Descoberta de conhecimento em redes sociais e bases de dados públicas

Trabalho de Formatura Supervisionado
Bacharelado em Ciência da Computação - IME USP
Aluna: Fernanda de Camargo Magano
Orientadora: Profª. Drª. Kelly Rosa Braghetto

16 de novembro de 2016

Introdução

- ➔ **Impacto das redes sociais** na sociedade em que estamos inseridos;
- ➔ Avanço de técnicas de **aprendizado de máquina**;
- ➔ **A opinião** das pessoas **importa** e pode ajudar a melhorar a qualidade de serviços existentes.
- ➔ O presente trabalho estuda: como 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.

Descoberta de Conhecimento em Bases de Dados

- Preocupa-se com o processo como um todo para obtenção de conhecimento;
- É um campo bastante interdisciplinar;
- Etapas do processo:
  - Coleta e seleção;
  - Pré-processamento;
  - Transformação;
  - Mineração de dados;
  - Interpretação e avaliação.

Mineração de opinião

- ➔ Também conhecida por **análise de sentimento** (*sentiment analysis*);
- ➔ Estudo computacional que envolve **opiniões, sentimentos e emoções** expressas em um conjunto de dados;
- ➔ Identificação da opinião expressa num conjunto textual ;
- ➔ Classificação em categorias de acordo com a sua **polaridade** na frase ou no texto analisado.

Coleta dos dados

- ➔ Uso da Streaming API e da OMDbAPI;
- ➔ O domínio escolhido é o de filmes do cinema;
- ➔ Busca por palavras-chave;
- ➔ *Track* para filtrar; espaços para o AND e vírgulas para o OR;
- ➔ Formato JSON, com muitas informações.

Coleta de dados


```
{"created_at":"Thu May 26 22:47:12 +0000 2016","id":735965539835641856,"id_str":"735965539835641856",
"text":"Going to see the new alice in wonderland \ud83d\ude2d\ud83d\ude2d",
"source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e",
"truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,
"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1358013224,"id_str":"1358013224",
"name":"cristiii\u2614\ufe0f","screen_name":"Cristiii0","location":"Dallas, TX","url":null,"description":"guess",
"protected":false,"verified":false,"followers_count":1062,"friends_count":638,"listed_count":0,"favourites_count":35563,
"statuses_count":38200,"created_at":"Tue Apr 16 22:27:50 +0000 2013","utc_offset":null,"time_zone":null,
"geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED",
"profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",
"profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png",
"profile_background_tile":false,"profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED",
"profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,
"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/735198499730423809\/OP43s4Pw_normal.jpg",
"profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/735198499730423809\/OP43s4Pw_normal.jpg",
"profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/1358013224\/1457763000","default_profile":true,
"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,
"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,
"entities":{"hashtags":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,
"filter_level":"low","lang":"en","timestamp_ms":"1464302832932"}
```

Pré-processamento e Extração de características

- ➔ Capitalização e remoção de *stopwords*;
- ➔ Tf-idf (*term frequency - inverse document frequency*);
- ➔ **Vetorização** dos dados: são transformados num vetor numérico de *features* (características) com as frequências de ocorrência de cada *token*.

Matriz de características

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \begin{array}{c} \overset{atributos}{\overbrace{\phantom{x_{11}\ x_{12}\ \cdots\ x_{1m}}}} \\ \left(\begin{array}{cccc|c} x_{11} & x_{12} & \cdots & x_{1m} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2m} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{11} & x_{12} & \cdots & x_{nm} & y_n \end{array}\right) \end{array} \overset{classes}{}$$

Algoritmos de mineração: NB e SVM

- ➔ Foram escolhidos dois algoritmos de classificação: ***Naive Bayes*** e **SVM** (Máquina de Vetores de Suporte, do inglês *Support Vector Machine*);
- ➔ *Naive Bayes* utiliza probabilidades condicionais (Teorema de Bayes);
- ➔ SVM procura maximizar as margens formadas de vetores de suporte;
- ➔ Ambos são considerados com bom desempenho para o domínio de filmes

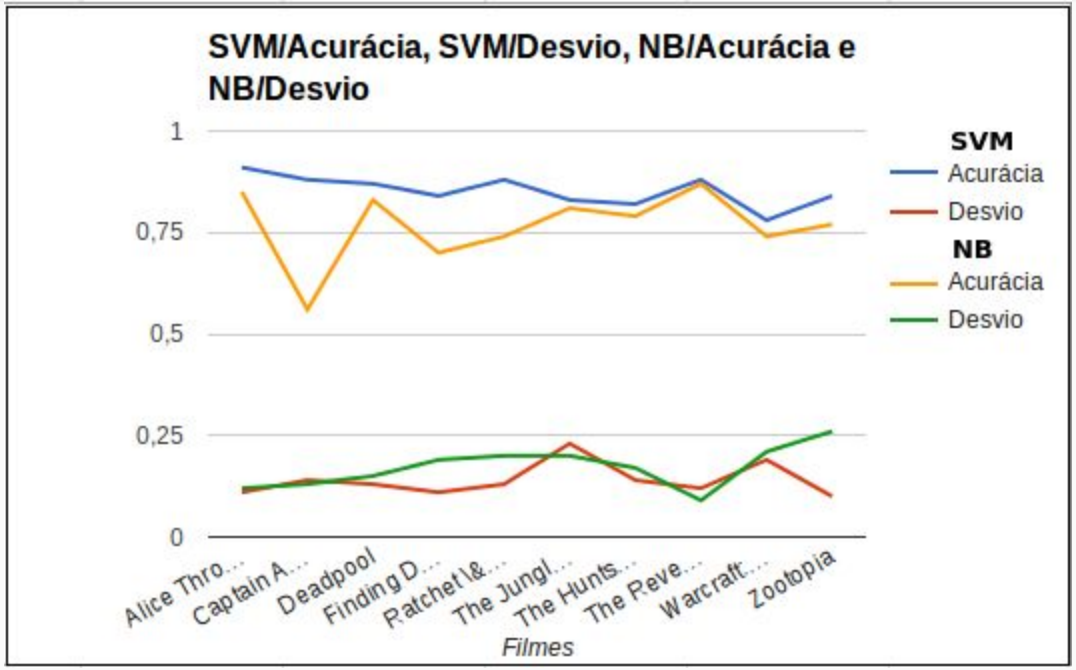
Corpus construído

- ➔ Total de 4.175.576 *tweets* em inglês de **10 filmes**;
- ➔ Desses, foram selecionados **2401** *tweets* para rotular;
- ➔ Manualmente rotulados em **seis categorias**: muito positivo, positivo, neutro, negativo, muito negativo e ruído;
- ➔ **Ruídos** podem ser propagandas e dados não relacionados com o que se busca;
- ➔ A escolha manual de rótulos é **subjetiva;**
  - ◆ Atividade trabalhosa, mas fundamental para a classificação de dados;
- ➔ **Quanto maior o conjunto de treinamento, melhor.**

Validação dos dados classificados

- ➔ Tarefa importante para garantir a qualidade dos dados classificados;
- ➔ Conjuntos de treinamento, de teste e de validação;
- ➔ Utilização de ***k-folds***, com k=10;
- ➔ Validação cruzada com **estratificação** para considerar o desbalanceamento das classes;
- ➔ Muitos *tweets* foram positivos, possivelmente por serem filmes bastante aguardados.

Comparação SVM e NB



Classificador de ruídos

Tabela de métricas calculadas na etapa de validação com SVM

CLASSES	precisão	cobertura	medida f
não ruído	0,84	0,96	0,90
ruído	0,92	0,71	0,80
média / total	0,87	0,87	0,86

Validação com duas classes

Tabela de métricas calculadas na etapa de validação com SVM

CLASSES	precisão	cobertura	medida f
neg	0,86	0,47	0,61
pos	0,83	0,97	0,90
média/total	0,84	0,84	0,82

Validação com cinco classes

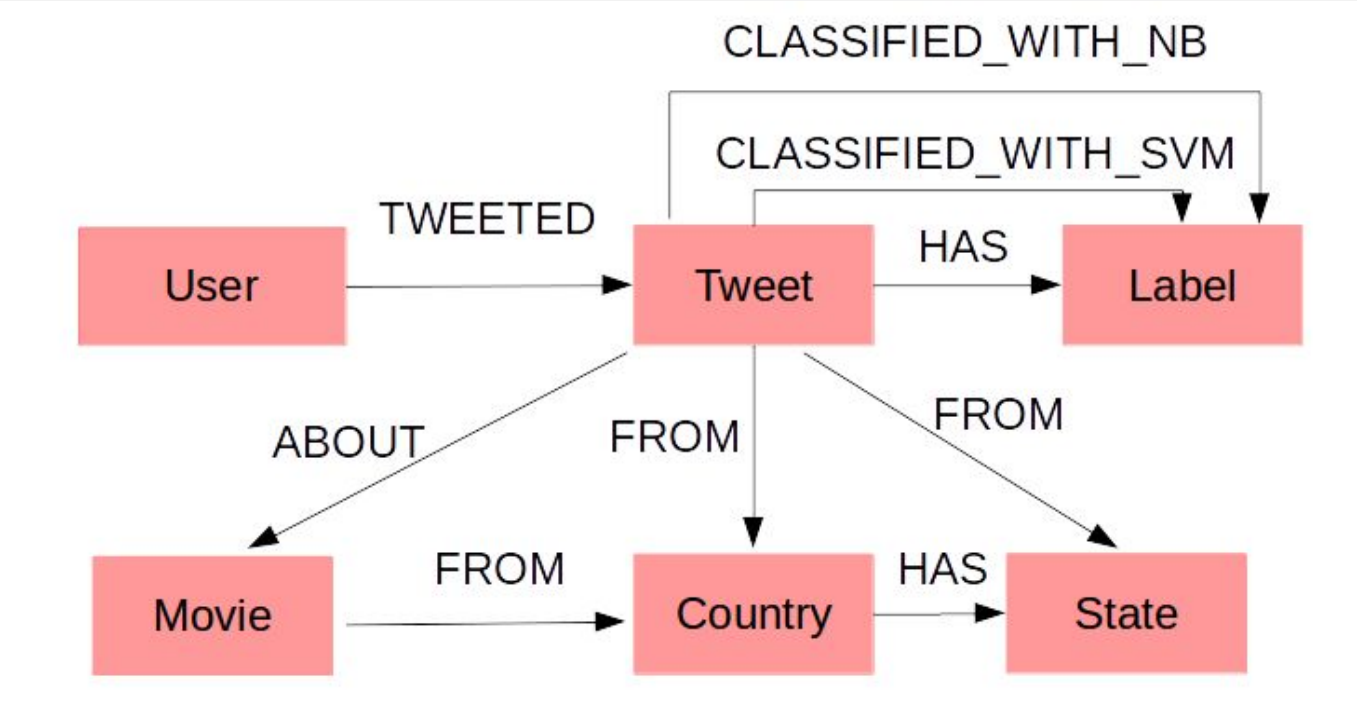
CLASSES	precisão	cobertura	medida f
muito neg	0,59	0,31	0,40
negativo	0,80	0,20	0,32
neutro	0,56	0,67	0,61
positivo	0,63	0,74	0,68
muito pos	0,70	0,63	0,66
média/total	0,63	0,62	0,61

Matriz de contingência para cinco classes

As linhas correspondem às classes reais e as colunas são os valores preditos

	muitoneg	*neg*	*neutro*	*pos*	*muitopos*
muitoneg	27	0	23	36	2
neg	5	4	3	7	1
neutro	7	0	91	33	5
pos	5	0	32	175	24
muitopos	2	1	13	27	73

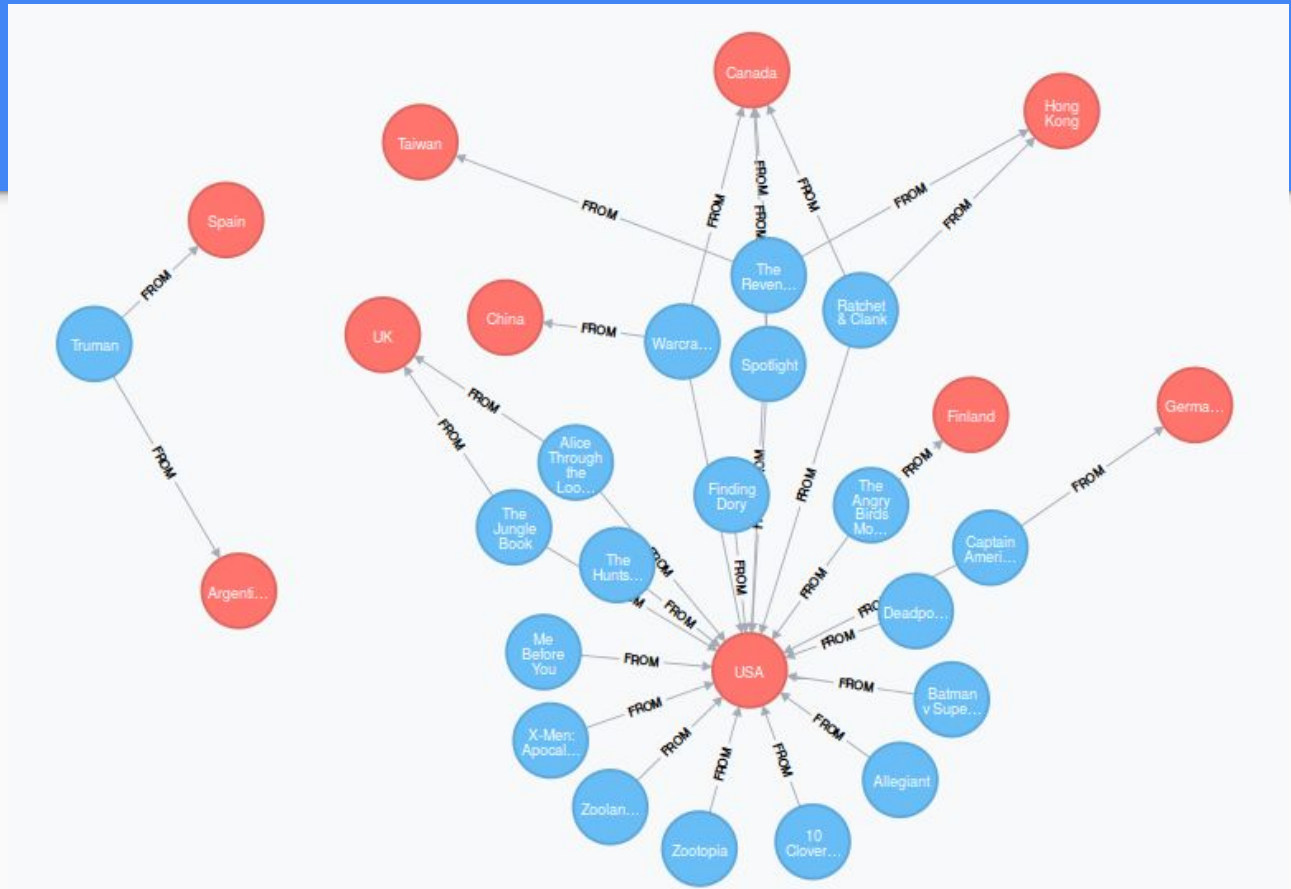
Estrutura do BD orientado a grafos



Visualização no banco de dados

Em vermelho, os países

Em azul, os filmes



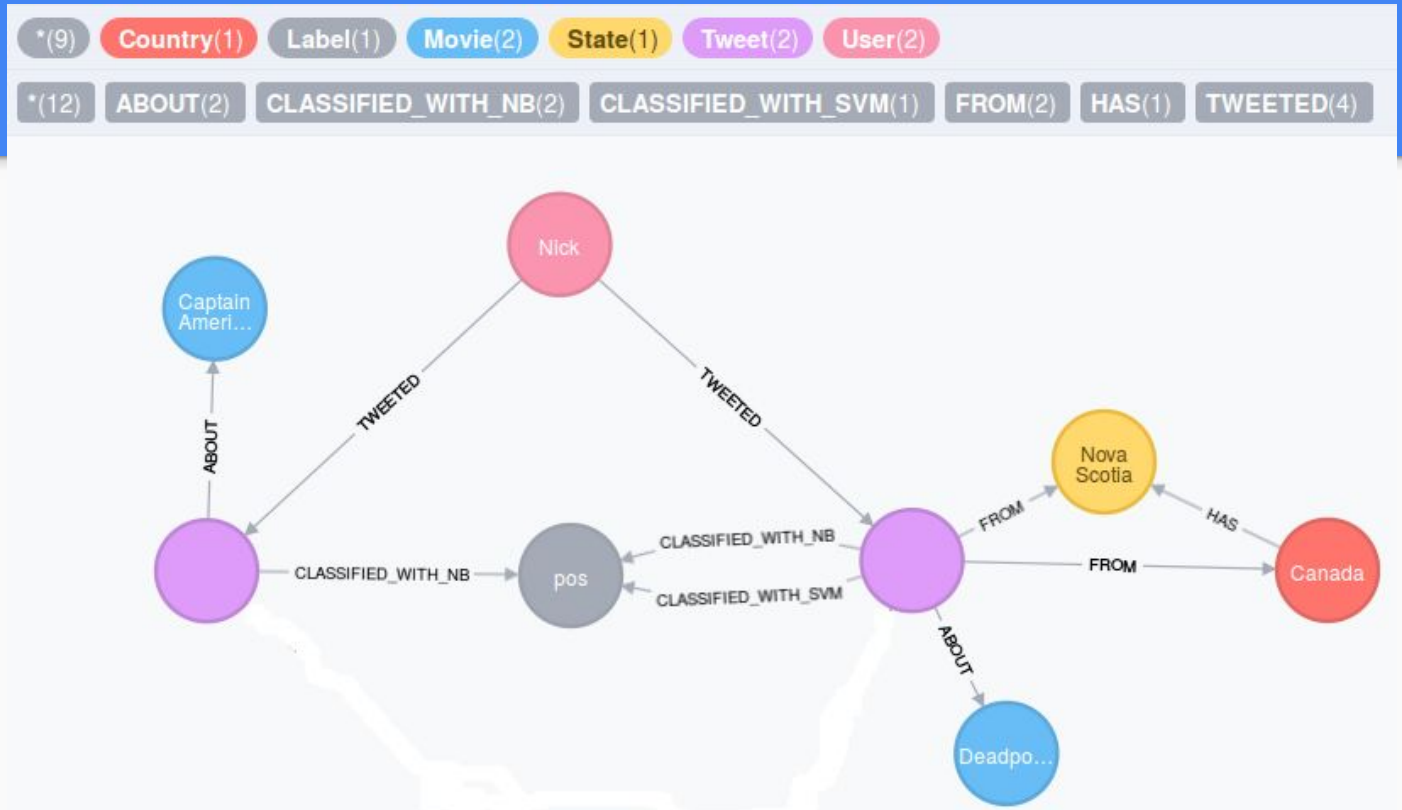
Visualização no banco de dados

Usuário

que postou

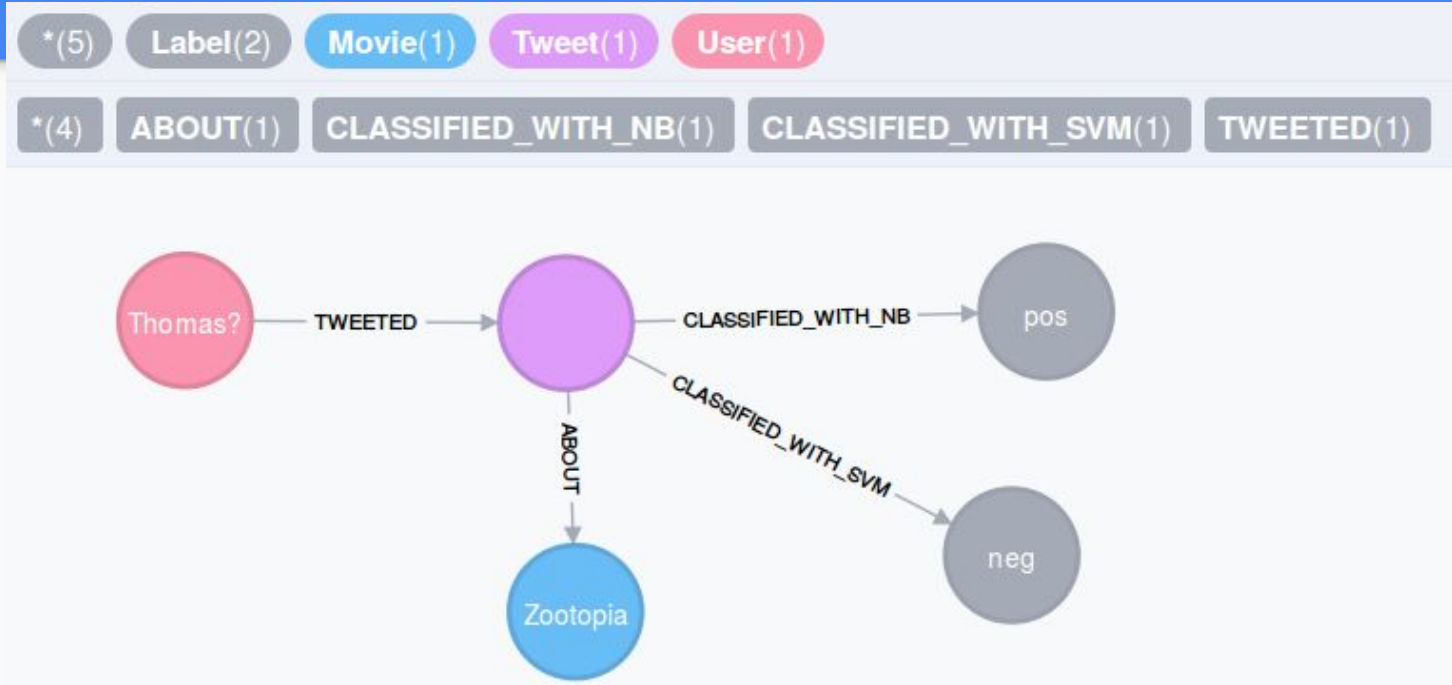
mais de uma

vez



Visualização no banco de dados

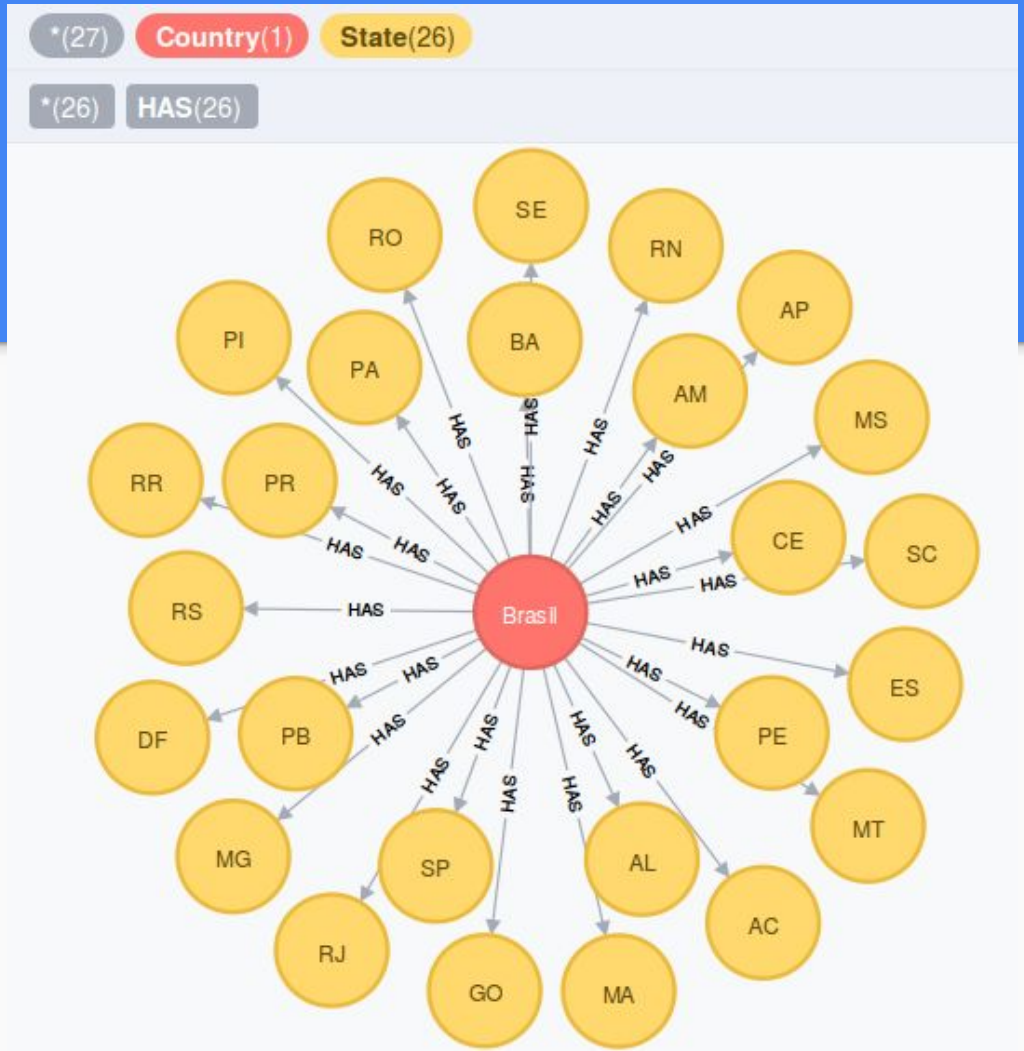
Nem sempre a classificação dos classificadores converge



Visualização no banco de dados

Estados brasileiros com

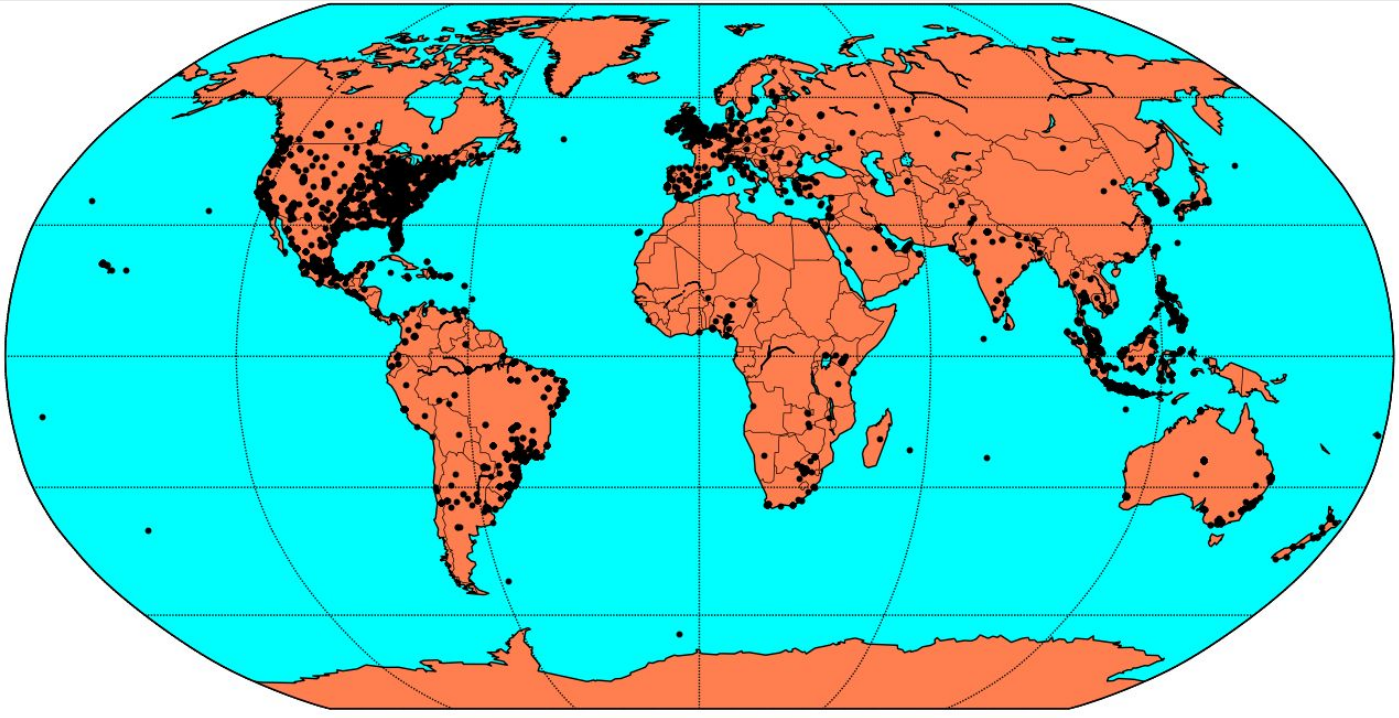
tweets em inglês



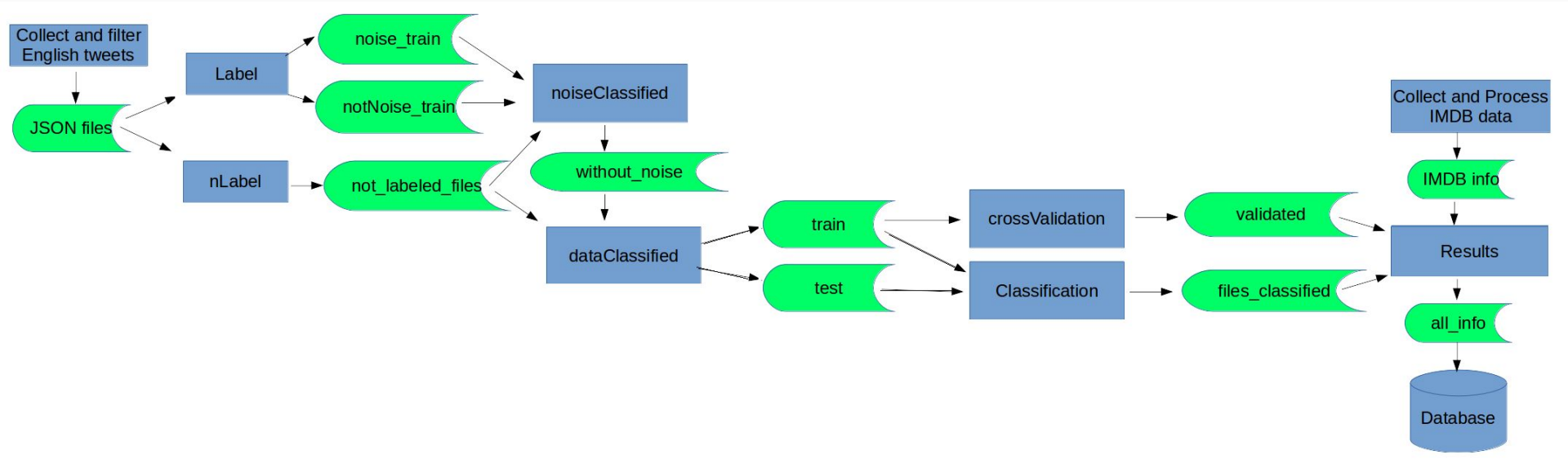
Visualização no banco de dados

m.title	m.genre	num_tweets
The Jungle Book	Adventure, Drama, Family	98
The Huntsman: Winter's War	Action, Adventure, Drama	142
Ratchet & Clank	Animation, Action, Adventure	151
The Revenant	Adventure, Drama, Thriller	242
Alice Through the Looking Glass	Adventure, Family, Fantasy	372
Warcraft: The Beginning	Action, Adventure, Fantasy	729
Zootopia	Animation, Action, Adventure	928
Deadpool	Action, Adventure, Comedy	1047
Finding Dory	Animation, Adventure, Comedy	4091
Captain America: Civil War	Action, Adventure, Sci-Fi	4746

Distribuição dos *tweets* pelo mundo



Workflow construído



Escolhas feitas e motivos

- **Linguagem** python 3
  - Utilização do scikit-learn e do matplotlib
- **Idioma** inglês;
- Escolha do **NB** e do **SVM** como métodos de classificação;
- **Twitter** como mídia social;
- **Domínio** de filmes.

Conclusão

- ➔ Os dados coletados passaram por uma etapa necessária de **pré-processamento**: eliminação de ruídos, filtragem do idioma e escolha dos *tweets* com geolocalização habilitada;
- ➔ *Naive Bayes* e SVM **obtiveram resultados similares** para as métricas analisadas, embora o desempenho do SVM tenha sido um pouco melhor;
- ➔ Foi possível **encontrar padrões** que geram conhecimento utilizando dados armazenados num banco de dados orientado a grafos;
- ➔ Trabalho bem abrangente, algumas **sutilezas** foram percebidas no processo de desenvolvimento.

Referências

- Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. In AI Magazine, volume 17, 1996.
- Bernardini F. C. Lunardi A. C., Viterbo J. Um Levantamento do Uso de Algoritmos de Aprendizado Supervisionado em Mineração de Opiniões. 2015.
- Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREc, volume 10, pages 1320–1326, 2010.
- S. H. Manning C., Raghavan P., An Introduction to Information Retrieval. Cambridge University Press, 2009.

Descoberta de conhecimento em redes sociais e bases de dados públicas

Trabalho de Formatura Supervisionado
Bacharelado em Ciência da Computação - IME USP
Aluna: Fernanda de Camargo Magano
Orientadora: Profª. Drª. Kelly Rosa Braghetto

16 de novembro de 2016