostrado que cada abordagem/técnica adotada é melhor em algum aspecto, porém não em todo o domínio. Isso significa que nenhum algoritmo é melhor em todas as possibilidades no domínio de problemas. Se um algoritmo é melhor que o outro em relação ao um domínio, então existem outros domínios que essa situação não acontecerá. No KDD, a solução de um problema utilizando uma abordagem pode produzir mais ou menos conhecimento do que utilizando outras abordagens no mesmo problema.

Em muitos domínios de aplicações, a generalização do erro de até mesmo os melhores métodos está longe do conjunto de treinamento. O objetivo é determinar um erro mínimo realizável. Se classificadores existentes não alcançarem esse nível, precisam-se de novas abordagens. Esse é um dos desafio da Mineração de Dados - não é apenas resolver, mas sim quantificar e entender melhor os problemas. Dentre as finitas soluções, o dilema é verificar qual a melhor estratégia para adquirir o conhecimento, seja utilizando um técnica ou a combinação de pontos fortes de cada uma.

Outro desafio da Mineração de Dados é em relação a quantidade de dados a serem analisadas. O que difere a MD com os métodos tradicionais é a escalabilidade dos dados.

Escalabilidade significa trabalhar com conjuntos com grandes quantidade de registros, dimensionalidade e classes, o que acarreta em problemas de tempo e memória. Grandes base de dados virou sinônimo de *terabytes*, até mesmo *pentabytes*. Apesar de serem um número imaginável de dados, este ramo tornou-se algo normal hoje em dia, no qual o uso de MD para reconhecer padrões é benéfico para empresas.

5.2 Mineração de Texto

A Mineração de Texto, considerada uma subárea da Mineração de Dados, é a procura por padrões em um ou conjunto de textos em linguagem natural e pode ser definido como o processo de análise de textos para extrair informações em um propósito em particular. Como os textos são dados não estruturados, o objetivo maior é transforma-lo em dados para análise (estruturado), por meio da aplicação do Processamento de Linguagem Natural (PLN), métodos estatísticos e técnicas de aprendizado de máquina.

Assim como a MD, a MT também está a associada um processo denominado *Knowledge Discovery from Text* (KDT) no que constituí além de etapas/técnicas semelhantes ao KDD para a busca do conhecimento a partir de análises textuais, inclui também qualquer técnica nova ou antiga que possa ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto. Muitos métodos tradicionais do KDD foram adaptados para suportar esse tipo de informação semi-estruturada ou sem estrutura.

A forma mais comum de armazenamento de informação é através de texto, logo o KDT, teoricamente, tem um potencial maior de utilização do que KDD, pois cerca de 80% das informações contidas nas organizações estão armazenadas em documentos textuais [10].

5.2.1 Etapas

De um modo geral, a MT pode ser dividida nas seguintes etapas: seleção de documentos, seleção da abordagem (análise semântica ou estatística), pré-processamento, indexação, mineração de dados e avaliação dos dados [29]. Estas etapas podem ser visualizadas no diagrama de atividades representado pelo Figura 2.

A seleção de documentos ou também denominado como coleta de dados é a primeira etapa do processo e corresponde a formação da base de dados de textos. Esses textos podem ser retirados de vários tipos de fontes, como artigos, jornais, notícias, blogs, fóruns e livros, dependendo da aplicação em que se deseja. Neste trabalho foram coletados textos de RSOs, mais especificamente *tweets* da mídia social Twitter. O armazenamento desses dados podem ser através de arquivos, banco de dados ou pela web, por meio de um sistema *crawler*, responsável pela extração de dados.

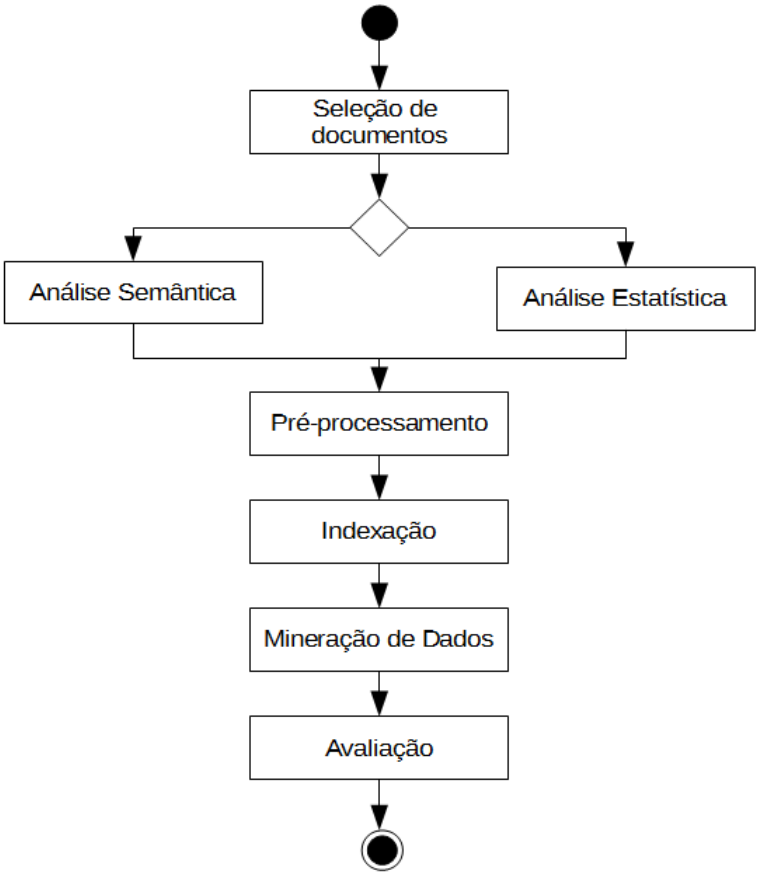


Figura 2 – Etapas do processo de Mineração de Texto.

A segunda etapa consiste na escolha de uma abordagem do problema. Esta abordagem consiste em duas formas:

- **Análise Semântica**. Fundamentada nas técnicas de Processamento de Linguagem Natural (PLN) identificar a um sequência de termos em um contexto de texto, no sentido de identificar qual a sua função.
- **Análise Estatística**. Considera como importante a frequência com que os termos ocorrem no texto, independente do idioma do texto.

A próxima etapa de um processo geral da MT corresponde ao pré-processamento de dados. Pré-processar dados tem como objetivo melhorar a qualidade dos dados disponíveis e organizá-los de uma forma mais fácil para a análise. Dentre os mecanismos existentes para pré-processamento em texto, os mais utilizados são a remoção de *stopwords* e lematização (*stemming*).

Stopwords são palavras comuns como artigos, pronomes e preposições que geralmente não são relevantes para análise de texto. Dependendo da objetivo da aplicação adotado, essas palavras podem ou não serem consideradas ruídos. Por possuírem uma

frequência muito alta em todos os documentos, não são capazes de discriminá-los, não devendo ser indexados.

A lematização ou *stemming* é responsável por reduzir palavras de acordo com seus radicais, ou seja, eliminando os prefixo e sufixos. Características como gênero, número e grau das palavras também eliminados. Como vantagem de se utilizar esse método, o número de índices é reduzido a 50% [29], porém em relação a precisão de buscas, usuários não conseguem pesquisar por palavras mais específicas.

A etapa de indexação leva em consideração que todos os termos não possuem a mesma importância no texto. Os termos mais frequentemente utilizados (com exceção das *stopwords*) costumam ter significados mais importantes, assim como as palavras constantes em títulos ou em outras estruturas. Substantivos e complementos também podem ser considerados mais relevantes que os demais termos. Existem várias técnicas para indexação de termos, sendo o *Term Frequency - Inverse Document Frequency* (TF-IDF) o mais utilizado. Um seleção desses termos pode ser necessária.

A próxima etapa se refere a Mineração de Dados. As áreas mais comuns são:

- **Classificação de Documentos**.
- **Recuperação de Informação**.
- **Clusterização e Organização de Documentos**.
- **Extração de Informação**.
- **Visualização**

Como última etapa tem-se a avaliação que compreende a interpretação dos padrões extraídos, a fim de constatar se os objetivos traçados foram alcançados. Para etapas, pode se utilizar desde medidas estatísticas como precisão e acurácia ou até mesmo o uso de um especialista humano.

5.3 Twitter

O Twitter, criado em 2006, é um serviço microblog capaz de oferecer aos seus usuários a possibilidade de escreveram breves atualizações de no máximo 140 caracteres. Os chamados *tweets* são pequenas mensagens que podem transmitir qualquer ideia, opinião e status de seus usuários, produtos, serviços e eventos. Por se tratar de poucos caracteres, torna-se uma leitura menos cansativa e mais direta.

De acordo com o site oficial da empresa [30], atualmente existem 316 milhões de usuários ativos mensalmente, sendo que 80% destes utilizam a mídia social por meio de

dispositivos móveis. São mais de 500 milhões de *tweets* por dia, o que torna uma enorme quantidade de dados a serem armazenadas.

Em meio ao *tweet*, usuários podem ser referenciados por meio do carácter "", possibilitando assim interações entre usuários. Para que um usuário consiga visualizar o conteúdo de outros usuários é necessário que este siga (*follow*) quem deseja. Quem o usuário segue é considerado *following* e quem segue ele, *follower*.

Outra característica importante dos tweets é a presença de *hashtags*. Determinados pelo caracter "#", as *hashtags* são marcações feitas pelo usuário que possibilita a "rotulação"dos tweets, determinando o assunto que ele possivelmente irá tratar. Dependendo do contexto que esta retratando, as *hashtags* podem alcançar um nível mundial, ou seja, pessoas de diferente lugares do mundo podem se conectar através do uso da mesma *hashtag*, explicitando qualquer informação acerca do contexto, seja ele um evento, produto ou serviço.

As grandes quantidades de dados públicos que fluem através das mídias sociais tem o potencial de oferecer uma nova visão valiosa para o comunidade acadêmica, agências de marketing, ONGs e outras organizações interessadas em entender o comportamento on-line.

5.4 AVDF - *Adaptative Distribution of Vocabulary Frequencies*

O modelo matemático *Adaptative Distribution of Vocabulary Frequencies* (AVDF), proposto por [1] tem como objetivo avaliar o nível de ruído de um conjunto de dados de uma mídia social, aperfeiçoando combinações de técnicas de pré-processamento para eliminação desses ruídos. Cada vez mais, o uso de técnicas de pré-processamento de dados em Mineração de Texto tem-se tornado fundamental para melhores precisões de resultados. Este método foi baseado na Lei de Zipf.

A Lei de Zipf [**??**] é uma medida clássica da literatura que estuda a distribuição de frequência dos termos de um conjunto de dados. Foi desenvolvida por George Kingley Zipf e trata-se de uma de lei de potências (Equação 5.1) da distribuição de frequências dos termos em relação a sua posição no *rank* em ordem decrescente, ou seja, o primeiro termo é o frequente da base inteira e o último, o menos frequente. Seja $f'(t)$ a frequência desejável de um termo t e $r(t)$, sua posição no *rank*.

$$f'(t) \sim \frac{1}{r(t)} \tag{5.1}$$

Isto significa que o segundo termo se repetirá aproximadamente com uma frequência da metade do primeiro, e o terceiro termos com uma frequência de 1/3 e assim sucessivamente.

Ao plotar o histograma dos termos e suas frequências em ordem decrescente é possível verificar a curva Zipf. Se o histograma for plotado em escala logarítmica, a curva passa a se tornar uma reta. A Lei de Zipf representa um padrão distribuição de probabilidade dos termos, no qual é possível perceber que está reta se adapta para todos os termos da distribuição, porém não deixa em evidência os ruídos do conjunto.

Geralmente, os termos mais frequentes de um texto são preposições, artigos e pronomes onde em MT e dependendo do objetivo da aplicação, são considerado *stopwords*. Com o objetivo de eliminar essas *stopwords*, a reta linear AVDF leva em consideração a evidência desses ruídos no histograma de frequência. A partir de dois ponto conhecidos no plano cartesiano, t_1 (termo mais frequente) e t_n (termo menos frequente), é possível encontrar uma reta linear bem como seu coeficiente angular α (Equação 5.2, onde $f(t)$ é a frequência real do termo t). Está nova reta não se adaptará tão bem a todos os termos porém a presença de ruído será evidente. A reta AVDF é dada pela Equação 5.3.

$$\alpha = \frac{log(r(t_1)) - log(r(t_n))}{log(f(t_1)) - log(f(t_n))} \tag{5.2}$$

$$AVDF(t) = \alpha(log(r(t))) + log(f(t_i)) \tag{5.3}$$

A Figura 3 ilustra a aplicação da Lei de Zipf e a reta AVDF em escala logarítmica.

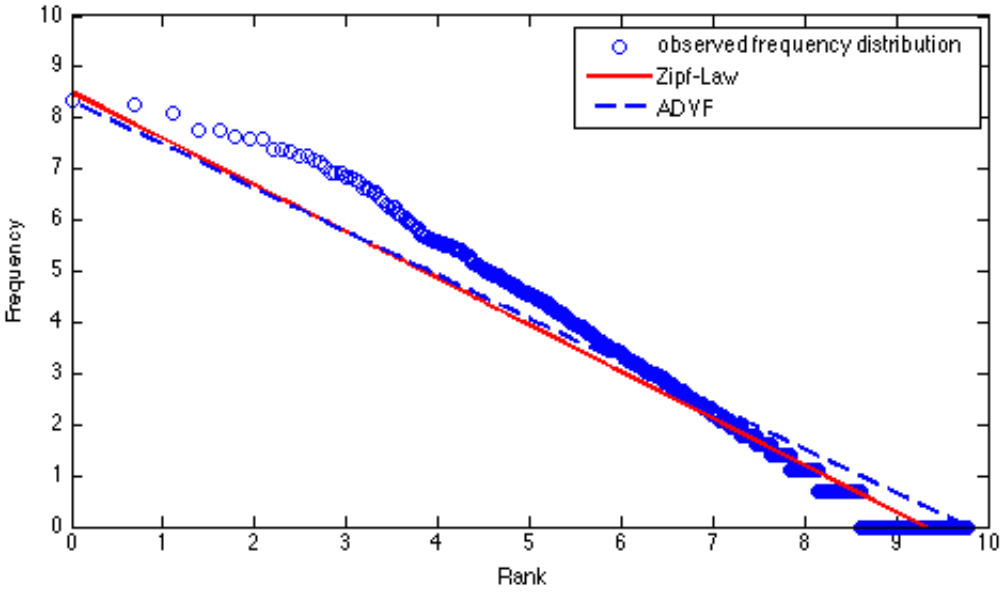


Figura 3 – Exemplo da reta de Zipf e AVDF [1].

5.5 Comunidades

O termo "comunidade"apresenta na literatura diferentes significados e conotações. Em relação as ciências sociais, comunidade se refere à um grupo de pessoas que com-

partilham o mesmo tipo de interesse ou atividade. Uma vez que as redes passaram a ser consideradas modelos para diversos sistemas reais, o conceito de comunidade expandiu, referenciando a estruturas de grupo em uma variedade de redes [31]. Uma nova noção de comunidades surgiu após o crescimento das mídias sociais, apresentando uma variedade de entidade onlines com diversas relações e interações dentre entidades.

O crescimento das mídias sociais proporcionou uma maior interação de seus usuários, de modo que novos tipo de redes são formadas. A grande variedade dessas redes nas mídias sociais chamou grande atenção de áreas como ciência da computação, psicologia, economia, marketing e ciência do comportamento [32]. Umas das principais tarefas é encontrar comunidades cujos os membros possuem uma maior interação com entre eles em relação ao membros de fora. As comunidades extraídas podem ser utilizadas para análise como visualização, marketing, formação e evolução de grupos, clusterização.

Em relação ao uso de grafos, uma comunidade é dado pela divisão de nós em grupos dos quais as conectividades interiores são densas porém entre eles são esparsas (Figura 4). A habilidade de encontrar e analisar esses grupos pode prover um entendimento e visualização mais ampla da estrutura da rede [2]. Para verificar se uma comunidade encontrada pelo algoritmo é boa ou não, este trabalho utilizará o conceito de modularidade, descrito na subseção a seguir.

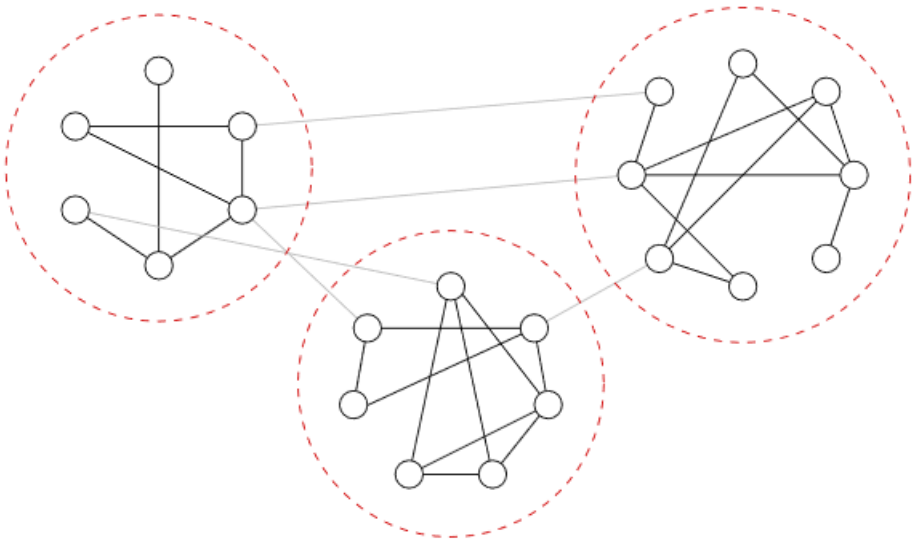


Figura 4 – Divisão de um grafo em comunidades [2].

5.5.1 Modularidade e método de Louvain

O conceito de modularidade [2] fornece uma medição da qualidade de uma comunidade dentro de um rede, quantificando uma força dada pela comparação da fração de arestas dentro da comunidade com arestas entre comunidades. Seja a modularidade Q um valor entre 0 e 1, quanto mais próximo de 1, mais forte são as conectividades dentro da

comunidade. Em redes com pesos, Q é definido de acordo com a Equação 5.4 [3]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{5.4}$$

onde A_{ij} representa o peso entre os nós i e j, $k_i = \sum_j A_{ij}$ é a somatória de pesos entre as arestas que ligam o nó i, c_i é a comunidade em que o nó i pertence, a função δ atribui 1 se as comunidades forem as mesmas, caso contrário 0 e $m = \sum_{ij} A_{ij}$.

De acordo com [3], em seu levantamento, existem diferentes métodos de descoberta de comunidades que utilizam a modularidade para comparar a qualidade dos grupos encontrados, que sempre objetivam uma função de otimização. A optimização de modularidade é computacionalmente caro, onde algoritmos de aproximação são necessários para lidar com grandes redes. Clauset et al. [33] desenvolveram um algoritmo de aproximação rápido para otimização da modularidade em grandes redes. O método consiste em mesclar comunidades recursivamente otimizando a produção de modularidade. Por se tratar de um algoritmo guloso, pode produzir valores menores do que realmente pode ser encontrado. Para contornar o problema citado anteriormente, em Blondel et al.[3] da Universidade Católica de Louvain desenvolveram um método utilizando o conceito de modularidade.

O método Louvain, como será referenciado o trabalho de Blondel, consiste em duas fases que se repetem iterativamente. Dado um grafo com peso de N nós, inicialmente, assume que cada nó é uma comunidade. Para cada nó i e seus vizinhos j, é avaliado o ganho de modularidade entre retirar i de sua comunidade e colocando nas comunidades de j. O nó i assume a nova comunidade onde o ganho de modularidade é máxima e positiva, caso contrário, i continua da comunidade de origem. A fase um é completa até que nenhum melhoramento possa ser alcançado para todos os nós, ou seja, o máximo local é atingido quando nenhuma movimentação pode melhorar a modularidade. O ganho da modularidade ΔQ obtendo a partir da movimentação de i para a comunidade C é demostrado na Equação 5.5:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_i n}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right], \tag{5.5}$$

onde $\sum_{in}$ é a soma dos pesos das arestas dentro de C, $\sum_{tot}$ é a soma dos pesos das arestas incidentes ao nós de C, k_i é a soma dos pesos das arestas incidentes ao nó i, $k_{i,in}$ é a soma dos pesos das arestas de i para os nós de C e m é a soma de pesos de todas as arestas do grafo. Na prática, ΔQ avalia a mudança de modularidade removendo i da comunidade e depois movendo-o para a comunidade vizinha.

A segunda fase constitui na construção de um novo grafo, onde as comunidades (nós agrupados) da fase um, passam a ser os novos nós. O peso das arestas entre dois novos nós é dado pela soma de pesos das arestas entre nós de duas comunidades. Após

a conclusão da segunda fase, a fase um pode ser ativada novamente denominando como "passo"a junção de da fase um e dois. Os passos são iterados até que nenhum ganho de modularidade é alcançavel. A Figura 5 ilustra o funcionamento do método de Louvain.

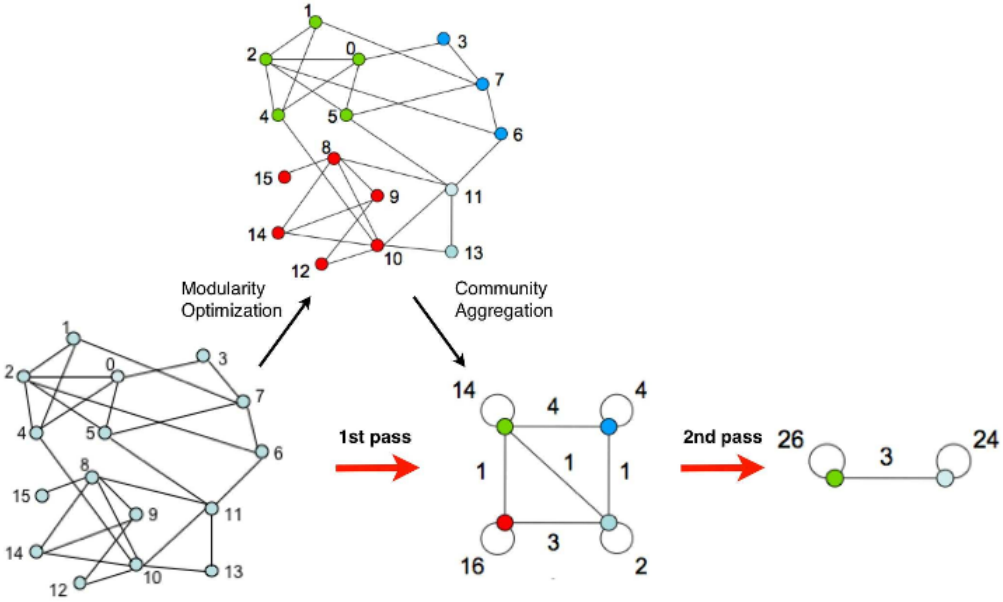


Figura 5 – Aplicação do método de *Louvain* com dois passos [3].

6 MODELO PROPOSTO

De acordo com a hipótese formulada neste trabalho para resolver o problema de extração de tópicos em microblogs, a Figura 6 ilustra os procedimentos adotados.

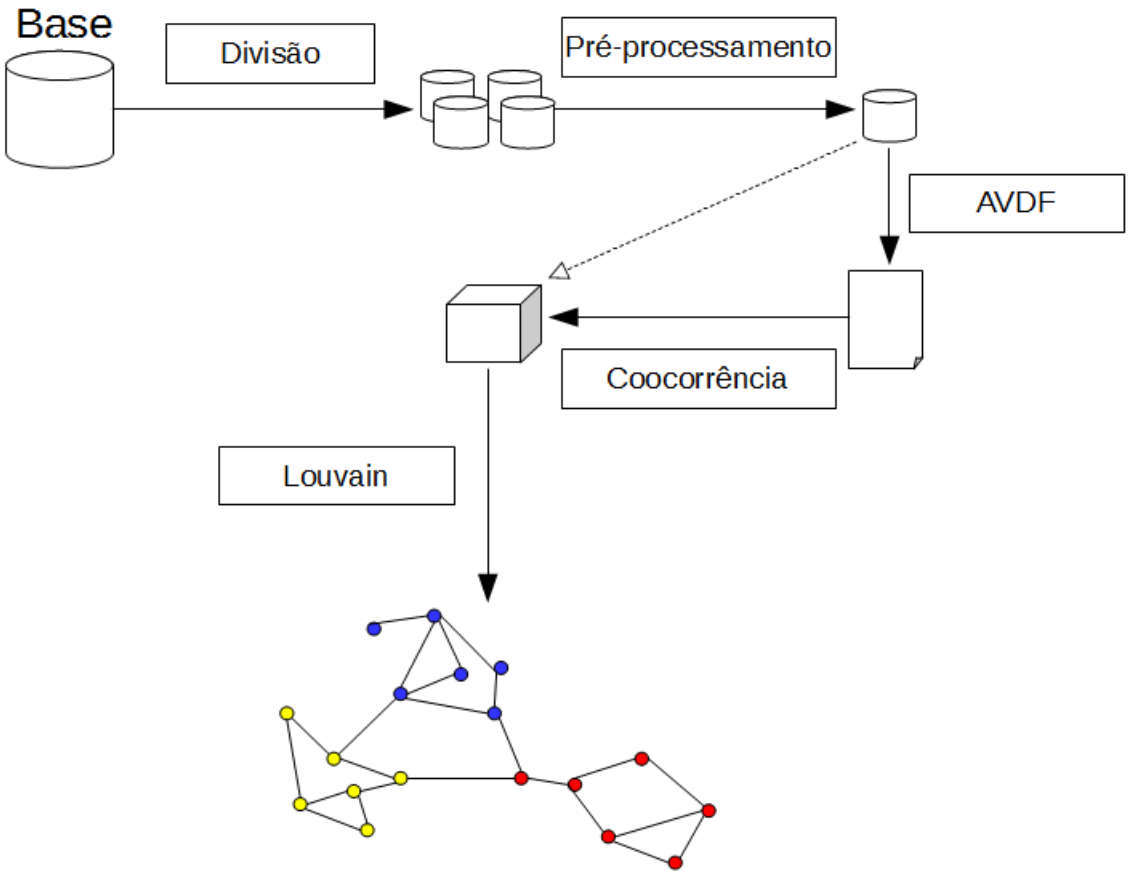


Figura 6 – Modelo Proposto.

O primeiro passo é a aquisição de dados. Em relação a Mineração de Texto, apenas os textos vão ser usados, portanto nenhuma informação adicional além dos *tweets* será necessária para a proposta. Estes dados serão coletados através de um sistema crawler.

Após o processo de coleta de dados e a formação de uma base, é necessário que esta passe por um processo de divisão de acordo com a metodologia utilizada, de maneira que possam ser feitos vários experimentos diferentes a cerca da mesma base de dados. Para cada sub-base adquirida com este processo, é aplicado o pré-processamento.

O pré-processamento adotado para este trabalho leva em consideração as características da estrutura dos *tweets*. O uso de técnicas tradicionais para texto em geral poderiam acarretar em desorganização dos dados e perdas de informações importantes. A base passará por um processo de limpeza (remoção de links, caracteres especiais, espa-

çamentos desnecessários) e pela remoção de *stopwords*. Por fim, as palavras passarão por um processo de tokenização.

Como descrito na Seção 5.4, o método AVDF coloca em evidência os termos que são considerados possíveis ruídos. Esses ruídos podem ser considerados termos que aparecem demasiadamente ou que apresentam poucas vezes devido à erros gramaticais. Após o processo de tokenização e de remoção de *stopwords*, para cada termo da lista de *tokens* verifica sua frequência em relação a todos os *tweets* coletados. A reta AVDF pode ser calculada a partir da Equação 5.3.

São selecionados N termos cuja a Distância Euclidiana (DE) da frequência real e a frequência descrita pelo AVDF de cada termo seja pequena. Quanto menor for a diferença entre a frequência real do termo com a frequência descrita pelo AVDF, maiores são probabilidades deste termo ser um tópico. Na Equação 6.1, calculá-se a Distância Euclidiana:

$$DE(V_i, V_j) = \sum_{m=1}^{n} (V_{im} - V_{jm})^{1/2}, \tag{6.1}$$

onde V_i é a frequência real, V_j é a frequência determinada pelo método AVDF e m é o m-ézimo termo do conjunto total n.

O método AVDF de Igawa et al. [1] é modelo proposto com origens na Lei de Zipf. Como o AVDF é um distribuição linear baseada apenas na frequência do primeiro termo, isso evita processamento extra, mantendo uma complexidade baixa de $O(n)$.

Os N termos selecionados da etapa anterior passam pelo processo de coocorrência. A cada par de termos, verifica-se a frequência deste em cada *tweet* da base, gerando uma lista de adjacência (exemplificado na Tabela 1). Os pares determinarão os vértices do grafos e frequência em pares em relação a base toda será o peso da aresta. Esta lista de adjacência será a entrada para construção do grafo.

Tabela 1 – Representação de lista de adjacência

Nó 1	Nó 2	Peso
palavra1	palavra2	3
palavra1	palavra3	5
palavra2	palavra3	2
...		

O algoritmo implementado para formação da lista de adjacência baseia-se na co-ocorrência dos termos. Para verificar quais pares de termos aparecem no mesmo termo, é utilizado uma busca binária de $O(log(n))$. Sendo m o total de *tweets* da base de dados, a complexidade assintótica do método de coocorrência utilizado é $O(m.log(n))$.

A última etapa do modelo proposto consiste na aplicação do método de Louvain (Equação 5.5) no grafo, dividindo em comunidade possíveis favorecendo a otimização da modularidade.

Embora a exata complexidade computacional do método de Louvain não é conhecida, este método nas maiorias das vezes se comporta como $O(nlog(n))$, onde a maioria do esforço está na primeira fase do algoritmo.

7 METODOLOGIA

Neste capítulo será tratado as definições e configurações de cada método utilizado neste trabalho. A Figura 7 representa a metodologia adotada para definição de tópicos em microblogs.

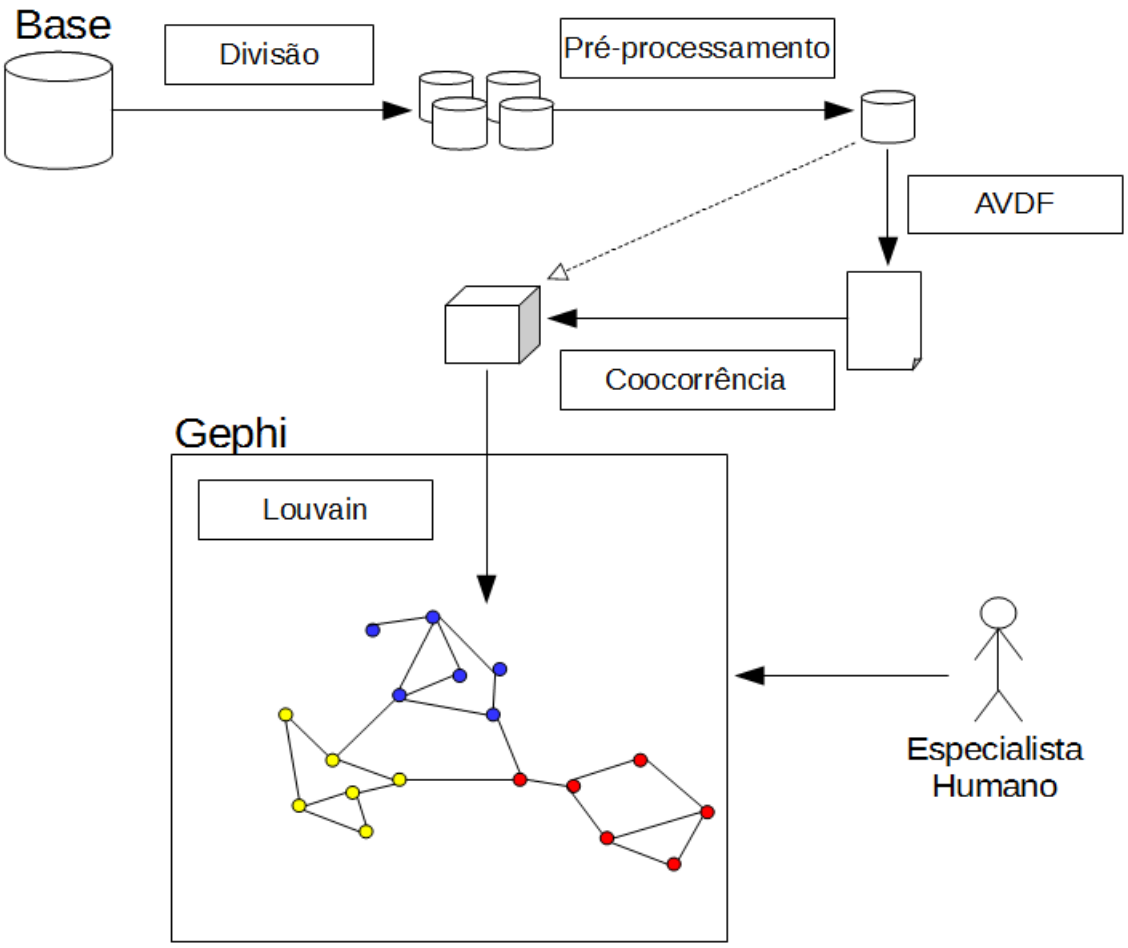


Figura 7 – Metodologia utilizada.

O Twitter foi o microblog escolhido para este trabalho por ser uma mídia social aberta e de conteúdo acessível. Para a aquisição de seus dados por um sistema *crawler*, é necessário o uso de uma API (Application Programming Interface) oficial do Twitter. Está API é disponibilizada gratuitamente para desenvolvedores que utilizam a mídia social, seja pela aquisição de seus dados ou publicação de novos conteúdos. A API do Twitter é dividida em três principais ferramentas: REST API que destina-se à manipulação dos dados dos usuários e as conexões entre eles, bem como o envio de mensagens. A Search API que é usada por produtos que permitem que um usuário consulte o conteúdo do Twitter,

pesquisando os tweets com palavras chaves específicas. E por último, a Streaming API que se destina à troca intensiva de dados em tempo real, sendo a mais adequada para a mineração de texto.

O sistema *crawler* deste trabalho utilizou a API Twitter4J[1] para aquisição dos dados. Por meio da entrada de palavras-chaves, a API recupera *tweets* de acordo com palavras-chaves especificadas. Os dados são armazenados em arquivos de extensão CSV (*Comma-Separated Values*) apresentando o conteúdo e a data de públicação dos tweets recuperados.

A base de dados formada pelo sistema *crawler* e utilizada neste trabalho é composta por 2.100 *tweets* referentes aos episódios do seriado inglês *Game of Thrones* (GoT) coletados entre os dias 24 de Maio a 01 de Junho de 2015.

O modelo proposto anteriormente foi realizado uma vez para cada divisão da base coletada. O primeiro experimento envolve toda a base (completo); os demais experimentos foram aplicados na base divida por dia de publicação do *tweet*. O lançamento de cada episódio do seriado acontecia aos domingos, portanto os dias que teriam maior número de *tweets* envolvendo o assunto seriam o próprio domingo e o dia seguinte. A partir desta lógica, foram realizados mais 4 experimentos de *tweets* dos dias 24 e 30 (domingos), 25 e 31 (segundas-feiras) de Maio.

Para a construção de uma formatação única, todos os caracteres são transformados em minúsculos. Através do uso de expressões regulares, os primeiros dados a serem retirados são as URLs e links. Tratando-se de extração de tópicos, esses links não representam um possível termo analisável. Como os *tweets* apresentam no máximo 140 caracteres, o uso de encurtadores de links são presentes, acarretando em links aleatórios sem possíveis informações acerca de seu conteúdo.

A maioria dos *tweets* apresentam uma linguagem mais coloquial, com a presença de gírias, uso de caracteres não-alfanuméricos e espaçamentos desnecessários. A partir deste ponto, o pré-processamento continua a partir das retiradas desses caracteres e espaçamentos. A retirada de espaçamentos desnecessários ajuda na etapa seguinte do pré-processamento, denominado de tokenização. Este processo transforma cada palavra de cada *tweet* em termos ou *tokens*. O processo de lematização não será adotado neste trabalho pois na próxima etapa, o uso do método AVDF pode eliminar variações de palavras não muito utilizadas.

Este trabalho utilizou a ferramenta *MathWorks* para realizar as técnicas de pré-processamento da base de dados. Verificando o padrão da base, esta apresenta vários idiomas em seu conteúdo, predominando o idioma inglês, logo em seguida o espanhol e o português. Por meio deste reconhecimento, foi utilizado listas de *stopwords* destes 3 idio-

[1] www.twitter4j.org

mas. Foram retiradas palavras como artigos, pronomes e preposições pois são considerados ruídos em relação a formação de tópicos.

Após o plotamento da reta AVDF, é adotado o retorno de 300 termos mais próximos da reta utilizando o distância euclidiana como medida (Equação 6.1). Este número representa uma boa quantidade de termos possíveis para representação de tópicos, sendo que nem todos eles serão utilizados.

Para a verificação da coocorrência, é adotado como corte para a lista de adjacência arestas com pesos maiores ou iguais a 2, devido ao elevado número de arestas com peso igual a 1. Esse corte justifica a eliminação de um grafo muito denso, sendo incapaz de ser analisado devido ao grau elevado dos nó.

Para o cálculo da modularidade e determinar as comunidades, o presente trabalho utilizou o software Gephi[2]. Este software código aberto (*open source*) possui módulos capazes de importar, visualizar, explorar, filtrar, manipular e exportar qualquer tipo de rede [34]. O método de Louvain também esta implementado no software Gephi.

Após a formação do grafo e a distinção em comunidades, a avaliação dos resultados deste trabalho será baseada em um especialista humano. A justificativa desta escolha se dá através das situações como:

- É necessário criatividade para avaliação.
- A velocidade da avaliação não precisa ser rápida.
- A análise deve ser baseada em um conhecimento pré-estabelecido a cerca do conteúdo.
- Pode ser necessário interação com o mundo exterior, outras fontes de conhecimento.

[2] Disponível em: www.gephi.github.io

8 RESULTADOS E DISCUSSÃO

Na Tabela 2, verifica-se o total inicial e final de termos e vocabulário de cada experimento realizado. Independente do experimento realizado, a porcentagem de redução antes e após o pré-processamento são próximas para todos os casos. Isso significa que dado um subconjunto de *tweets* de um conjunto total, este apresentará o mesmo padrão de estrutura de palavras que a base inteira, ou seja, o uso de *stopwords* é necessário e constante em toda a base. Pode-se perceber que a redução do vocabulário segue o mesmo padrão que o total de termos.

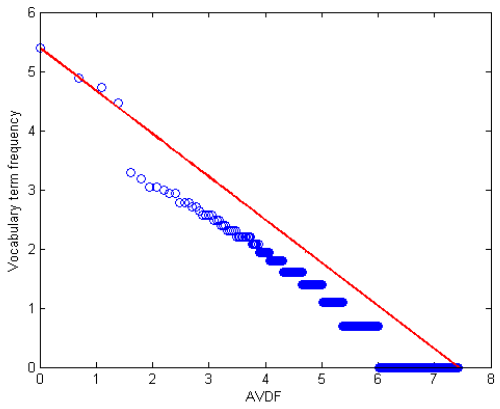
Tabela 2 – Características da base de cada experimento.

Nº	Base	Total de termos Inicial - Final	Vocabulário Inicial - Final	Redução dos termos	Redução do Vocabulário
1	GoT completo	20896 - 15034	5865 - 5523	28%	5,8%
2	GoT 24/05	4815 - 3450	1903 - 1672	28,3%	12,1%
3	GoT 25/05	4888 - 3552	2061 - 1847	27,3%	10,3%
4	GoT 30/05	4882 - 3464	2069 - 1857	29,0%	10,2%
5	GoT 31/05	6311 - 4568	1990 - 1770	27,6%	11%

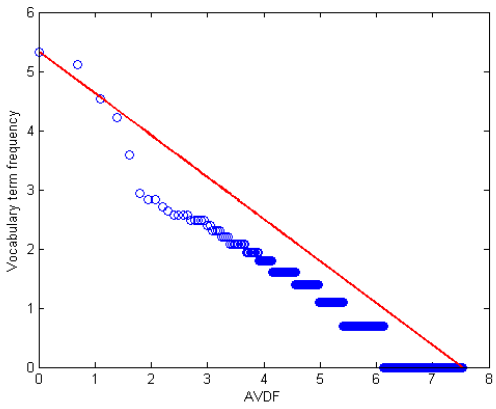
Um detalhe importante a ser considerado sobre os experimentos a cima é em relação a base 5. As bases 2, 3 e 4 apresentam quantidades de total de termos semelhantes assim como seus respectivos vocabulários. Porém na base 5, por apresentar uma quantidade maior de total de termos, seu vocabulário está próximo das bases anteriores. Isso significa que muitos *tweets* da base 5 apresentam conteúdos semelhantes. Considerando centenas ou milhares de *tweets* em uma base de dados, a tarefa de publicação de conteúdo iguais é dado por *bots*, sejam eles fraudulentos ou não, concluindo que a base 5 apresenta grande quantidade de textos *spams* causado por esses agentes.

Aplicando o método AVDF nestas bases, temos como resultados a Figura 8, mostrando o comportamento para cada experimento realizado. Com exceção da base 5, pode-se perceber que nos demais gráficos, a reta AVDF estipulada está distante das frequências reais de cada termo. Este fato é causado pela retirada das *stopwords* e também pelo fator sintático e semântico da base. Ao contrário de um texto tradicional em que todas as frases referenciam uma ideia em comum, a base de dados utilizada neste trabalho é formada por *tweets* de diferentes origens concatenados formando uma estrutura única. Como a base 5 apresenta um conteúdo mais padronizado por causa da ação de *bots* em produzirem textos *spams*, seus termos e suas respectivas frequências ficam próximo ao esperado pela reta AVDF. A probabilidade de termos produzido por *spams* serem retornados como tópicos é alta. Este tipo de comportamento deve ser analisado pelo especialista humano nas etapas seguintes.

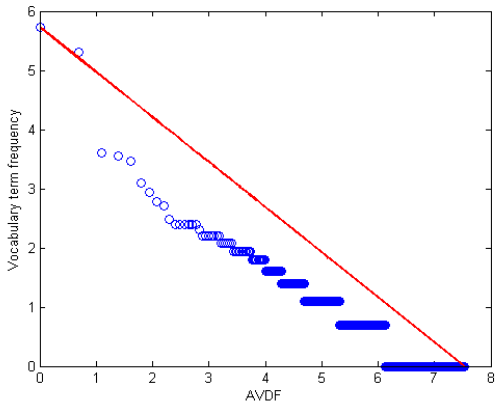
Retorna-se os 300 termos mais próximos de cada bases de dados da reta AVDF utilizando a distância euclidiana como comparativo. Os termos retirados geralmente são os primeiros e os que se encontram na porção central do rank.



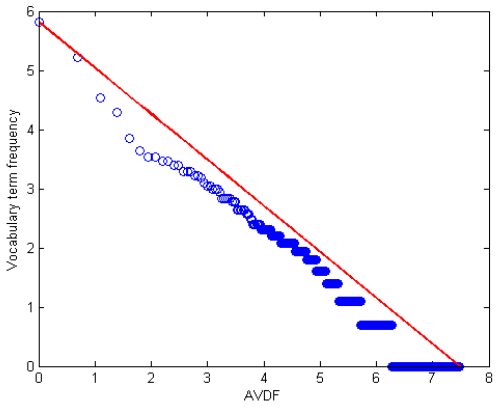
(a) GoT 24/05



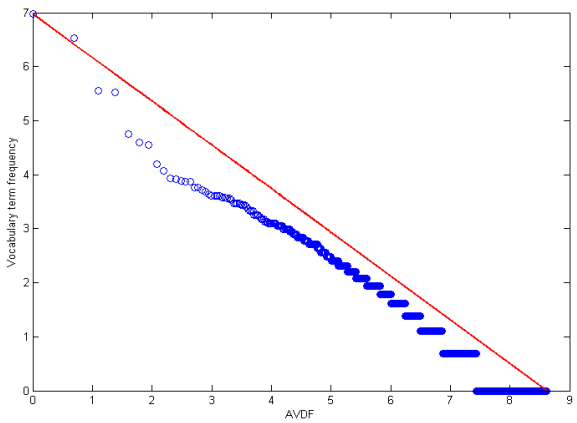
(b) GoT 25/05



(c) GoT 30/05



(d) GoT 31/05



(e) GoT completo

Figura 8 – Aplicação do método AVDF nos experimentos

8.1 Grafos

Utilizando a ferramenta Gephi, para cada base de dados foi gerado um grafo em relação a coocorrência dos 300 termos mais próximos da reta da AVDF. Começando pela Figura 9, é ilustrado o grafo final da base 1 da Tabela 2. A coocorrência dos termos gerou uma lista de adjacência de 242 nós e 378 arestas com peso maior que 2. Após a aplicação do método de Louvain, tem-se a formação de 18 comunidades resultantes. Nesta imagem podemos perceber comunidade que se encontram afastadas e não possuem ligação com grafo central. Normalmente essas comunidades são formadas por *tweets* que não fazem sentido com o contexto da base, geralmente formado por *bots* na propagação de *spams*.

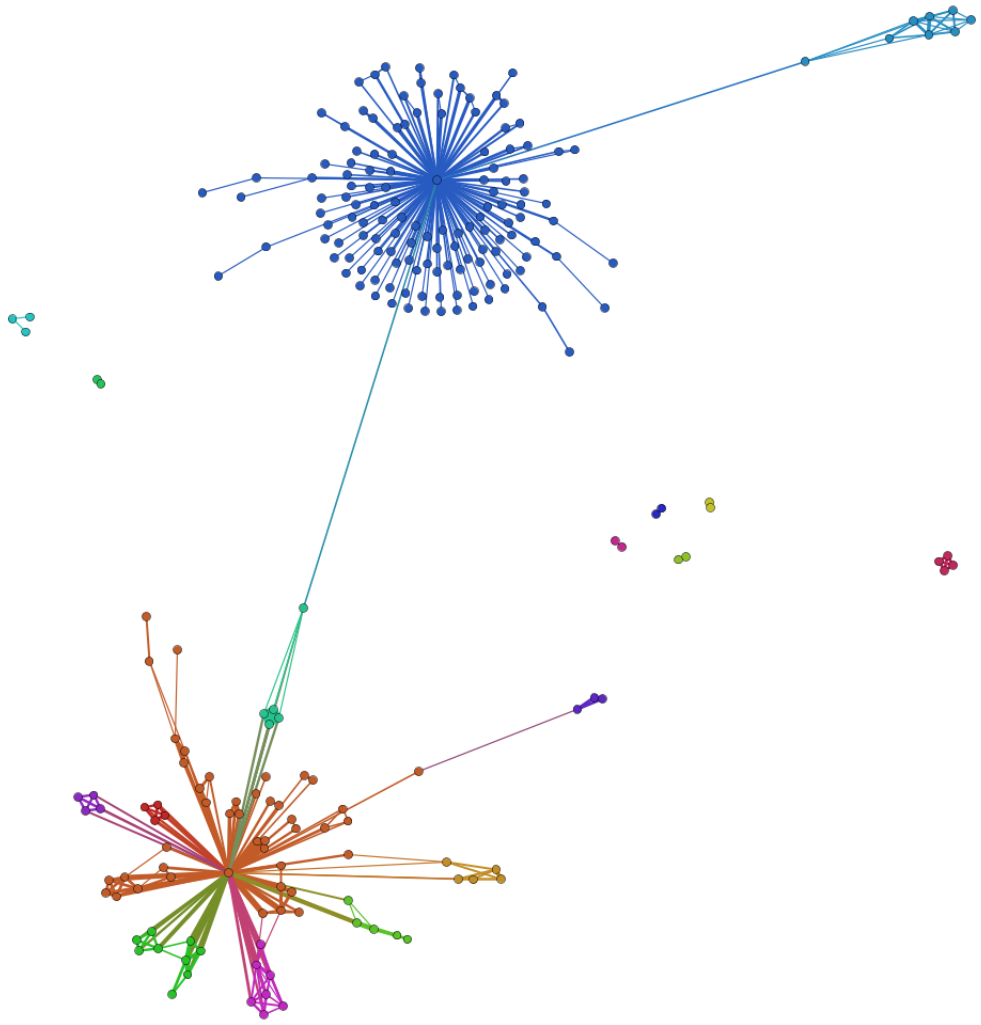


Figura 9 – Grafo GoT completo.

Começando a analise dos resultados pela maior comunidade formada, ilustrada na Figura 10, esta apresenta-se termos referenciando diretamente o conteúdo do seriado inglês *Game of Thrones*, ou seja, palavras específicas referente a lugares, pessoas e eventos que aconteceram no seriado. Neste tipo de análise são necessários especialistas humanos

que possuem conhecimento acerca do tema. É possível verificar que o termo central desta comunidade é a sigla inicial do seriado (got).

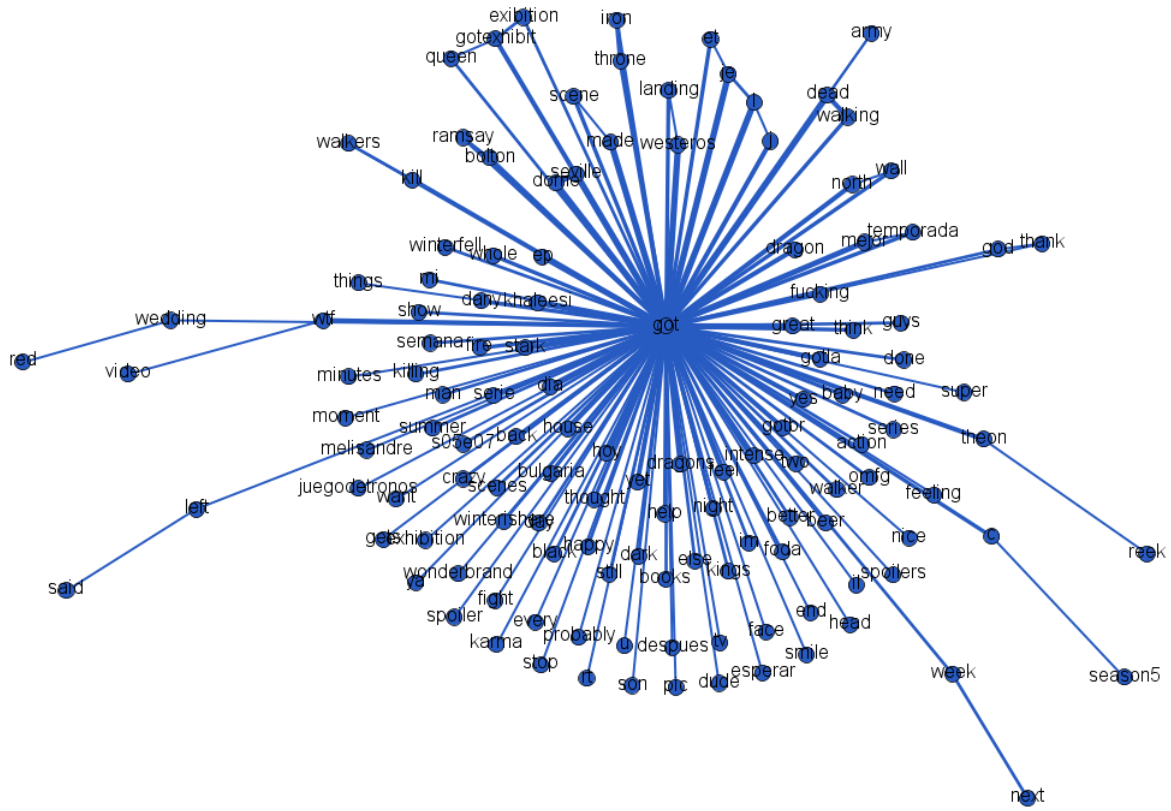


Figura 10 – Comunidade 1 GoT completo.

Grande parte dos termos desta comunidade apresentam apenas uma ligação, sendo diretamente relacionada com o centro do grafo. Ao aplicar novamente o método de Louvain apenas para os nós desta comunidade, obteve uma formação de novas comunidades, porém pequenas. Como a maioria dos nós apresentam apenas uma ligação, este não há escolha em não fazer parte da mesma comunidade que sua adjacência. As novas comunidade que surgiram estão relacionadas a fatos que aconteceram na série. Como exemplo, os dois nós "red"e "wedding"faz referência a um trágico evento que aconteceu no seriado. Os nós "iron"e "throne"se relacionam ao contexto do seriado, assim como "reek"e "theon", ambos se referem ao mesmo personagem. Os nós "next"e "week"pode se referir ao próximo episódio da temporada.

Em relação as outras comunidade formadas, a comunidade da imagem a cima se encontra distante das demais. As outras comunidades reconhecidas não possuem vínculo com as características do seriado, como a da Figura 11, a segunda maior comunidade.

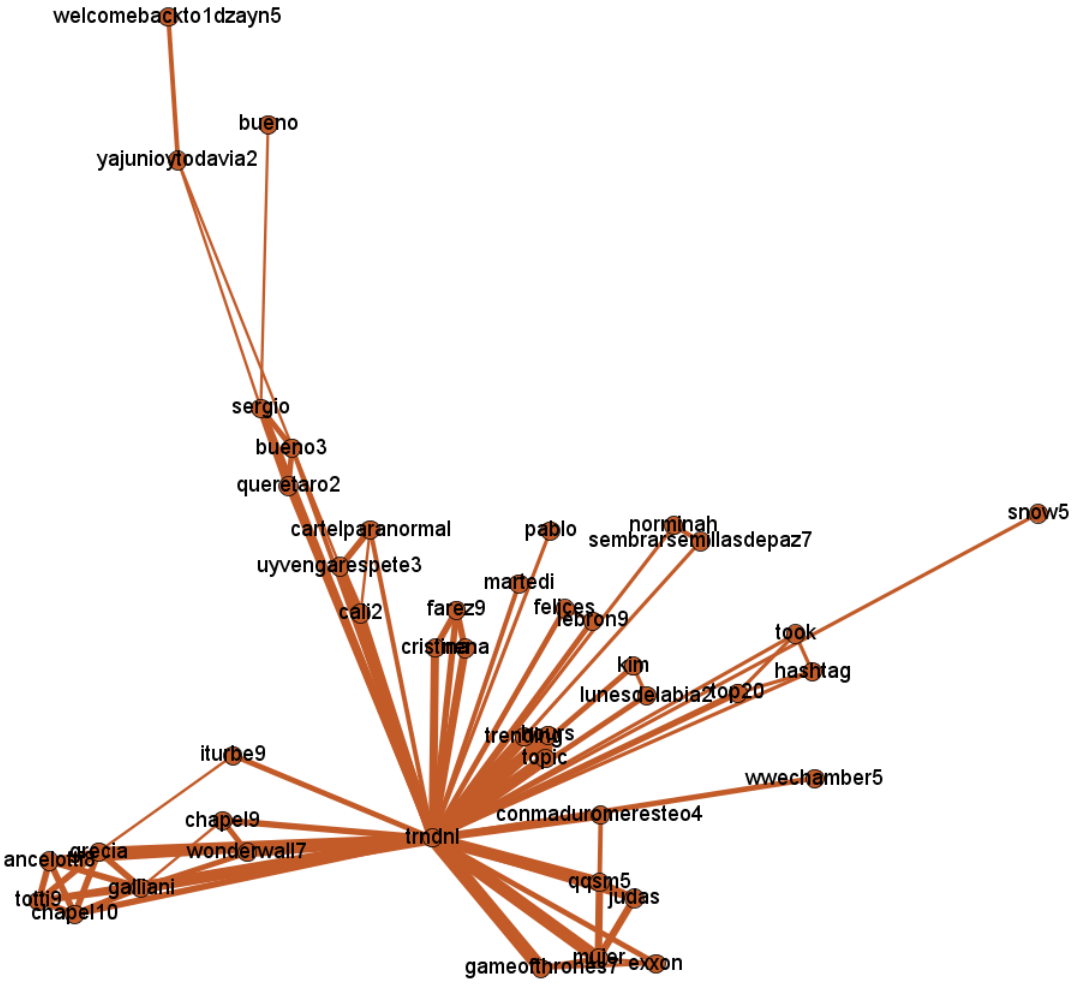


Figura 11 – Comunidade 2 GoT completo.

O termo central "trndnl"refere-se a "Trendinalia", um sistema que analisa a classificação de *Trend Topics* (delimitado pelo uso do "#") no Twitter. Esse sistema efetua publicações automáticas com as *hashtags* mais utilizadas no momento. Se muitos usuários utilizam, por exemplo, a *hashtag* "#gameofthrones", essa tag seria captada pelo sistema Trendinalia por ser muito utilizada, que automaticamente publicaria um tweet com esse conteúdo. Por apresentar apenas um termo característico, esse *tweet* é reconhecido pelo sistema *crawler*, inserindo-o na base de dados. A Tabela 3 refere-se a *tweets* relacionados a este problema. Pode-se considerar que os termos destes *tweets* são considerados ruídos em relação ao tema abordado e que devem ser descartados pelo especialista humano.

Utilizando a base 2 da Tabela 2, obteve-se o grafo da Figura 12, formado por 100 nós e 173 arestas, com a distinção de 11 comunidades. Por apresentar uma quantidade menor de dados em relação à base 1 porém retornando a mesma quantidade de termos após a aplicação do método AVDF, através de um especialista humano é possível obter dados mais específico acerca do conteúdo.

Tabela 3 – Exeplos de *tweets* ruidosos.

Nº	Conteúdo
1	6. #GameOfThrones7. Leon Larregui8. Santos y Queretaro9. Chivas10. John Nash 2015/5/25 04:14 CDT #trndnl http://t.co/IN7801UqsL
2	6. Leopoldo Lopez7. Ceballos8. Pastor Maldonado9. Cersei10. Exxon 2015/5/25 04:44 VET #trndnl http://t.co/TZZWvFfY1p
3	1. 1. #charliecharliechallenge2. #MeCaesMalSi3. #GameOfThrones4. #camrenfeels5.#30MVamosTodos http://t.co/TZZWvFfY1p

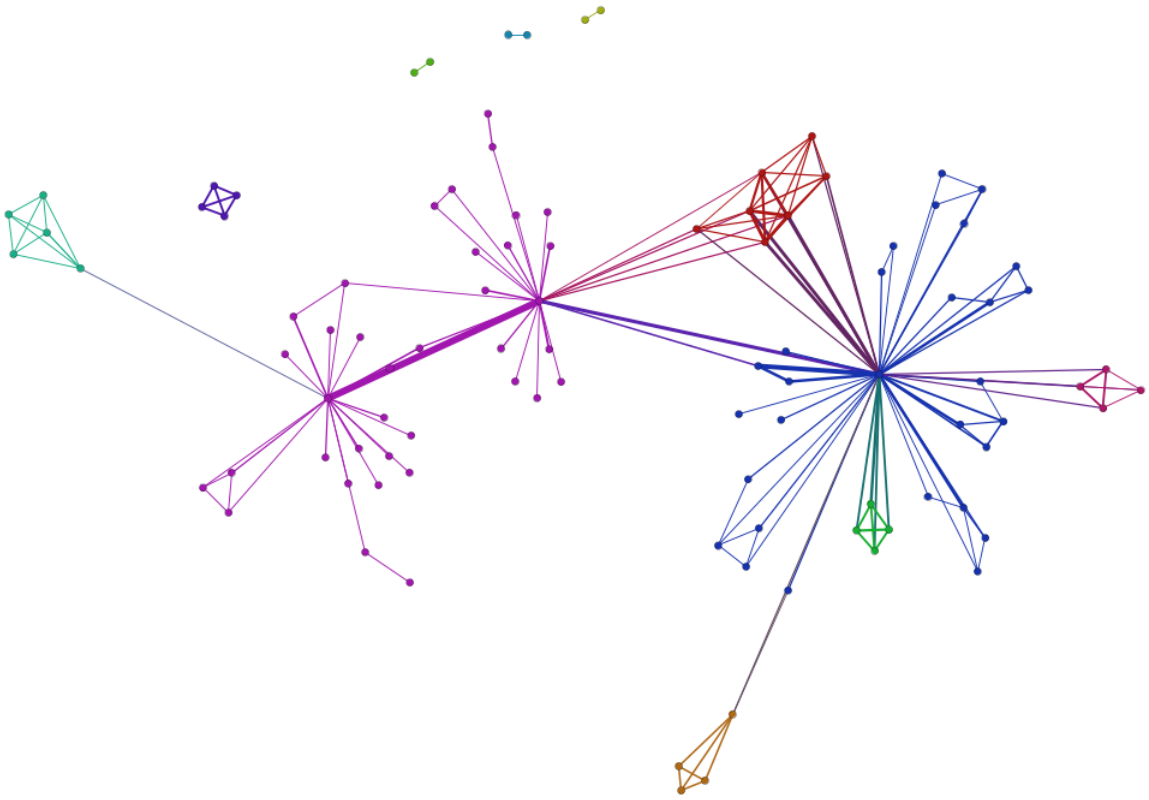
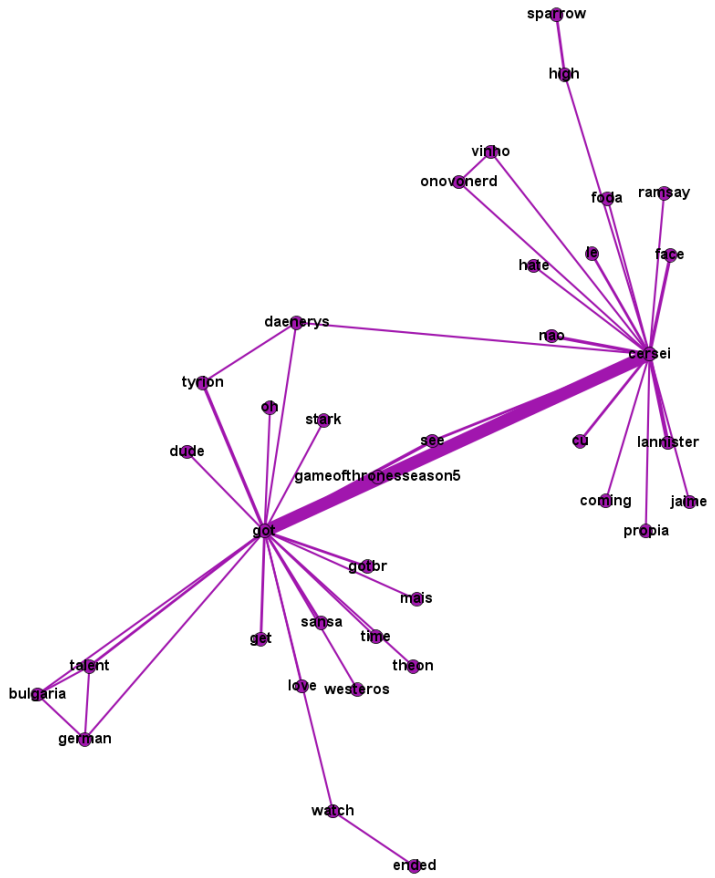
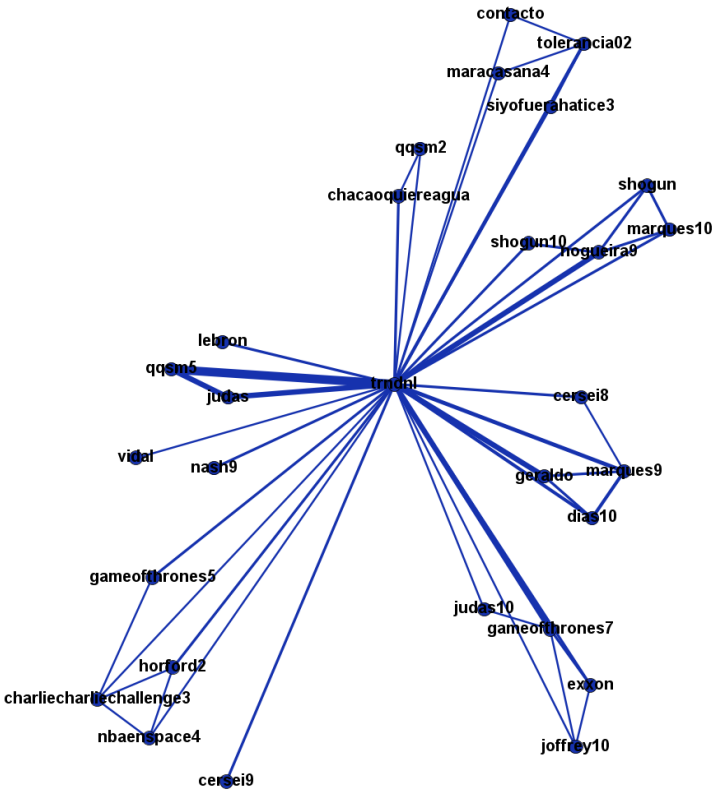


Figura 12 – Grafo GoT 24/05

Não muito diferente do grafo anterior, este grafo apresenta duas grandes comunidades: A Figura 13a apresenta dois nós centros: "got"e "cersei", possuindo uma forte ligação entre eles com "gameofthronesseason5". Há presença de nós com personagens importantes da série como "jaime", "tyrion", "daenerys", "sansa", "ramsay"e "theon". Na Figura 13b, a situação se assemelha com a Figura 11, sendo o "trndnl"como nó central e seus termos *spams* ao redor.



(a) Comunidade 1



(b) Comunidade 2

Figura 13 – Comunidades GoT 24/05.

Na base 3 da Tabela 2 referente aos *tweets* do dia 25 de Maio, tem-se um total de 75 vertices e 142 arestas, com a distinção de 10 comunidades. Na Figura 14 tem-se o grafo referente a esta base de dados. É possível perceber a divisão de 2 grupos. O primeiro grupo é representado pela comunidade (Figura 15) que se relaciona ao contexto da base. É possível extrair informações, como por exemplo, o termo "s05e07"se refere ao episódio atual que se está comentando, além de outros personagens como John Snow referentes ao termo "jon"e "snow". As outras comunidades não se referem ao tema do seriado e são termos *spams*.

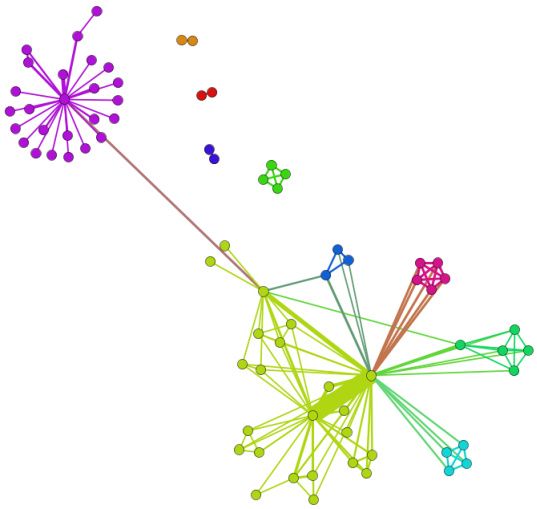


Figura 14 – Grafo GoT 25/05

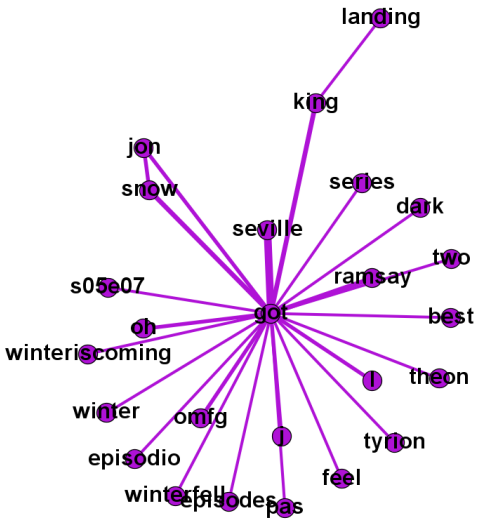


Figura 15 – Comuninadade GoT 25/05

Em relação a base 4, referente ao dia 30 de Maio, na Figura 8c, pode-se perceber que a grande maioria do termos se encontram muito distantes da reta AVDF. Ao gerar a coocorrência do 300 termos selecionados, apenas 26 deles foram determinados, com

37 arestas. A Figura 16 ilustra o grafo formado por esta base, com a distinção de 5 comunidades.

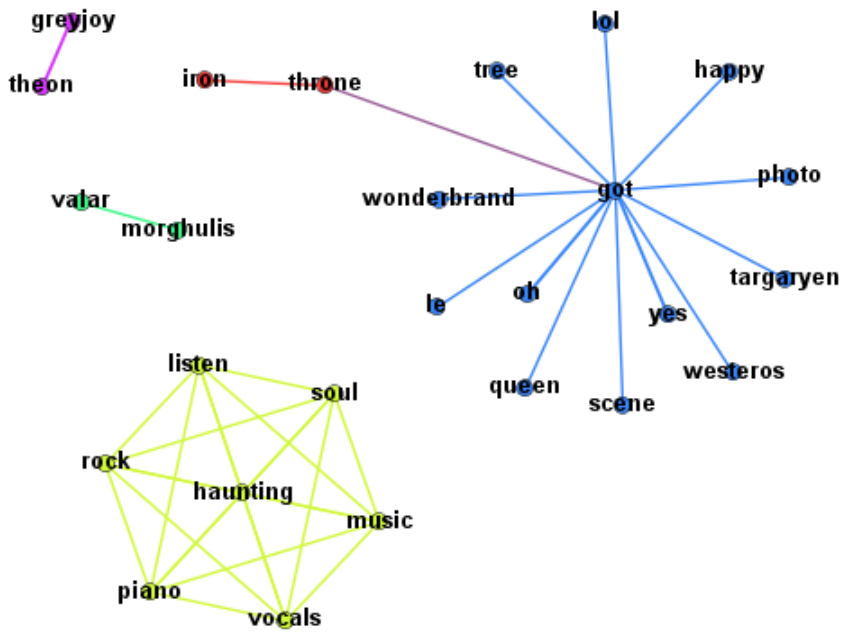


Figura 16 – Grafo GoT 30/05

Por último e não menos importante, tem-se o gráfico da base 5, referente ao dia 31 de Maio. Considerando ao que foi comentando anteriormente em relação a base 5, a Figura 17 ilustra o grafo composto por 192 vértices e 461 arestas, com a distinção de 9 comunidades.

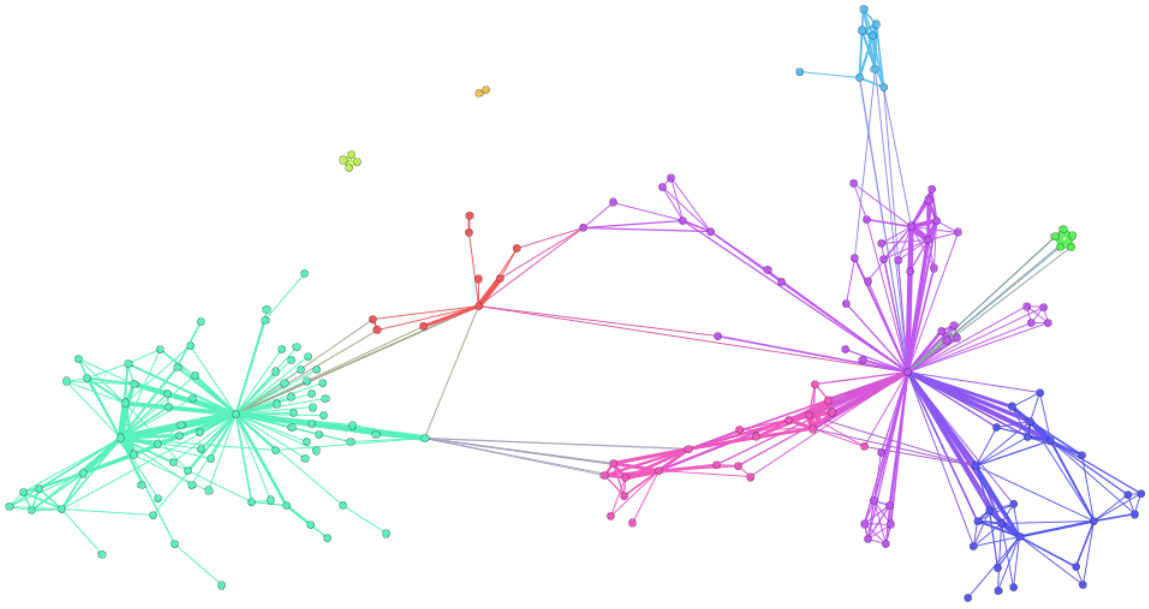


Figura 17 – Grafo GoT 31/05

A Figura 18 ilustra a comunidade referente ao contexto do seriado inglês. Com exceção desta, as demais comunidades são constituídas por termos *spams*.

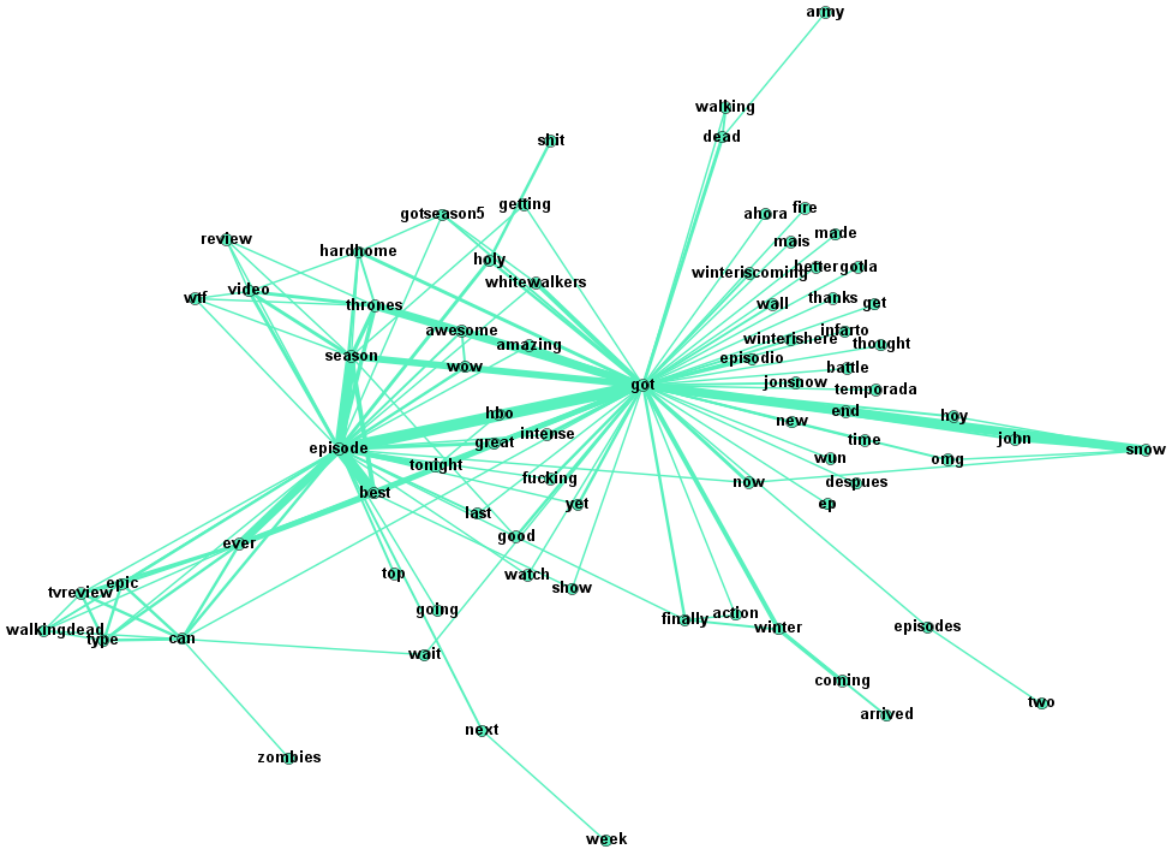


Figura 18 – Comuninadade GoT 31/05

Levando em consideração que todos os grafos dos experimentos a cima possuíram grande quantidades de termos *spams* em seu conteúdo, ainda foi possível analisar as comunidades restantes e verificar termos como possíveis tópicos em relação ao contexto do base de dados. O uso de um especialista humano foi capaz de associar os termos com base em seus conhecimentos e/ou através de fontes externas, tomando melhores decisões. Neste trabalho, o agente humano foi capaz distinguir que muito termos não possuia relação nenhuma com o assunto, devido aos *tweets* de *bots*. Talvez uma máquina não seria capaz de distinguir rapidamente. Seria necessário um maior conhecimento pré-estabelecido com um tempo maior de processamento.

Em relação a grande quantidade de *tweets spams*, uma possível solução é verificar quais dados serão analisados. Encontrar um padrão nesse tipo de texto e retirá-los, proporcionará um base de dados mais rica em relação ao contexto, ou seja, seria possível extrair mais termos/tópicos específicos.

9 CONCLUSÃO

Este presente trabalho tem como objetivo propor um modelo capaz de extrair tópicos em texto de mídias sociais. O problema relatado trata-se das diferenças entre textos de RSOs e textos tradicionais como artigos, reportagens, entre outros, por não apresentarem as mesma caracteristicas em suas estruturas. É necessário uma adaptação das técnicas existentes ou criação de modelos capazes de analisarem textos de RSOs, levando em consideração características como abreviações, gírias e erros gramaticais.

A mídia social utilizada neste trabalho foi o microblog Twitter. Utilizando a API Twitter4J disponibilizada pelo desenvolvedores, foi possível coletar *tweets* de acordo com palavras-chaves escolhidas. A base de dados utilizada neste trabalho possui cerca de 2.100 *tweets* em relação ao tema *Game of Thrones*, seriado inglês, coletados entre o dia 24 de Maio à 1 de Junho de 2015. Esta dividida e utilizada em 5 tipos de experimentos. O pré-processamento leva em consideração a transformação em caracteres minúsculos, retirada de links, caracteres especiais e espaçamentos desnecessários. No fim desta etapa, são retirados as *stopwords* e depois, passa por um processo de tokenização.

Após o pré-processamento, os termos são avaliados por suas frequências e pelo método AVDF. São selecionados o 300 termos mais próximo da reta proposta. Posteriormente, estes termos passam por uma verificação de suas coocorrências em pares. As listas de adjacências resultantes desses processos são responsáveis pelas formações dos grafos a serem analisados.

Por meio do Gephi, aplicou-se o método de Louvain em cada grafo resultante dos experimentos a fim de delimitar comunidades. Após a análise dos resultado, foi possível perceber que o modelo proposto atingiu o objetivo em delimitar termos e tópicos específicos acerca do contexto da base. Através de um especialista humano para analisar os resultados, pode-se perceber a presença de demasiados termos *spams* causados pela presença de publicações de *bots* no Twitter. Esses termos não apresentavam nenhuma relação com o tema base, sendo considerados como ruídos.

Como trabalho futuros, o objetivo seria encontra padrões nestes tipos de *tweets spams* afim de identificá-los e retirá-los da base de dados, com isso os resultados seriam mais precisos com termos mais específicos. A utilização de um classificador de usuários no Twitter como humano ou *bot* poderia ajudar na identificação destes padrões. Seriam retirados todos os *tweets* do usuário classificado como *bot*.

REFERÊNCIAS

[1] IGAWA, R. et al. Adaptive distribution of vocabulary frequencies: A novel estimation suitable for social media corpus. In: *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*. [S.l.: s.n.], 2014. p. 282–287.

[2] NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical review E*, APS, v. 69, n. 2, p. 026113, 2004.

[3] BLONDEL, V. D. et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, v. 2008, n. 10, p. P10008+, jul. 2008. ISSN 1742-5468. Disponível em: <http://dx.doi.org/10.1088/1742-5468/2008/10/p10008>.

[4] VERMA, V.; RANJAN, M.; MISHRA, P. Text mining and information professionals: Role, issues and challenges. In: *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on*. [S.l.: s.n.], 2015. p. 133–137.

[5] CHITRA, K.; SUBASHINI, B. Data mining techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, Citeseer, v. 3, n. 8, p. 219–226, 2013.

[6] SOELISTIO, Y. E.; SURENDRA, M. R. S. Simple text mining for sentiment analysis of political figure using naïve bayes classifier method. In: *Proc. the 7th International Conference on Information and Communication Technology and Systems*. [S.l.: s.n.], 2013.

[7] CHOI, D. et al. Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, Elsevier, v. 38, p. 16–21, 2014.

[8] CAMILO, J. C. d. S. C. O. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. [S.l.], 2009.

[9] HAN, J. *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 1558609016.

[10] AKILAN, A. Text mining: Challenges and future directions. In: IEEE. *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*. [S.l.], 2015. p. 1679–1684.

[11] HE, J. Advances in data mining: History and future. In: IEEE. *Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on*. [S.l.], 2009. v. 1, p. 634–636.

[12] FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. [S.l.]: Cambridge University Press, 2007.

[13] ZAPPAVIGNA, M. Ambient affiliation: A linguistic perspective on twitter. *New media & society*, SAGE Publications, v. 13, n. 5, p. 788–806, 2011.

[14] IGAWA, R. A. et al. Recognition of compromised accounts on twitter. 2015.

[15] TSAI, F. S. A tag-topic model for blog mining. *Expert Systems with Applications*, v. 38, n. 5, p. 5330 – 5335, 2011. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417410011693>.

[16] ZENG, J. et al. Topics modeling based on selective zipf distribution. *Expert Systems with Applications*, v. 39, n. 7, p. 6541 – 6546, 2012. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417411017222>.

[17] LI, H. et al. Mining user interest in microblogs with a user-topic model. *Communications, China*, v. 11, n. 8, p. 131–144, Aug 2014. ISSN 1673-5447.

[18] CHEN, Y. et al. Emerging topic detection for organizations from microblogs. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2013. (SIGIR '13), p. 43–52. ISBN 978-1-4503-2034-4. Disponível em: <http://doi.acm.org/10.1145/2484028.2484057>.

[19] TAN, S. et al. Interpreting the public sentiment variations on twitter. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 26, n. 5, p. 1158–1170, 2014.

[20] LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, Taylor & Francis, v. 25, n. 2-3, p. 259–284, 1998.

[21] HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, Springer, v. 42, n. 1-2, p. 177–196, 2001.

[22] CHEN, Z.; LIU, B. Mining topics in documents: Standing on the shoulders of big data. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2014. (KDD '14), p. 1116–1125. ISBN 978-1-4503-2956-9. Disponível em: <http://doi.acm.org/10.1145/2623330.2623622>.

[23] HUANG, S. et al. Topic detection from microblog based on text clustering and topic model analysis. In: IEEE. *Services Computing Conference (APSCC), 2014 Asia-Pacific*. [S.l.], 2014. p. 88–92.

[24] SOBKOWICZ, P.; KASCHESKY, M.; BOUCHARD, G. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, v. 29, n. 4, p. 470 – 479, 2012. ISSN 0740-624X. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011). Disponível em: <http://www.sciencedirect.com/science/article/pii/S0740624X12000901>.

[25] PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: *LREC*. [S.l.: s.n.], 2010. v. 10, p. 1320–1326.

[26] CHU, Z. et al. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, IEEE, v. 9, n. 6, p. 811–824, 2012.

[27] HUANG, S. et al. Topic detection from microblog based on text clustering and topic model analysis. In: *Services Computing Conference (APSCC), 2014 Asia-Pacific*. [S.l.: s.n.], 2014. p. 88–92.

[28] BLEI, D. M. Probabilistic topic models. *Communications of the ACM*, ACM, v. 55, n. 4, p. 77–84, 2012.

[29] MORAIS E. A. M. AMBRÓSIO, A. P. L. *Mineração de Textos*. [S.l.], 2007.

[30] TWITTER. *USO DO TWITTER / FATOS SOBRE A EMPRESA*. 2015. Disponível em: <https://about.twitter.com/pt/company>.

[31] PAPADOPOULOS, S. et al. Community detection in social media. *Data Mining and Knowledge Discovery*, Springer US, v. 24, n. 3, p. 515–554, 2012. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1007/s10618-011-0224-z>.

[32] TANG, L.; WANG, X.; LIU, H. Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, Springer US, v. 25, n. 1, p. 1–33, 2012. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1007/s10618-011-0231-0>.

[33] CLAUSET, A.; NEWMAN, M. E.; MOORE, C. Finding community structure in very large networks. *Physical review E*, APS, v. 70, n. 6, p. 066111, 2004.

[34] BASTIAN, M.; HEYMANN, S.; JACOMY, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.