IEC - Instituto de Educação Continuada
Pós-Graduação em Business Inteligence Analytics

Recuperação da Informação na Web e em Redes Sociais

Análise de sentimentos das músicas de 10 artistas internacionais mais tocados

Aluno: Camila Moraes
Jaqueline Gama
Professor: Zilton Cordeiro Jr.

Abril
2019

IEC - Instituto de Educação Continuada
Pós-Graduação em Business Inteligence Analytics

Projeto Final

Análise de sentimentos das músicas de 10 artistas internacionais mais tocados

Trabalho apresentado ao Instituto de Educação Continuada (IEC) da pós-graduação em Business Inteligence Analytics da PUC Minas, como requisito parcial para a obtenção de créditos na disciplina de Recuperação da Informação na Web e em Redes Sociais.

Aluno: Camila Moraes
Jaqueline Gama

Professor: Zilton Cordeiro Jr.

Abril
2019

Conteúdo

1 Resumo

O trabalho final da matéria propõe escolhermos um tema de livre escolha para análise de dados, utilizando a ferramenta Knime Analytics que permite uma análise de mineração de dados.

Para tanto nosso tema é relacionado com letras de músicas.

Vamos fazer a análise de sentimentos das músicas dos artistas internacionais mais tocados de 2019, escolhendo 10 deles.

2 Introdução

O estilo musical de hoje em dia se tornou muito eclético. A elaboração de uma música contém elementos que são o ritmo, a melodia (sequência de notas) e a harmonia.

Há vários artistas e estilos de músicas (rock, pop rock, sertanejo, funk, samba, pagode, clássica, etc) e é difícil hoje em dia falar que existe um estilo de música ou artista mais tocado, são fases e épocas, altera a todo momento, a ecleticidade tomou conta das rádios, dos sites de músicas, aplicativos, etc.

Neste trabalho resolvemos fazer uma análise, do site https://www.vagalume.com.br/.

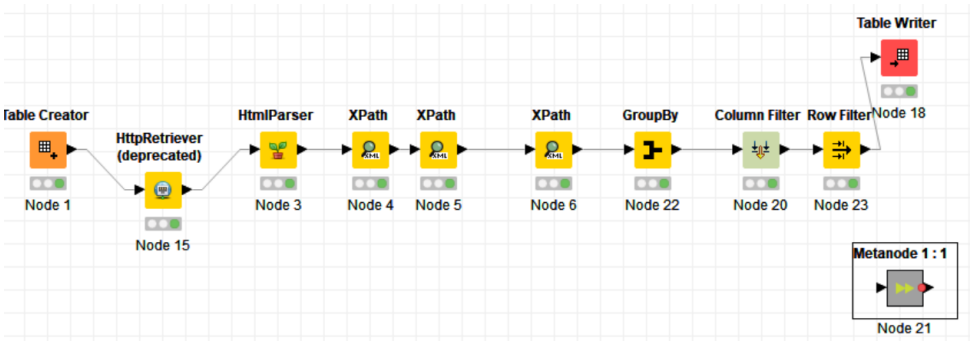
Vamos pegar os 100 artistas mais tocados no mês de março do ano de 2019 e a partir dele descobrir o "top 10" destes artistas. Queremos analisar na época atual, quais são os que estão mais "bombando". Após este filtro faremos um estudo de sentimentos das letras destes artistas selecionados, analisando suas músicas.

O Software usado para chegar a análise final será o Knime que por ser uma ferramenta mais dinâmica e simples permite trabalhar com fluxos de trabalhos de maneira mais rápida e intuitivamente para chegar a soluções e análises finais de dados.
A sua interface permite a montagem de nós ou nodes (que são entradas ou saídas de informação, sendo cada um com uma função diferente) para processamento de ETL (dados).

3 Descrição das Atividades

Abaixo vamos descrever as etapas do fluxo do trabalho e no final do trabalho, no anexo, você pode encontrar o link de acesso do Kmine.

3.1 Encontrando os artistas



A primeira parte do fluxo do nosso trabalho será baseada na extração de dados dos TOP 100 dos artistas internacionais. Para isto temos que informar qual o site que queremos realizar a coleta, e para este trabalho será retirado do site https://www.vagalume.com.br/top100/artitas/internacional/2019/03/. Usamos o node Table Creator que tem a URL a ser coletada, para que pudéssemos pegar o HTML e retirar as informações necessárias.

Para realizarmos o download da página e criar a tabela com o conteúdo no servidor automaticamente usamos o node HttpRetrivier (deprecated), pois é ele que recebe a URL informada no Table Creator.

Posterior a isto fizemos o tratamento para recuperar as URLS de cada um dos artistas, ou seja, extraímos desta URL principal, os 10 artistas “top” utilizando o node HTMLParser que extrai o conteúdo em HTML, baixado do HttpRetrivier e transforma em uma tabela com o conteúdo em XML.

O próximo passo foi extrair somente o conteúdo de nosso interesse. Para isto usamos alguns XPath que permite a navegação entre os atributos e elementos do HTML. Fizemos os tratamentos/filtros como a extração das classes, o posicionamento do objeto e um para HREF para colocar o link. O objetivo é identificar qual página você será direcionado, afim de chegarmos a página de cada artista, que é o que precisamos, ou seja, recuperar as URLS deles.

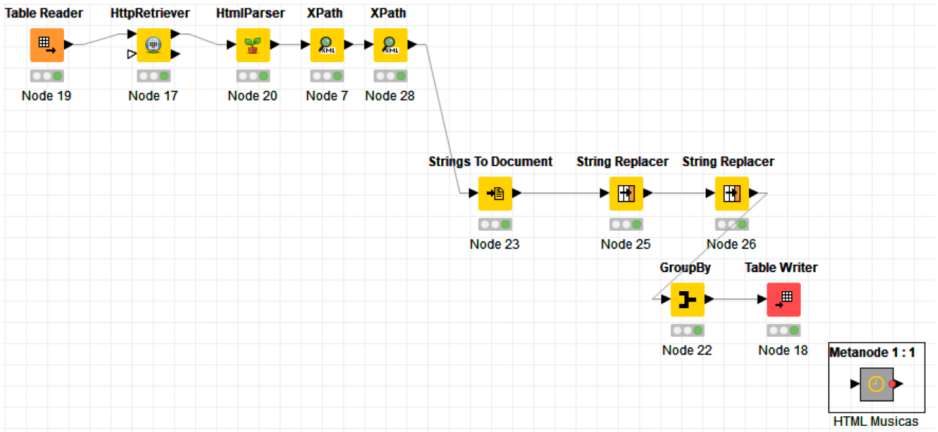
Utilizamos o node Groupby para retirar as linhas repetidas e node Column Filter para obtermos uma tabela apenas com as colunas de link que queremos trabalhar. Para finalizarmos este primeiro fluxo usamos o Rowfilter para filtrar somente os 10 artistas

3 Descrição das Atividades

que mais aparecem, em uma amostra de 100 artistas que estão na página.
No final desta análise, temos, conforme tabela abaixo, a descrição dos TOP 10 artistas internacionais.

Row ID	S headlin...
Row8	Lady Gaga
Row4	Ed Sheeran
Row1	Ariana Grande
Row10	Queen
Row5	Eminem
Row3	Coldplay
Row7	Imagine Dra...
Row9	Maroon 5
Row6	Foster The ...
Row2	Bruno Mars

3.2 Letras de músicas



A continuação do nosso trabalho, se dá com a segunda parte do fluxo (imagem acima) que buscará as letras de músicas dos artistas “top 10” selecionados na primeira parte deste fluxo.

Para isto usamos inicialmente o Table Reader que fará o armazenamento em formato de tabela dos 10 artistas e usamos alguns XPath para fazermos os tratamentos / filtros para buscar as músicas relacionadas aos 10 artistas.

3 Descrição das Atividades

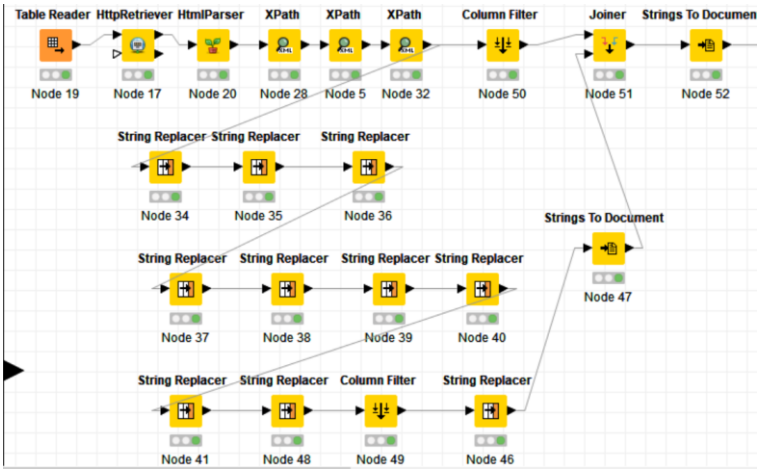
Utilizamos o node String To Document para transformar no objeto tipo “documento”. Outro ponto a se considerar é que para a nossa análise queríamos apenas letras de artistas internacionais (originais), porém o site fazia a tradução de algumas letras nacionais o que não era de nosso interesse. Para fazermos este tratamento usamos o node String Replace para tirarmos a tradução que o site fazia de algumas letras.

Agrupamos a coluna Headline (URL) para retirarmos as linhas duplicadas do node GroupBy e finalizamos com o node Table Writer para gravar uma tabela de dados em um formato interno pronto para ser lido.

Com este fluxo conseguimos observar que os artistas internacionais “Top 10”, tinham mais de 50 músicas publicadas no site analisado.

3.3 Sentimentos das músicas dos artistas “TOP 10”

Porém, queríamos também analisar os sentimentos positivos e negativos que mais se destacam entre as músicas dos artistas internacionais TOP 10, para isto fizemos a terceira parte do fluxo.



Acima está o print desta 3ª parte do fluxo com vários nodes para primeiramente pegarmos as URL das músicas, posteriormente fizemos o tratamento para retornar as informações como o nome do artista, a música e a letra das músicas. Aqui não vamos explicar cada node aplicado a este fluxo, pois muitos são repetidos das explicações que fizemos neste trabalho, nos 2 primeiros fluxos.

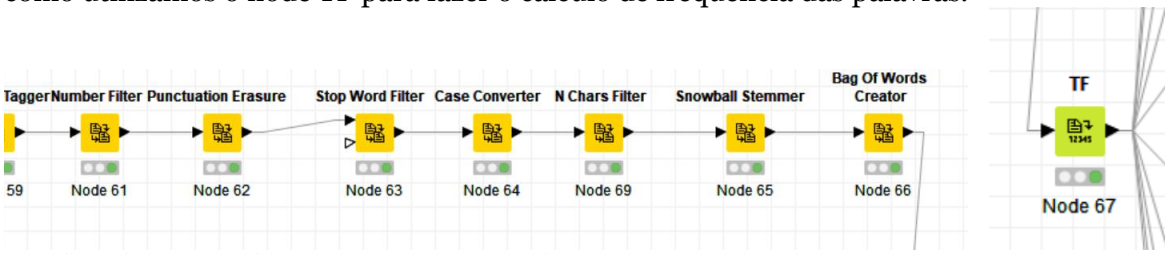
O que temos de diferente neste fluxo é a utilização do node Join que foi usado para unificar as colunas do autor, nome e letras das músicas.

3 Descrição das Atividades

Usamos o node POS Tagger e o Stanford Tagger, que detectou as classes de palavras da língua inglesa que retorna em uma coluna as informações. Posteriormente, utilizamos o node OpenNLP Tagger para reconhecer as categorias no texto.

Colocamos 2 nodes de Dictionary para análise de sentimentos. Usamos 2 arquivos de palavras de sentimentos, sendo um positivo e outro negativo e foi feito um tratamento para podermos utilizar as palavras que remetem a estes sentimentos nas músicas.

Fizemos ainda um tratamento para retirada de pontuações, palavras pequenas e vogais. Bem como utilizamos o node TF para fazer o cálculo de frequência das palavras.



Por fim, fizemos o filtro por artista, para fazermos a análise de sentimento das suas músicas. Para cada artista foi gerado uma TEG com o sentimento positivo e negativo das palavras que mais aparecem.

Fizemos alguns tratamentos e filtros como por exemplo, determinar exemplos negativos com a cor vermelha e os positivos com a cor verde, conforme pode ver alguns exemplos abaixo:

Bruno Mars

3 Descrição das Atividades

Lady Gaga

3.4 Análise de 2 dos artistas "TOP 10" no Twitter

Queríamos para finalizar o trabalho fazer uma análise de 2 dos artistas, que consideramos os mais populares, sendo eles Lady Gaga e Bruno Mars, para gerar um gráfico dos usuários que mostram conexão entre eles.

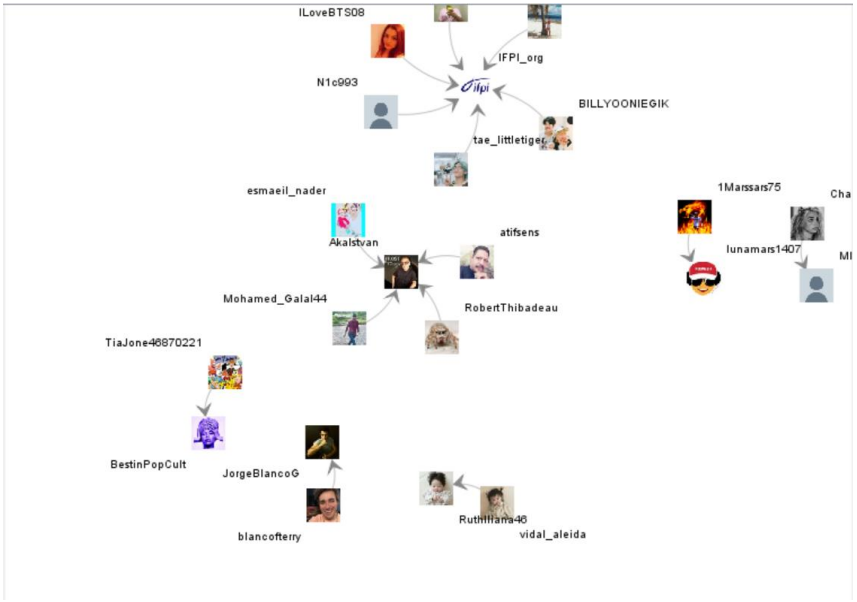
Para iniciar este último fluxo utilizamos o node para conectar no Twitter. Usamos o Twitter Search para realizar e extrair as informações mais relevantes em um banco de dados.

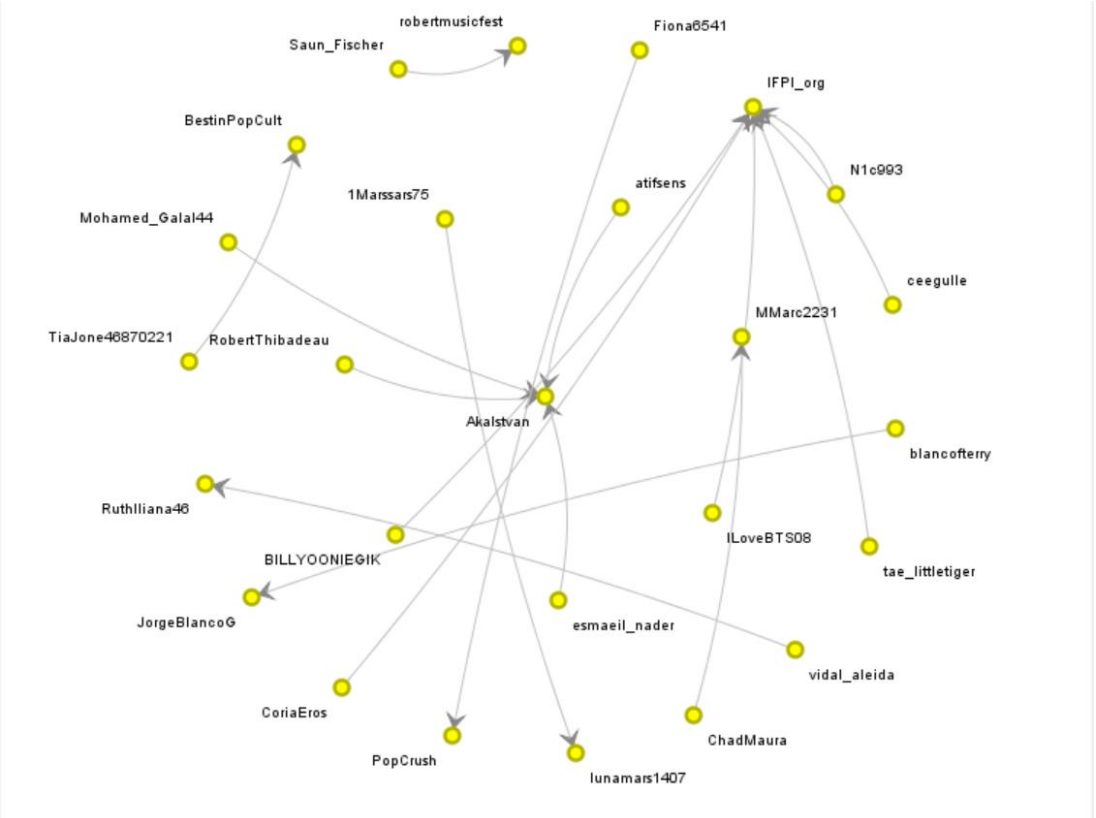
Foi criado um banco de dados utilizando o nó Database Writer e SQLite, pois com o banco de dados local a consulta fica mais eficiente.

Usamos o node Row Filter com os "retweet from" da coluna para excluir os missing. Já com o node Group By agrupamos os "retweet from" e os usurários.

No final do fluxo usamos o node Muti Feature Inserter para filtrar os perfis dos usuários e usamos o node Network Viewer para gerar o grafo final de conexão entre os usuários que fizeram retweet com os 2 artistas selecionados.

3 Descrição das Atividades





4 Análise dos Resultados

Ao finalizar o fluxo podemos ter a visão dos artistas TOP 10 internacionais e analisando os sentimentos de suas músicas observamos que o sentimento positivo que mais aparece nas letras é a palavra **LOVE**. Vale ressaltar que dos 10 artistas somente em 2 deles ela não aparece como destaque.

Já dos sentimentos negativos, observamos 3 palavras que mais se destacam, sendo elas: **Lost, Bad e Cold.**

Buscamos com a análise do Twitter ver a quantidade de usuários que twettam e retweetam sobre os 2 artistas ao mesmo tempo. Porém, observamos que as pessoas que comentam sobre eles, é pequena considerando uma análise de 10 mil dados.

O gráfico final mostra a rede de conecção entre os usuários que twettam sobre estes artistas.

Se analisarmos as pessoas que comentam dos artistas em separados observamos uma quantidade bem maior de twitter.

5 Trabalhos Futuros

Pensando em trabalho futuros, poderíamos fazer a mesma análise com o próximo ano para verificar se a análise de sentimentos e artistas varia de ano a ano, ou se os artistas sem mantem como os preferidos.

Uma análise com artistas nacionais para comparar por exemplo se os sentimentos são semelhantes aos internacionais, seria interessante.

Outra análise seria das músicas TOP 100 quais seriam os estilos mais tocados, seja rock, samba, pop, sertanejo, etc.

Seria possível também pegar estes termos que mais se apresentam para fazer uma análise de sentimentos no Twitter relacionado aos artistas gerando um grafo dos usuários para os Twitter positivos e negativos.

BIBLIOGRAFIA

VAGALUME MÍDIA. Top 100. Disponível em: <https://www.vagalume.com.br/top100/artitas/internacional/2019/03/> Acesso em: 22 de abril de 2019

KNIME. Disponível em: < https://www.knime.com/knime>
Acesso em: 22 de abril de 2019

Anexo

Google Drive

https://drive.google.com/open?id=1bxjqA4fGtaEwRboe5ZBMpurbxuz1xK89

DropBox

https://www.dropbox.com/sh/m82j9s21oog20i7/AAD2yWz78ICgjkHmqPhvAFnwa?dl=0