UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

PROPOSTA DO TRABALHO DE CONCLUSÃO DE CURSO

MAC0499

Autora:
Fernanda DE CAMARGO MAGANO

Supervisora:
Kelly Rosa BRAGHETTO

A proposta do TCC visa a descrever e contextualizar como será o trabalho de conclusão de curso e qual será a direção tomada

April 7, 2016

Contents

1 Contextualização:

As mídias sociais representam uma forma de comunicação muito utilizada nos dias atuais. Cada vez mais pessoas, de diferentes faixa etárias e opiniões, têm acesso a essas mídias e apresentam interesse em usá-las. Para citar algumas: Twitter, Facebook, Whatsapp, LinkedIn, entre outras, ocupam um espaço importante na sociedade.

De acordo com as informações do Pew Research Center [1], 74% dos adultos que usavam internet em janeiro de 2014 acessavam sites de redes sociais. Além disso, na última década houve um grande salto no número de jovens de 18-29 anos que utilizam redes sociais. Em setembro de 2013, 90% dos jovens nessa faixa etária já utilizavam redes sociais. Eles eram apenas 9% em fevereiro de 2005.

Outras estatísticas mais recentes, coletadas do Smartinghts [2], comprovam a importância das redes sociais na atualidade: em janeiro de 2016, a população mundial contava com 7,395 bilhões de pessoas, sendo 3,149bi usuários da internet e 2,307 bi usuários de mídias sociais.

Dessa forma, quase um terço da população mundial participa de redes sociais, expondo e compartilhando ideias e opiniões, fazendo comentários positivos ou negativos sobre as mais diversas experiências de suas vidas.

Diversos tipos de informações relevantes podem ser automaticamente extraídas a partir do enorme volume de dados gerados diariamente nas redes sociais, por meio de técnicas como as de mineração de dados e opiniões.

2 Caracterização do Problema

2.1 Mineração de dados e de opiniões

A mineração de dados, de acordo com o livro *Data Mining and Analysis - Fundamental Concepts and Algorithms* [3] trata-se de extrair conhecimento sobre um grande conjunto de dados, descobrindo padrões e modelo preditivos.

Na área de mineração, uma série de conhecimentos de diferentes campos são envolvidos. Para citar algumas áreas: recuperação de informação, para obtenção de dados; estatística, inteligência artificial e aprendizado de máquina, para a classificação dos dados (como, por exemplo, separar aspectos positivos e negativos sobre determinado assunto).

Uma subárea da *mineração de dados* é a mineração de opinião ou análise de sentimentos (*sentiment analysis*). Trata-se de extrair opiniões das pessoas através dos textos e, portanto, a extração de dados subjetivos, usando técnicas de processamento de linguagem natural, análise textual e linguística [4, 5, 6].

Na mineração de opiniões costuma-se atribuir uma certa polaridade à informação, isto é, classifica-se como positiva, negativa ou neutra. Por exemplo, se o dado analisado é uma resenha de um filme, com base em palavras-chave tenta-se concluir se o filme é avaliado como bom ou não.

2.2 Dificuldades associadas à mineração de opiniões

As redes sociais são fontes de dados heterogêneas. Cada rede social tem as suas características quanto à forma de interação. Por exemplo, algumas

restringem o tamanho do texto dos *posts*; outras permitem a mistura de emoticons e símbolos nos textos.

A qualidade dos dados também é heterogênea, já que exitem vários tipos de pessoas e organizações que utilizam mídias sociais e, assim, nem sempre a fonte de informação é confiável, ou seja, nem tudo é verídico. Coletar opiniões trata-se de obter dados subjetivos, então depende de uma série de fatores, como o meio em que o indivíduo está inserido, suas crenças pessoais e experiências de vida.

Trabalhar com grandes volumes de dados gerados continuamente (*Big Data*) é complexo das perspectivas da coleta, armazenamento e análise. O volume de dados gerado por uma rede social em um único ano atinge facilmente Terabytes ou até mesmo Petabytes.

Na análise dos dados para a mineração de opiniões, o processamento de linguagem natural é particularmente bem trabalhoso, já que posts possuem abreviações, gírias, sarcasmo (o que é difícil de ser detectado por uma máquina), emoticons, além do conteúdo poder estar em várias línguas. Assim, o tratamento dos dados tem que levar todos esses aspectos em consideração.

3 Proposta

Este projeto tem por objetivo estudar como diferentes tipos de técnicas de mineração de dados podem ser combinadas para a descoberta de conhecimento a partir do cruzamento de informações coletadas de mídias sociais e de bases de dados públicas.

Para isso, foi escolhido um domínio com bastante disponibilidade de dados abertos online: o domínio de cinema. Será usada a base de dados do Internet Movie Database (IMDB [7]) que disponibiliza informações sobre os filmes, e dados extraídos de *posts* na rede social Twitter.

O Twitter [5] é uma das redes sociais mais usadas na atualidade (possuía 320 milhões de usuários em janeiro de 2016, de acordo com statista [8]).

O Twitter caracteriza-se por apresentar mensagens com limite de 140 caracteres chamadas de *tweets*, os quais possuem *hashtags* que são palavras-chave para a informação contida no *tweet*.

Esses dados dos *tweets* são públicos e podem ser coletados gratuitamente de maneira *on-the-fly*, isto é, são fornecidos dados daquele momento de diferentes países do mundo. Se o desejado é fazer a análise desses dados por um período de tempo, cabe ao desenvolvedor o armazenamento desse volume de informações.

Com a identificação da polaridade (i.e., aspectos negativos e positivos) sobre os dados será possível estudar as preferências dos usuários com respeito aos filmes e fazer estudo por regiões, comparando duas cidades diferentes, uma do interior e uma metrópole, por exemplo, e estudar interesses e influências das mídias e indústria cinematográfica nesses locais.

O projeto utilizará bancos de dados orientados a grafos para armazenar e relacionar esses conhecimentos descobertos.

O domínio de cinema é apenas uma das possíveis áreas de aplicação para o uso das técnicas que serão estudadas neste trabalho. É possível citar outros exemplos de domínios: epidemias, mobilidade urbana, política e economia.

4 Ferramentas e métodos

Algumas tarefas a serem realizadas são: coleta dos dados, filtragem e limpeza de algumas informações e, então, será feita a mineração de opinião.

Para fazer coleta dos dados, processamento de linguagem natural e análise semântica, é necessário usar um conjunto de ferramentas porque cada uma delas é aplicada para uma parte do problema. Todas terão sua importância, mas somente juntas poderão alcançar os objetivos almejados.

Desta forma, é preciso criar um *workflow* combinando o uso das ferramentas mais apropriadas para lidar com cada etapa do problema. Com um *workflow*, tarefas independentes executadas por ferramentas diferentes podem ser realizadas paralelamente. Contudo, se uma tarefa for dependente da outra (por exemplo, para fazer análise sobre dados, precisa-se primeiro coletá-los), o *workflow* funciona como um *pipeline*, saída da primeira tarefa funciona como a entrada da segunda.

Uma das ferramentas necessárias é um analisador léxico e semântico que faça distinção entre substantivos, verbos, adjetivos, artigos, entre outros e possa fazer associações entre palavras sinônimas e antônimas.

Essas características são importantes, visto que artigos podem ser considerados *stopwords*, isto é, palavras bastante comuns num texto e que, portanto, não agregam muito significado. Adjetivos e advérbios podem ser usados na polaridade de textos por parte dos classificadores, uma vez que exprimem uma opinião referente a um assunto. Portanto, fazer tratamento léxico é parte fundamental para fazer posterior análise. Uma possível ferramenta com essas funcionalidades é o Wordnet [9].

Precisa-se também fazer a parte de mineração de opinião, classificando o teor e semântica associados a uma palavra. Atribuição de pontuação e análise de sentimentos inerentes ao texto podem ser feitos com ferramentas como Sentiwordnet [10].

Uma linguagem prática para fazer parsers e com bibliotecas úteis para processamento de linguagem natural é python, que pode ser usado com outras linguagem de *script*, como bash, por exemplo, para fazer filtros.

Por fim, ocorrerá a organização das informações em um banco de dados orientado a grafos que visa a mostrar os relacionamentos entre as pessoas que vão ao cinema, se uma pessoa segue a outra, têm interesses próximos e se conhecem/ou se admiram (no caso de um artista, por exemplo). Além de agrupar por região geográfica.

Para a organização dos dados será necessário um Sistema Gerenciador de Banco de Dados orientado a grafos, de modo a facilitar também o entendimento e melhor visualização das informações obtidas. Dentre os vários SGDBs exitentes com esse intuito, um que poderá ser utilizado é o Neo4j [11] que é altamente escalável, suportando centenas de milhares de transações ACID por segundo, além de possuir armazenamento e processamento nativo para grafos.

Outra forma de visualização de dados utilizada serão gráficos ilustrativos, com informações relevantes, como número de *tweets* por língua no contexto de uma *query* pesquisada.

5 Resultados esperados

Depois de realizar os métodos por intermédio de ferramentas cujas funcionalidades foram descritas na seção anterior, conseguiremos chegar a algumas conclusões.

A repercussão de um filme, seja pela expectiva, seja depois de assisti-lo, pode ser positiva ou negativa. Empresas de marketing podem se interessar por resultados nesse sentido. O gênero associado aos filmes de destaque e renome podem ajudar a reconhecer padrões e descobrir quais os assuntos que entretêm e interessam à população.

Outro uso dos dados é associar interesses por localização. Por exemplo, duas cidades, uma metrópole e uma cidade do interior, podem ter um público diferente no quesito cinema. Além do fato de que a quantidade de pessoas que vão ao cinema têm proporções diferentes (uma metrópole tem maior quantidade de cinemas e de pessoas).

Alguns *tweets* têm a opção de geolocalização ativas e, assim, pode-se saber o país de onde a pessoa utilizou a rede social. Contudo, isso já limita os *tweets* que serão utilizados na análise, porque só os que estão com essa opção habilitada serão interessantes.

O *workflow* que será montado pode servir como modelo para outras pessoas interessadas em fazer processos de automatização na área de mineração de dados e opiniões. Isso porque a série de etapas que serão seguidas são bem definidas e frequentemente utilizadas, além de fundamentais em áreas que trabalham com grande volume de dados.

6 Cronograma

As tarefas a serem realizadas seguem descritas a seguir:

1- Implementação/adaptação de programas para coleta automatizada dos dados sobre filmes que estão em cartaz no cinema e que estão sendo comentados nas redes sociais;

2- Implementação/adaptação de parser e limpeza para remoção de ruídos, como *stopwords*, e ficar com dados de maior qualidade;

3- Implementação/adaptação de programas para coleta automatizada dos dados do IMDB, como gênero, classificação indicativa (faixa etária mínima), atores que participaram;

4- Fazer a parte de *sentiment analysis*, extraindo informações com ferramentas adequadas para tratamento léxico e semântico dos dados;

5- Montar o banco de dados orientado a grafos com as informações já coletadas, tratadas e analisadas;

6- Desenvolvimento de formas de visualização, como grafos e gráficos;

7- Escrever a monografia;

8- Montar pôster e os slides para apresentação.

Meses	Tarefas							
	1	2	3	4	5	6	7	8
Março	x	x					x	
Abril	x	x					x	
Maio	x	x	x	x			x	
Junho	x	x	x	x			x	
Julho		x	x	x			x	
Agosto			x	x	x		x	
Setembro				x	x	x	x	
Outubro					x	x	x	x
Novembro						x	x	x

FIGURE 1: Cronograma

Bibliography

[1] Pew Research Center. http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/. [Accessed 2015-04-02].

[2] Smartinsights. http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [Accessed 2015-04-02].

[3] Mohammed J Zaki and Wagner Meira Jr. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[4] Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

[5] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Lrec*. Volume 10, 2010, pages 1320–1326.

[6] Karin Becker and Diego Tumitan. Introdução à mineração de opiniões: conceitos, aplicações e desafios. *Simpósio brasileiro de banco de dados*, 2013.

[7] IMDB website. http://www.imdb.com/. [Accessed 2015-04-05].

[8] Statista. http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed 2015-04-02].

[9] George A Miller. Wordnet-about us. *Wordnet. princeton university*, 2009.

[10] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: a publicly available lexical resource for opinion mining. In *Proceedings of lrec*. Volume 6. Citeseer, 2006, pages 417–422.

[11] Neo4j. http://neo4j.com/. [Accessed 2015-04-06].