Trabalho Final de Recuperação da Informação Na Web e Redes Sociais

Diego Felipe Berg Maurício Pardin*

2017, v1

Resumo

O trabalho apresentado mostra a utilização da ferramenta KNIME para recuperação de informações da API do Spotify e Genius e da rede social Twitter para avaliar se o sentimento que as bandas e artistas passam nas letras tem relação com o que os usuários do twitter escrevem sobre elas.

1 Introdução

O trabalho apresentado mostra a utilização do KNIME para recuperação de informações da API do Spotify e Genius e da rede social Twitter para avaliar se o sentimento que as bandas e artistas passam nas letras tem relação com o que os usuários do twitter escrevem sobre elas.

2 Desenvolvimento

Para o desenvolvimento do trabalho foi utilizado a KNIME (BERTHOLD et al., 2007), que é uma ferramenta de mineração e processamento textual. Ela foi escolhida por ser *open source*, ter uma curva de aprendizado baixa e pela eficiência e flexibilidade. O KIME trabalha utilizando o conceito de *workflow* de dados, e pode ser estendido utilizando módulos para execução de códigos em diversas linguagens.

A Figura 1 apresenta o fluxo dos dados no KNIME. Nela identificamos a interação com a API do Spotify, API do Genius e do Twitter. Também podemos observar as fases de pré-processamento, enriquecimento e análise do sentimento das letras de música e dos *tweets*.

2.1 Spotify WEB API

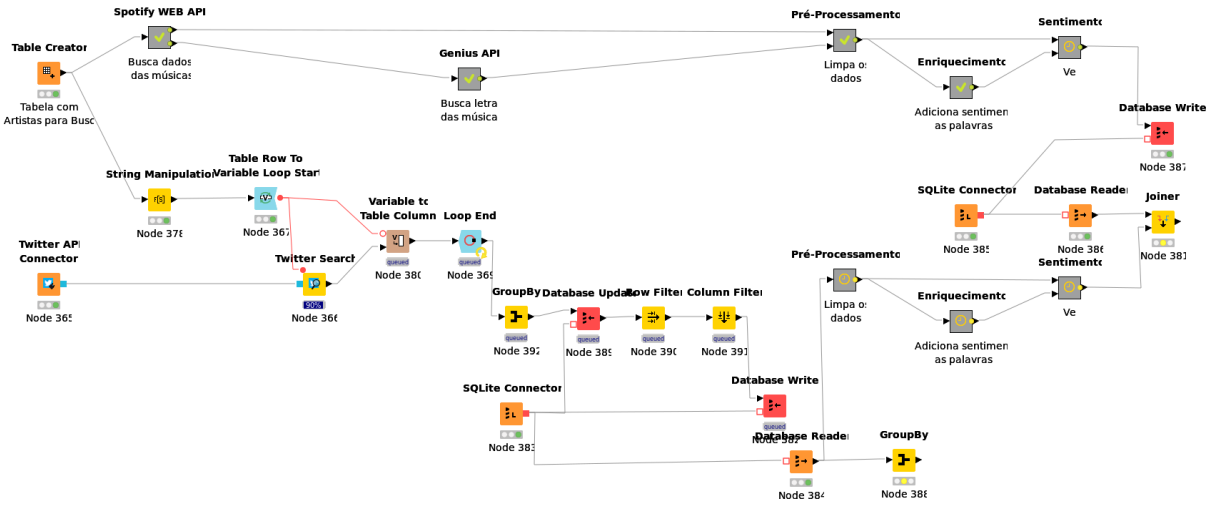
O Spotify é um serviço online de *streaming* de músicas e disponibiliza uma API que forneça informações sobre artistas e sua músicas. Dentro das informações disponibilizadas das músicas, a valência é definida por Spotify (2017) como:

> *A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).*

O KNIME conta com um grupo de nodos denominado *REST Web Services*, porém a API disponibilizada pelo serviço utiliza autenticação de uma forma que não é suportada por este grupo de nodos. Com isto foi utilizado o nodo *Python Script* para realizar as requisições ao serviço. No Algoritmo 1 é apresentado o código em *Python* utilizado para solicitação do *token* de autenticação na API. O nodo utilizada como entrada e saída de dados um *dataframe* da biblioteca pandas.

*Pós-graduando em Ciência de dados e *Big Data*

Figura 1 – Fluxo criado no KNIME



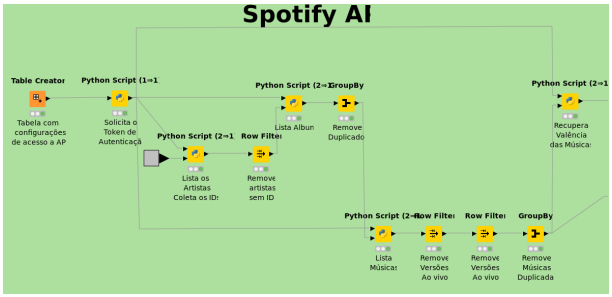
Fonte: Do autor

Algoritmo 1 – Código Python para autenticação com a API do Spotify

```
import base64
import requests

output_table = input_table.copy()

cred = base64.b64encode('%s%s%s' %
      (input_table['client_id'].any(), ':',
      input_table['client_secret'].any()))

headers = {'Authorization': 'Basic %s' % cred,
      'Content-Type': 'application/x-www-form-urlencoded'}

r =
      requests.post("https://accounts.spotify.com/api/token",
      headers=headers,
      data=input_table['request_body'].any())

tokens = r.json()

output_table['token_type'] = [tokens['token_type'],]
output_table['access_token'] = [tokens['access_token'],]
```

A Figura 2 apresenta o fluxo completo de trabalho para extração das informações da API do Spotify. Como primeiro passo, e são solicitadas as credenciais de acesso a API. Após isto é feita uma busca por artista, caso não seja encontrado o ID do artista na busca, ele é removido em um passo posterior. Após isto são listados os álbuns dos artistas, e então são removidas as entradas duplicadas. Esta ação é necessária pois há variação nos álbuns de acordo com o mercado consumidor. Após isto são buscadas as músicas destes álbuns, e removidas as versões ao vivo e as músicas duplicadas. Após esta limpeza é recuperada a valência das músicas.

Figura 2 – Fluxo de interação com Spotify



Fonte: Do autor

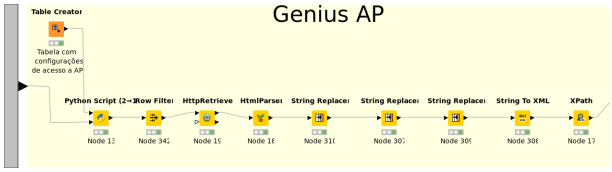
2.2 Genius Lyrics API

O Genius é um serviço de disponibilização de letras de música e também possui uma API para acesso das suas informações. Os dados obtidos da API do Spotify também foram utilizados para alimentar as requisições feitar na API do Genius. Enquando a API do Spotify retorna todos os dados disponíveis nos formatos JSON e XML, na API do Genius não é possível recu-

perar todos os dados neste formato. É possível buscar apenas a url de acesso as letras de música. Com isto foram utilizados os nodos *HTTPRetriver*, *HTMLParser* e *XPath* para extração dos dados. O também foi necessário utilizar o node *Python Script* para recuperação dos dados.

A Figura 3 apresenta o fluxo de busca das letras no Genius. Primeiro é feita uma busca por artista e música, após isto são removidos os registros que não possuem uma url para buscar as letras. Após isso é feita o download das páginas retornadas na busca. Após o *parser* do HTML são removidos utilizando expressões regulares os espaços em branco sequencias e quebras de linha para não prejudicar as buscas no *XPath*. Também foi necessário alterar um elemento do tipo *anchor* do HTML utilizado para apresentar os comentários e interpretações feitas para os trechos das músicas, esta medida foi necessária pois as palavras utilizadas para as observações das estrofes das músicas iria influenciar no resultado do sentimento atribuído música.

Figura 3 – Fluxo de interação com Genius



Fonte: Do autor

No fase de limpeza dos dados, foram convertidas as palavras para minúsculas com o node *Case Converter*. Após isto utilizando expressões regulares foram removidos os marcadores de tipo de estrofe da música e números e caracteres inválidos. O sentimento atribuído as palavras nas letras com o sentimento atrelado são apresentados na figura Figura 4 com vermelho para as palavras negativas e azul para palavras positivas.

2.3 Twitter API

O Twitter é uma rede social onde os usuários podem enviar mensagens de até 140 caracteres. Para acesso aos dados é disponibilizada a API do Twitter, para seu uso não foi necessário nenhum codificação pois o KNIME possui nodos para acesso. Na API do Twitter foram buscados apenas mensagens na língua inglesa, devido ao *corpus* utilizado para avaliação do sentimento das palavras. Foram capturados 5909 distribuídos conforme a Tabela 1.

Figura 4 – Nuvem de palavra das letras das música

Fonte: Do autor

No pré-processamento dos *tweets*, foram removidas as menções *@USERNAME* pois elas podiam ser consideradas palavras e influenciar no sentimento da mensagem. Não foi necessário remover urls e emoticons pois o *corpus* não deve ser influenciado por estes elementos. A figura Figura 5 apresenta a classificação de sentimento das palavras contidas nas mensagens, sendo apresentadas em vermelho as palavras negativas e em azul as palavras positivas.

Figura 5 – Nuvem de palavras das mensagens

Fonte: Do autor

2.4 Análise de sentimento

Os dados recuperados da API do Spotify e Genius foram enriquecidos utilizando o nodo *Dictionary tagger*, Foram utilizados dois dicionários de *corpus*, um contendo as palavras classificadas como positivas e outro com as palavras classificadas negativas.

Para a classificação de sentimento das letras das músicas e tweets foi calculada a moda do sentimento das palavras.

3 Resultados

A tabela Tabela 2 apresenta a moda do sentimento das letras da musicas, média da valência e a moda do sentimento dos tweets. Algumas bandas não possuem estratégias que usem as redes sociais, então a quantidade de mensagens obtidas foi pequena.

Tabela 1 – Tweets capturados por artista

Artista	Tweets
Adele	761
Amy Winehouse	209
Black Sabbath	283
Creedence	26
Elvis Presley	204
Faith no more	64
Hammerfall	72
Led Zeppeling	74
Metallica	85
Motörhead	362
Nelly	387
Pink Floyd	95
R.E.M.	180
Ramones	150
Rihana	155
The Rolling Stones	516
bob marley	240
manoWar	43
nofx	105
pixies	507
prince	761
radiohead	173
weezer	457
Total	5909

Fonte: Do autor

4 Conclusão

Não foi possível avaliar com precisão alguns resultados devido ao baixo nível de tweets coletados. Na maioria dos resultados foi visto que a valência média e a moda do sentimento não estão relacionadas. Apesar das letras de Ramones apresentarem palavras muitas negativas a melodia é positiva. Já no caso de Adele, temos letras positivas e melodias melancólicas. The Rolling Stones tem as letras mais tristes e a melodia mais melancólica em média, enquanto os Pixies tem as letras e melodias mais positivas. A coleta de uma base de dados maior pode alterar estes

Tabela 2 – Resultados

Artista	Letras	Valência	Tweets
radiohead	POSITIVE	0.309	POSITIVE
Amy Winehouse	POSITIVE	0.498	POSITIVE
prince	POSITIVE	0.502	POSITIVE
Metallica	NEGATIVE	0.271	POSITIVE
pixies	POSITIVE	0.575	POSITIVE
R.E.M.	POSITIVE	0.274	POSITIVE
manoWar	POSITIVE	0.137	POSITIVE
Ramones	NEGATIVE	0.634	POSITIVE
Motörhead	NEGATIVE	0.326	POSITIVE
Faith no more	POSITIVE	0.488	POSITIVE
weezer	POSITIVE	0.382	POSITIVE
Black Sabbath	NEGATIVE	0.407	POSITIVE
nofx	POSITIVE	0.415	POSITIVE
Adele	POSITIVE	0.300	POSITIVE
Pink Floyd	NEGATIVE	0.261	POSITIVE
The Rolling Stones	NEGATIVE	0.159	POSITIVE
Elvis Presley	POSITIVE	0.441	POSITIVE
Nelly	POSITIVE	0.413	POSITIVE
Hammerfall	NEGATIVE	0.125	POSITIVE

Fonte: Do autor

resultados, ou algum acontecimento específico com relação a um artista como um lançamento de álbum ou envolvimento em alguma polêmica.

Como trabalho futuro o modelo pode ser aprimorado utilizando os *emoticons* para avaliação além de outras redes sociais e também a importância das letras e profundidade.

Referências

BERTHOLD, M. R. et al. KNIME: The Konstanz Information Miner. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. [S.l.]: Springer, 2007. ISBN 978-3-540-78239-1. ISSN 1431-8814. Citado na página 1.

SPOTIFY. *Spotify Web API*. 2017. Disponível em: <https://developer.spotify.com/web-api/>. Citado na página 1.