Descoberta de Conhecimento em Redes Sociais e Bases de Dados Públicas

Aluna: Fernanda de Camargo Magano
Orientadora: Profa. Dra. Kelly Rosa Braghetto
Universidade de São Paulo, Instituto de Matemática e Estatística
fernanda.magano@usp.br — https://linux.ime.usp.br/~nanda/mac0499/

Introdução

Com o impacto das redes sociais na sociedade em que estamos inseridos e o avanço de técnicas de aprendizado de máquina, a opinião das pessoas acerca de determinados assuntos pode ser obtida para melhorar a qualidade de serviços existentes. O presente trabalho estuda 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. Essas técnicas permitiram identificar a polaridade dos *tweets*, ou seja, classificar os *posts* como sendo uma opinião positiva ou negativa. Para armazenar os dados e facilitar a visualização das informações foi utilizado um banco de dados orientado a grafos.

A escolha do domínio de cinema se deve à grande disponibilidade de dados abertos online. Utilizou-se a base de dados do *Internet Movie Database* (IMDB), que disponibiliza informações sobre os filmes, e dados extraídos de *posts* na rede social *Twitter* [3].

Descoberta de Conhecimento em Bases de Dados (KDD)

O *KDD* se preocupa com o processo como um todo para obtenção de conhecimento, desde a coleta até como os dados serão interpretados e visualizados. É um campo bastante interdisciplinar, abrangendo conhecimentos de aprendizado de máquina, reconhecimento de padrões, inteligência artificial, estatística, entre outros. A mineração de dados é uma parte fundamental do processo *KDD*.

As etapas do processo de descoberta de conhecimento em bases de dados são:

1. Coleta e seleção;
2. Pré-processamento;
3. Transformação;
4. Mineração de dados;
5. Interpretação e avaliação.

Mineração de opinião

Mineração de opinião ou análise de sentimento (*sentiment analysis*) é definida como o estudo computacional que envolve opiniões, sentimentos e emoções expressas em um conjunto de dados. Consiste na identificação da opinião expressa num conjunto textual e a classificação em categorias de acordo com a sua polaridade na frase ou no texto analisado.

Métricas utilizadas

Precisão (P): medida de relevância dos resultados e refere-se a uma baixa taxa de falsos positivos.

$$P = \frac{T_p}{T_p + F_p}$$

Cobertura ou revocação (R): medida de quantos resultados relevantes foram retornados, dentre os possíveis, referindo-se à baixa taxa de falsos negativos.

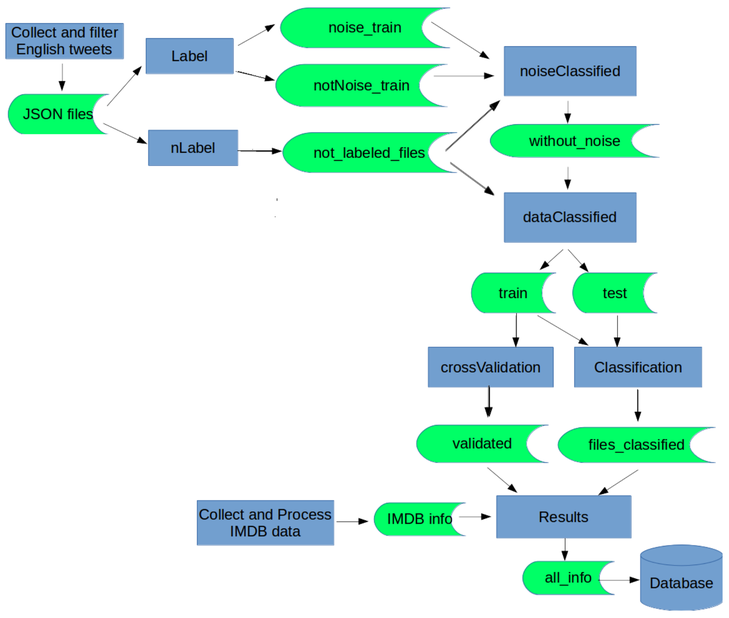
$$R = \frac{T_p}{T_p + F_n}$$

Medida F_1: média harmônica da precisão e da cobertura.

$$F_1 = \frac{2PR}{P + R}$$

Workflow construído

Figura 1: Workflow construído



Localização e distribuição dos tweets

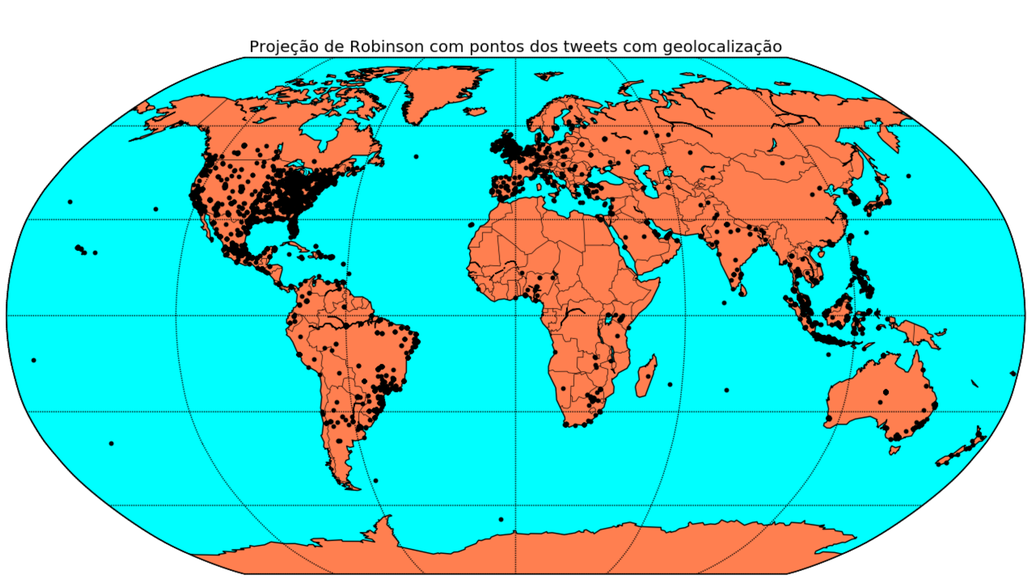


Figura 2: Mapa construído para ilustrar a distribuição dos tweets com geolocalização habilitada

Classificadores e validação

Foram utilizados dois classificadores para categorizar os dados coletados: *Naive Bayes* e *SVM*. Ambos são conhecidos por apresentar bom desempenho para análise de sentimentos [2].

A tabela abaixo ilustra um dos resultados da validação cruzada para o *SVM* realizando-se estratificação, isto é, considerando-se o desbalanceamento do tamanho das classes.

Filmes	Classe	Precisão	Cobertura	Medida f
Alice Through ...	neg	1,00	0,54	0,70
	pos	0,86	1,00	0,93
Finding Dory	neg	0,75	0,27	0,40
	pos	0,68	0,94	0,79
The Jungle Book	neg	1,00	0,20	0,33
	pos	0,81	1,00	0,89
The Revenant	neg	1,00	0,18	0,31
	pos	0,79	1,00	0,88
Warcraft: The ...	neg	0,67	0,29	0,40
	pos	0,81	0,96	0,88
Zootopia	neg	1,00	0,14	0,25
	pos	0,86	1,00	0,92

Figura 3: Validação feita por filme com as métricas de recuperação de informação: precisão, cobertura e medida f.

Estrutura do banco de dados

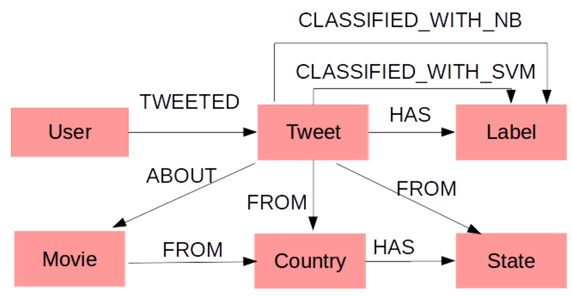


Figura 4: Estrutura do banco de dados orientado a grafos

Padrões reconhecidos

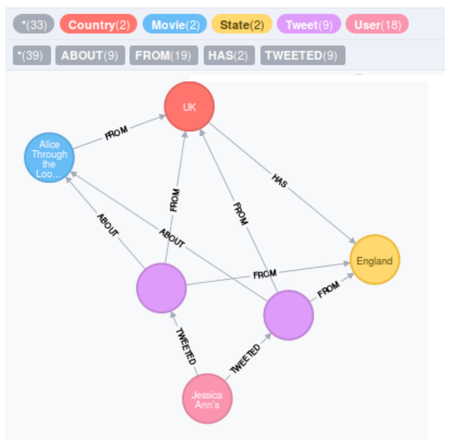


Figura 5: Exemplo de informação e visualização obtida a partir da organização e armazenamento dos dados no Neo4J

A figura 5 ilustra o caso de um usuário que postou dois *tweets* diferentes para o mesmo filme. A identificação de padrões nos dados permite a construção de conhecimento.

Conclusão

Os dados coletados passaram por uma etapa necessária de pré-processamento para a eliminação de ruídos, filtragem do idioma e escolha dos *tweets* com geolocalização habilitada - informação relevante para possibilitar a visualização da distribuição dos *tweets* do idioma inglês pelo mundo.

Foram utilizados os classificadores *Naive Bayes* e *SVM* que obtiveram resultados similares para as métricas analisadas, embora o desempenho do *SVM* tenha sido um pouco melhor nos testes feitos no processo de validação. Com o rótulo de cada *tweet* armazenado no banco de dados foi possível encontrar padrões que geram conhecimento, como exemplo, identificar usuários realizando duas postagens sobre um filme. Por ser um trabalho bem abrangente, algumas sutilezas foram percebidas no processo de desenvolvimento, ao classificar, validar e analisar os dados.

Referências

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. In *AI Magazine*, volume 17, 1996.

[2] Bernardini F. C. Lunardi A. C., Viterbo J. Um Levantamento do Uso de Algoritmos de Aprendizado Supervisionado em Mineração de Opiniões. 2015.

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