

Dióginis Carvalho Pinheiro

# **Análise de preferências musicais baseada no Twitter**

Campos dos Goytacazes, RJ

16 de dezembro de 2019

Dióginis Carvalho Pinheiro

## **Análise de preferências musicais baseada no Twitter**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para a obtenção do título de Bacharel em Ciência da Computação, sob orientação de Prof. Luis Antonio Rivera Escriba.

Universidade Estadual do Norte Fluminense Darcy Ribeyro – UENF

Centro de Ciência e Tecnologia – CCT

Laboratório de Ciências Matemáticas – LCMAT

Curso de Ciência da Computação

Orientador: Prof. Luis Antonio Rivera Escriba

Campos dos Goytacazes, RJ

16 de dezembro de 2019

---

Dióginis Carvalho Pinheiro

Análise de preferências musicais baseada no Twitter/ Dióginis Carvalho Pinheiro. – Campos dos Goytacazes, RJ, 16 de dezembro de 2019-  
65 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Luis Antonio Rivera Escriba

Monografia (Bacharelado) – UENF-CCT-LCMAT-Ciência da Computação, 16 de dezembro de 2019.

CDU 004.41 : 004.4'2 :

---

Dióginis Carvalho Pinheiro

## **Análise de preferências musicais baseada no Twitter**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para a obtenção do título de Bacharel em Ciência da Computação, sob orientação de Prof. Luis Antonio Rivera Escriba.

Trabalho aprovado. Campos dos Goytacazes, RJ, 16 de dezembro de 2019:

---

**Prof. Luis Antonio Rivera Escriba**  
Orientador

---

**Profa. Dra. Annabell del Real Tamariz**  
Membro da Banca

---

**Prof. Banca02**  
Membro da Banca

Campos dos Goytacazes, RJ 16 de dezembro de 2019

*Este trabalho é dedicado à minha mãe e irmã, que sempre me incentivaram e acreditaram nos meus sonhos.*

# Agradecimentos

Agradeço primeiro aos colegas de classe Arthur Rangel, Ian Carlos Bertoncello, Gabriel Fiorese, Livia Freitas e Rafael Ghossi, que tiveram papel importante nessa conquista. Ao longo dos anos nos tornamos uma família, o apoio e convivência diária tornou tudo mais fácil.

A minha família que tem sido a base de todas as minhas conquistas, que sempre apoiou todas as minhas decisões e me encorajou nos momentos de dúvidas.

A todos os professores por todos os ensinamentos durante todos esses anos na UENF, foram anos de troca de informação e aprendizados.

*"Once you stop learning, you start dying".  
(Albert Einstein)*

# Resumo

A extração de dados em redes sociais tem se tornado cada vez mais frequente, pois os dados existentes nelas são importantes e podem ter grande valor para quem os busca estudar e entender. O Twitter destaca-se como um grande provedor de dados, pois se trata de um microblog onde diversas opiniões sobre os mais diferentes tópicos podem ser encontradas. Através da extração dos dados e da mineração de textos é possível encontrar padrões sobre esses dados e obter conhecimento sobre as informações relevantes. Portanto, o presente trabalho buscou milhares de tweets sobre música, utilizando uma palavra-chave para obter informações sobre o que os usuários do Twitter estão escutando e quais informações podem ser conhecidas a partir desses tweets fazendo uma análise sobre esses dados e obtendo padrões que levam à informações dos artistas e seus respectivos gêneros.

**Palavras-chaves:** Redes Sociais. Mineração de Textos. Música.



# Abstract

The extraction of data on social networks has become increasingly frequent because the data on them are important and can have great value for those seeking to study and understand. Twitter stands out as a great data provider because it is a microblog where different opinions on different topics can be found. Through data extraction and text mining it is possible to find patterns on this data and gain knowledge about the relevant information. Therefore, the present paper sought thousands of tweets about music, using a keyword to get information about what Twitter users are listening to and what information can be known from these tweets by analyzing this data and getting patterns that lead to it. to artists' information and their respective genres.

**Key-words:** Social Media. Text Mining. Music.

# Lista de ilustrações

Figura 1 – Fases do processo de KDD . . . . .	18
Figura 2 – Espaço utilizado para o criação de mensagens no Twitter . . . . .	23
Figura 3 – Processo de análise de perfil de usuário proposto . . . . .	29
Figura 4 – Funcionamento da REST API do Twitter . . . . .	30
Figura 5 – Funcionamento da Streaming API do Twitter . . . . .	31
Figura 6 – Tokenização de Tweets utilizando NLTK . . . . .	36
Figura 7 – Tokenização de Tweets utilizando NLTK . . . . .	37
Figura 8 – Gráfico de gêneros musicais mais escutados baseado na quantidade de Tweets . . . . .	39
Figura 9 – Exemplos de pontuação . . . . .	45
Figura 10 – Remoção de pontuação e números . . . . .	45
Figura 11 – Exemplo de stop-words em inglês e português . . . . .	46
Figura 12 – Teste na versão online do software de Stanford . . . . .	48
Figura 13 – Teste na versão online do software do site Monkey Learn . . . . .	49
Figura 14 – Tabela com top 10 de idiomas dos tweets . . . . .	54
Figura 15 – Gráfico com top 10 de idiomas dos tweets . . . . .	55
Figura 16 – Tabela com top 10 fonte dos tweets . . . . .	56
Figura 17 – Gráfico com top 10 fontes dos tweets . . . . .	56
Figura 18 – Word cloud de localização dos tweets . . . . .	57
Figura 19 – Mapa dos locais dos tweets . . . . .	58
Figura 20 – Artistas nacionais mais ouvidos na busca dos meses de outubro e no- vembro . . . . .	59
Figura 21 – Top 10 Artistas mais mencionados em outubro . . . . .	60
Figura 22 – Top 10 Artistas mais mencionados em novembro . . . . .	60
Figura 23 – Gêneros musicais mais ouvidos baseado no top 10 de artistas de outubro e novembro . . . . .	61
Figura 24 – Gêneros musicais mais ouvidos baseado nos 150 artistas e nas menções que receberam nos meses de outubro e novembro . . . . .	62

## Lista de tabelas

Tabela 1 – Estrutura de Tweets com a hashtag #NowPlaying . . . . .	35
Tabela 2 – Tabelas de Tweets no arquivo CSV . . . . .	44
Tabela 3 – Primeira limpeza dos textos . . . . .	46
Tabela 4 – tweets tokenizados e sem stop-words . . . . .	50

## Lista de abreviaturas e siglas

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>14</b>
1.1	Problemática	15
1.2	Hipótese	15
1.3	Objetivos	15
1.4	Justificativa	15
1.5	Método	16
1.6	Estrutura do Trabalho	16
<b>2</b>	<b>ANÁLISE DE PERFIL DE USUÁRIO</b>	<b>17</b>
2.1	Descoberta de Conhecimento em Base de Dados	17
2.2	Mineração de dados	19
2.3	Mineração de Texto	21
2.4	Redes Sociais	21
2.4.1	Twitter	22
2.4.2	Twitter API	23
2.4.3	Uso de Hashtags	24
2.4.4	Bibliotecas do Twitter	24
2.5	Perfil de usuário	24
2.6	Trabalhos relacionados	25
2.7	Discussão das técnicas	26
<b>3</b>	<b>MODELO DE ANÁLISE DE PERFIL</b>	<b>28</b>
3.1	Obtenção de dados	29
3.2	Tweets como fonte de dados	33
3.3	Limpeza de Dados	34
3.4	Tokenização	35
3.5	Banco de características musicais	37
3.6	Clusterização	38
3.7	Análise Visual	38
<b>4</b>	<b>MODELO IMPLEMENTADO</b>	<b>40</b>
4.1	Twitter e exigências	40
4.2	Obtenção de Tweets	41
4.3	Pré-processamento de Dados	44
4.4	Tokenização	49
4.5	Caracterização	50

<b>4.6</b>	<b>Classificação</b>	<b>51</b>
4.6.1	Clusterização	51
<b>4.7</b>	<b>Atributos e padrões</b>	<b>51</b>
4.7.1	Características Musicais	52
4.7.2	Padrões	52
<b>4.8</b>	<b>Análise Visual</b>	<b>52</b>
<b>5</b>	<b>RESULTADOS</b>	<b>53</b>
	<b>Conclusão</b>	<b>63</b>
<b>5.1</b>	<b>Dificuldades</b>	<b>63</b>
<b>5.2</b>	<b>Trabalhos futuros</b>	<b>63</b>
	<b>REFERÊNCIAS</b>	<b>65</b>

# 1 Introdução

Com o crescente fluxo de dados que circulam na internet nos dias de hoje, as atenções estão cada vez mais voltadas para as redes sociais, já que as mesmas são provedoras de grande parte desses dados. Com essa expansão das redes sociais e de seus respectivos dados, muitas empresas têm procurado maneiras de tratar esses dados, buscando informação para utiliza-los na melhora de seus serviços, ou seja, moldando os para atender melhor seus clientes. O tratamento desses dados têm sido feito de diversas formas e para propósitos diferentes, uma das técnicas utilizadas é a mineração de dados, que segundo [Junior Eric Rommel \(2008\)](#) é o processo de pesquisa em grandes quantidades de dados para extração de conhecimento, utilizando técnicas de Inteligência Computacional para procurar relações de similaridade ou discordância entre dados, com o objetivo de encontrar padrões, irregularidades e regras, com o intuito de transformar dados, aparentemente ocultos, em informações relevantes para a tomada de decisão e/ou avaliação de resultados.

Esses dados aparentemente ocultos podem conter informações cruciais para analisar o comportamento dos usuários e para empresas que tem presença online pois podem direcionar as propagandas e conteúdo dos seus produtos, aumentando assim o desempenho de suas vendas e fazendo também com que o usuário tenha um experiência melhor.

A mineração de dados pode ser utilizada de várias formas e permite com que empresas possam obter maior informação sobre os clientes, agir com agilidade quando é preciso que decisões sejam tomadas pois através desses dados é possível identificar o comportamento e a preferência dos clientes fazendo com que se tenha uma previsão do que o cliente se espera daquele serviço ou recomendações de produtos e afins. Com o crescimento das redes sociais, grandes empresas têm seu foco em monitorar e analisar as interações e comportamento dos usuários nas mesmas e uma plataforma que têm se destacado pelo número expressivo de informações que são diariamente publicados pelos seus usuários é o Twitter, que é uma rede de microblog que conta com cerca de 500 milhões de tweets diários.

O twitter tem provado sua importância no mercado e mostrado como a extração de dados em sua plataforma têm sido determinante no posicionamento de empresas e de como os clientes podem ter experiências personalizadas em diversos setores. O que faz com que seja interessante trazer essa experiência na área de preferência musical, fazendo com que o usuário tenha músicas que sejam recomendadas para o mesmo baseadas em suas preferências e no que previamente foi escutado.

A intenção do trabalho é utilizar técnicas e aplicar a mineração de dados para descobrir atributos que possam ser utilizados na definição do perfil do usuário para uma

aplicação em uma base de dados coletada do Twitter que poderia contribuir para um sistema de recomendação visto que ao fazer a descoberta desses dados através da mineração, será possível construir perfis de usuários. O trabalho utiliza os dados dos usuários no Twitter se baseando em hashtags e assim constrói perfis através do estilo musical e do conteúdo que eles ouviram e compartilharam previamente nos tweets.

## 1.1 Problemática

A questão de criar uma análise para que ocorra uma criação de perfil de usuário é o que acontece para que haja uma personalização para o usuário e para que o mesmo possa ter uma experiência completamente direcionada para seus interesses e gostos pessoais. O problema a ser resolvido é lidar com o excessivo número de informações que existem nas redes sociais e que fazem com que muitas vezes se tenha acesso a informações que não são úteis ou de interesse do usuário.

## 1.2 Hipótese

Uma ferramenta tecnológica de detecção de perfis dos usuários baseado na mineração de textos de mensagens compartilhadas por redes sociais permitiria conhecer os gostos musicais, por exemplo, para melhoria de gestores nessa atividade. É de conhecimento público o quanto empresas têm trabalhado cada vez mais para criar experiências únicas para o usuário, essa ferramenta seria o passo primordial para essa personalização.

## 1.3 Objetivos

O *objetivo geral* deste trabalho é criar um modelo de perfil de usuário no contexto musical baseado em mensagens postadas pelos mesmos no Twitter. Essas mensagens seriam a base para começar o processo de criação de perfil, pois elas trazem informações da preferência musical do usuário. Através dessas mensagens será possível aplicar o processo de limpeza e mineração de dados para tornar os dados aptos para serem tratados e conhecidos.

## 1.4 Justificativa

Quando se discute sobre uma base de dados qualquer que seja é preciso conhecer e analisar para que se tenha o melhor aproveitamento possível dos respectivos dados. É exatamente por esse motivo que a personalização de uma experiência para o usuário se faz necessária, pois apesar de tanta informação nas redes sociais, é possível fazer com que se tenha algo que seja feito exclusivamente para um usuário específico, ou melhor, que tenha



um foco nos gostos pessoais e que traga ao usuário uma melhor experiência se tratando de redes sociais.

## 1.5 Método

O primeiro passo é a extração de dados que deverá ser feita através da API do Twitter até que se tenha um banco de dados suficiente para que ocorra um estudo. Após a extração, será executada a limpeza desses dados, a começar pelos links presentes nos tweets e depois o excesso de informações presentes nesses tweets. Em seguida, serão aplicados técnicas de mineração de texto para conhecer esses dados e agrupar os dados que tenham similaridade. Após o agrupamento será realizada uma análise para encontrar padrões e mostrar esses dados em forma de gráfico ou diagramas para que se tenha entendimento claro deles.

## 1.6 Estrutura do Trabalho

Este trabalho está estruturado em seis capítulos da seguinte maneira:

No Capítulo 2, apresenta-se a fundamentação teórica onde será discutido a mineração de dados e textos, redes sociais e os dados necessários para uma possível criação de perfil de usuário com análise de dados através da mineração de texto nas redes sociais.

No Capítulo 3, é apresentado o modelo proposto para realizar a análise de perfil dos usuários através das mensagens que eles postaram em suas contas.

No Capítulo 4, é apresentada a implementação do modelo proposto e os passos de limpeza de dados até que se tenha os dados prontos para serem utilizados.

No Capítulo ??, é apresentado os resultados do trabalho.

E por fim, a conclusão.

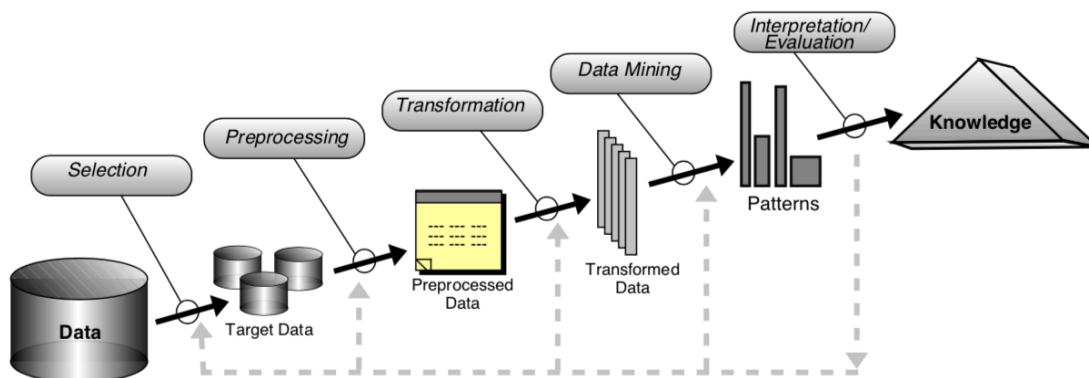
## 2 Análise de perfil de usuário

Em todas as redes sociais que criamos um perfil, além das informações que são cedidas diretamente pelos usuários ao realizar cadastro, ao longo do uso, outras informações são coletadas sem que os usuários percebam e isso faz com que a rede social tenha acesso a por exemplo, como o usuário se comporta, o que o mesmo está pensando, produtos que sejam de seu interesse e etc. Para que se tenha conhecimento da análise de perfil de usuário, é preciso entender alguns conceitos básicos que precedem essa análise. Isso implica em um domínio de técnicas de descoberta de conhecimentos em bancos de dados contendo informações que intercambiam os usuários de redes sociais. Essa descoberta de conhecimentos relacionado a um contexto, neste caso preferência musical, demanda grandes operações de dados com técnicas de mineração. No entanto, a mineração manipula informações matematizadas, que as informações de texto do contexto foram transformadas em atributos operáveis. Assim, a informação crua e sem formato, nas expressões comuns devem ser processadas e transformadas. A Figura 1 ilustra a sequência adotada nessa categoria de trabalho.

### 2.1 Descoberta de Conhecimento em Base de Dados

O Knowledge of Discovery in Databases (KDD) conhecido como descoberta de conhecimento em base de dados, como o próprio nome diz possibilita a descoberta e o conhecimento sobre um banco de dados qualquer, do menor ao maior. Existem diversas definições para KDD e podemos citar (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 2000) que define como o processo não trivial de identificar em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis, essa descoberta é feita através de um processo definido por Fayyad, Piatetsky-Shapiro e Smyth (2000) como as fases do processo de KDD.

Figura 1 – Fases do processo de KDD



Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 2000)

1. O primeiro passo nas etapas do processo de conhecimento de dados é de fato, o conhecimento e entendimento do negócio. É a análise a fundo dos objetivos e as metas do negócio, esse passo inicial é um dos principais fatores no que chamamos de definição do problema.
2. Nesse passo o conjunto de dados que será trabalhado e escolhido para que seja utilizado na análise que é pretendida. Esses dados podem ser obtidos e escolhidos através de diversas fontes diferentes como por exemplo, questionários, planilhas, api de aplicativos e etc. Assim como existem diversas formas de se obter esses dados, é esperado que os mesmos apresentem formatos diferentes, o que inclusive pode ser levado em conta na qualidade do resultado do processo.
3. Essa etapa chamada de pré processamento, pode ser considerada como uma limpeza de dados, pois é nessa parte que dados que podem interferir na qualidade do processo são eliminados, como por exemplo, dados redundantes e inconsistentes. Além disso, é considerado como serão analisados os dados incompletos e os dados que não estão dentro do padrão que é esperado pela base de dados utilizada. Outro fator importante e bastante decisório nos resultados dos algoritmos é de quantidade de variáveis em questão, por isso são utilizados métodos de redução ou transformação para facilitar o entendimento e otimizar o processo.
4. Com os dados transformados e armazenados devidamente, é possível estabelecer o método que será utilizado baseado nos objetivos que foram estabelecidos na primeira etapa do processo. Existem diversos métodos de mineração de dados e eles podem ser mais eficazes dependendo do objetivo, temos por exemplo, classificação, regressão, clusterização e etc. Esse é um passo crucial para se obter sucesso nos resultados do processo pois cada método tem seus benefícios e com isso é possível decidir quais modelos e parâmetros poderão ser trabalhados.

5. A mineração de dados consiste na busca de padrões em um determinado banco de dados seguindo os passos anteriores e utilizando os algoritmos que melhor se adequam ao problema e ao objetivo desejado, pois como dito anteriormente são diversos tipos e devem ser escolhidos de forma correta para que se tenha um resultado satisfatório.
6. A interpretação dos dados pode ser feita de várias formas, e deve ser apresentada de maneira que a visualização dos padrões encontrados seja de fácil interpretação e que possibilite análise clara caso seja necessário retornar a um dos passos anteriores.
7. A consolidação consiste em documentar e reportar o conhecimento às partes interessadas para que sejam utilizadas de forma produtiva pelo cliente, esses documentos podem ser em forma de relatório.

Esse processo é primordial e apesar de ter sido desenvolvido há alguns anos, continua sendo um guia atual e necessário a se seguir quando se trata de minerar dados pois o processo apresenta de forma clara cada uma das etapas que deve ser executada até que se obtenha o resultado desejado. Portanto, a criação de modelo de perfil usuário baseado na preferência musical de usuários utilizando como fonte a API do Twitter é o resultado da análise dos objetivos e metas, esses dados serão coletados de forma direta da API e serão pre-processados e transformados até que se tenha os dados polidos para executar a mineração de dados, seus resultados e o devido conhecimento sobre esses dados.

## 2.2 Mineração de dados

Segundo [Fayyad, Piatetsky-Shapiro e Smyth \(2000\)](#), a ideia de achar padrões úteis em dados tem sido dado uma variedade de nomes como mineração de dados, extração de conhecimento, descobrimento de informação, coleta de informações, arqueologia de dados e processamento de padrão de dados. O termo mineração de dados tem sido utilizado por estatísticos, analistas de dados e administradores de sistemas de informação.

A mineração de dados é parte do processo de descobrimento de dados e tem a função de minerar, que significa extrair algo, o que nesse caso vem a ser os dados. Essa extração faz com que se tenha uma informação melhor e mais detalhada dos dados analisados, o que acaba gerando uma descrição ou uma previsão do comportamento de um futuro fenômeno. Cada uma delas traz um sentido diferente para essa análise, a descrição é baseada em informações tal quais nos fornecem uma explicação para os resultados obtidos no processo, enquanto a previsão baseia-se em valores passados para prever comportamentos futuros esperados.

A mineração de dados consiste em escolher o algoritmo adequado ao processo, baseando-se nos objetivos e analisando como cada um desses algoritmos se relacionam

com o trabalho desejado. Alguns dos mais conhecidos são: classificação, regressão, clusterização, sumarização, modelagem de dependência e detecção de desvio.

1. Classificação é a forma de associar ou classificar um item de acordo com uma ou mais classes que foram previamente definidas, utilizando técnica estatística conhecida como análise discriminante. [Junior Eric Rommel \(2008\)](#). Podemos citar como exemplo empresas de notícias que querem dividir os interesses de seus clientes em nichos específicos, como esporte, política, música e etc. A função da classificação é mapear os textos para encontrar possíveis características que permita alocar o que foi encontrado em alguma classe.
2. Regressão é a forma mais comum de verificar um relacionamento entre variáveis, apesar de serem similares, a regressão se trata de números e não um valor categórico como a classificação. O objetivo entre a comparação de variáveis é obter uma previsão do que pode ser feito futuramente, como por exemplo, em casos onde os gastos mensais dos últimos meses de um usuário são analisados e com base no que foi estudado ter uma previsão dos novos gastos desse usuário.
3. Clusterização é o agrupamento, tem a tarefa de identificar dados similares entre si e que são diferentes de outros dados de outros grupos. Apesar da ideia ser parecida com classificação, o que difere um do outro é que no agrupamento esses dados são agrupados sem precisar que previamente tenha criado classes. Além disso, esse método não tem intenção de classificar ou estimar o valor de uma variável, ele é utilizado apenas para identificar grupos similares.
4. Sumarização determina uma descrição com dispersão reduzida para um dado subconjunto no pré-processamento dos dados, freqüentemente utilizadas na análise de descobrimento de dados. Podemos citar como exemplos simples de sumarização de dados, as medidas de posição e variabilidade. [Junior Eric Rommel \(2008\)](#)
5. Modelagem de dependência consiste em descrever dependências existentes entre variáveis e são encontrados em dois níveis estruturado e quantitativo. O estruturado é apresentado em forma de gráficos exibindo quais variáveis são dependentes, enquanto a quantidade apresenta o grau de dependência entre elas.
6. Detecção de desvio tem por objetivo encontrar conjuntos de dados com características divergentes de outro conjunto.

Após analisar as técnicas de mineração, podemos afirmar que os dados serão trabalhados de duas formas, uma delas é utilizando algoritmos de clusterização para agrupar dados que são semelhantes e também fazer uso de algoritmos de classificação, pois através de informações que são resgatadas desses tweets, iremos classificá-los de acordo com gênero musical.

## 2.3 Mineração de Texto

Segundo [Gupta e Lehal \(2009\)](#) a mineração de texto, também conhecida como análise inteligente de texto, mineração de dados de texto ou descoberta de conhecimento em texto (KDT), geralmente se refere ao processo de extrair informações e conhecimentos interessantes e não triviais de textos não estruturados. A mineração de texto é um jovem campo interdisciplinar que se baseia na recuperação de informações, mineração de dados, aprendizado de máquina, estatística e linguística computacional. Como a maioria das informações (cerca de 80%) é armazenada como texto, acredita-se que a mineração de texto tenha um alto valor potencial comercial. O conhecimento pode ser descoberto a partir de muitas fontes de informação, no entanto, os textos não estruturados continuam a ser a maior fonte de conhecimento prontamente disponível.

O processo de mineração de texto é similar e conta com estrutura parecida com o da mineração de dados. Sendo assim, através da mineração de texto será possível analisar as palavras nos Tweets dos usuários podendo ter melhor ideia do que cada uma significa e auxiliando na classificação do perfil de usuário. É uma das etapas mais importantes nesse processo pois é onde os dados que foram transformados estão prontos para serem analisados e assim encontrar padrões e similaridade entre si.

Segundo [Bezerra \(2010\)](#) existem três tipos de processamento voltado para análise de conteúdo:

- Processamento Léxico e Sintático: Envolve o reconhecimento de tokens(termos), a normalização de termos e a construção de linguagem;
- Processamento Semântico - Envolve a extração do significado inerente aos textos. Requer a extração de entidades nomeadas tais como nomes de pessoas, nomes de organizações, locais, etc.
- Processamento de Características Extra-Semânticas - Mais complexo, envolve a identificação de sentimentos nos textos analisados. Por exemplo: sarcasmo, melancolia, alegria, etc.

O processo de mineração de texto é parte fundamental do processo de realização deste trabalho, pois ajudará a conhecer melhor os dados que serão extraídos de uma rede social.

## 2.4 Redes Sociais

Rede social é um termo atualmente conhecido, mas o conceito é antigo, as redes sociais estão presentes na forma em como nos relacionamos uns com os outros

e aprendemos mais sobre o ambiente em que vivemos, por isso podemos dizer que é um sistema aberto permanentemente e que é construído tanto individualmente quanto coletivamente, pois é através dos círculos sociais que adquirimos vivência de uma vida em sociedade e constante aprendizado de práticas utilizadas no nosso cotidiano.

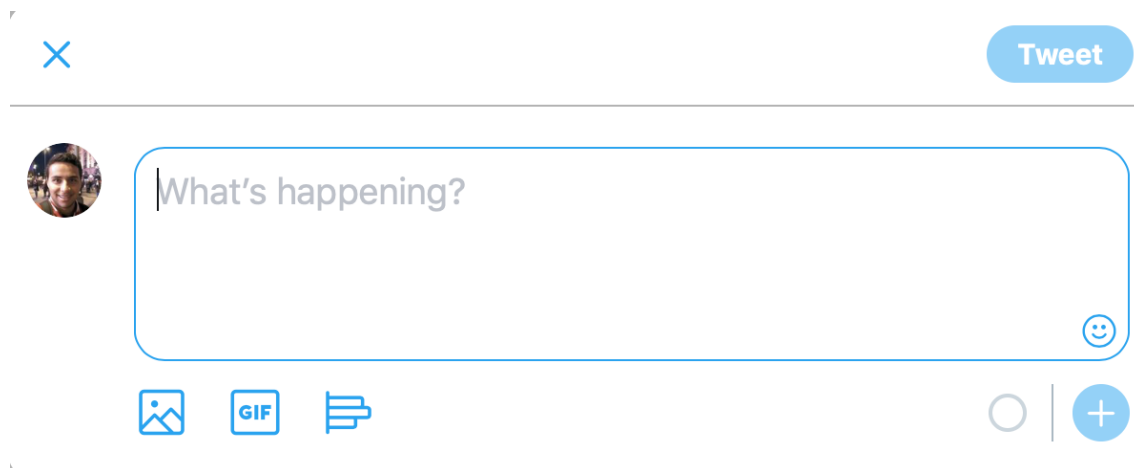
As redes sociais em si evoluíram com o tempo e hoje em dia temos um conceito diferente para as formas que nos relacionamos, grande parte disso é por causa da introdução da Internet e da forma em que o usuário pode estar conectado de qualquer lugar e dialogando com qualquer pessoa do mundo. As redes sociais online permitem que os usuários criem perfis dentro de uma plataforma onde poderão compartilhar informações para um grupo de amigos ou para qualquer usuário da plataforma. Dentro de seus perfis os usuários escolhem compartilhar o que querem e muitas vezes eles utilizam para coisas que não falaria pessoalmente, as redes sociais fazem com que os usuários percam a inibição e se sintam livres para abordarem temas que normalmente não seriam abordados pelos mesmos.

Todas as mudanças na forma como nos relacionamos e na influência cada vez maior da Internet no comportamento dos usuários têm sido abordado em diversas áreas e colocado em pauta para discussões. Com o aumento de interatividade online, as empresas tem sido as maiores beneficiadas pois estão sempre em busca de entender melhor o comportamento do usuário direcionando melhor os seus produtos para seu público alvo, isso acontece em centenas de áreas diferentes e o foco desse trabalho é o de mercado fonográfico e serviços de streaming. Com o compartilhando de suas músicas preferidas, é possível analisar esses dados e saber o comportamento dos usuários e as preferências musicais dos mesmos. Uma rede social que cresceu significativamente nos últimos anos e têm sido utilizada para compartilhar essa informações é o Twitter, pois devido a suas informações serem disponibilizadas em tempo real grande parte de seus usuários se sentem confortáveis em compartilhar com seus seguidores, o que estão fazendo, comendo, assistindo e ouvindo.

### 2.4.1 Twitter

O Twitter ficou conhecido como uma rede social onde informações curtas poderiam ser compartilhadas, no início os usuários contavam apenas com 140 caracteres para expor suas ideias, mas há 2 anos os criadores aumentaram o número de caracteres para 260, o que fez com que os usuários pudessem detalhar mais as informações que são compartilhadas. É uma das maiores redes sociais nos dias atuais e conta com um número de cerca de 500 milhões de tweets diários, tweets esses que podem conter notícias do cotidiano, informações pessoais e etc. Além disso, a rede de microblog conta com diferentes formas de disponibilizar tais informações como por exemplo, através de textos, imagens, gifs e links.

Figura 2 – Espaço utilizado para o criação de mensagens no Twitter



Fonte: Interface Gráfica do Twitter

Embora a rede social tenha passado por algumas mudanças ao longo dos anos, a característica que predominou e destacou o twitter desde sempre é o acesso a informações de forma rápida, curta e predominantemente através de texto, o que o difere de outras redes como Instagram que prioriza o compartilhamento de imagens e o Facebook que apesar de aceitar diferentes tipos de maneiras para criar informação, ainda assim grande parte dos compartilhamentos são através de fotos e vídeos. Além dos fatores anteriores, outro motivo para escolha do Twitter para esse trabalho é justamente por utilizar informações em tempo real e com dados que podem ser buscados de diversas formas através de pesquisas no próprio site ou através da extração feita na API.

#### 2.4.2 Twitter API

O twitter permite que desenvolvedores que se cadastrarem em uma área exclusiva do site tenham acesso a serviços e ferramentas que auxiliam no desenvolvimento de aplicações, para isso é preciso realizar o cadastro de uma aplicação, o que gera então permissões necessárias para acessar informações presentes na rede social. Assim como existe a facilidade ao acesso da aquisição de informação, o Twitter também possui um padrão nos tipos de informações que facilitam o processo de limpeza e a identificação do tipo de informação presente nos tweets.

Na área de desenvolvedor apresentada pelo Twitter é possível ter acesso a dois tipos de APIs que têm se destacado nessa busca de informações: Twitter API Streaming e Twitter Rest API. Enquanto o Twitter Streaming API funciona captando as mensagens que estão sendo postadas em tempo real, o que exige que se tenha uma conexão HTTP (Hypertext Transfer Protocol) sempre aberta e em funcionamento para garantir o êxito do processo, o Twitter Rest API capta as mensagens através de requisições em massa e possui outros tipos de funcionalidades como por exemplo captação de informações da timeline



de um determinado usuário, além de ter acesso as mensagens que o mesmo retweetou e obter acesso a mensagens captadas através de buscas. Ambas APIs se destacam na busca de informações e te suas vantagens, mas a escolhida para ser utilizada nesse trabalho é a API Streaming .

### 2.4.3 Uso de Hashtags

Uma das características do Twitter é a utilização de hashtags (#) que facilitam o acesso a informações e classificam tweets de acordo com o assunto que está sendo discutido, o que faz com que seja possível encontrar esses tweets de forma mais rápida e precisa. Através de pesquisas, é possível buscar tópicos específicos como a por exemplo, as hashtags #NowPlaying e #NP que serão utilizadas nesse trabalho para encontrar tweets de usuários que compartilharam músicas que estão ouvindo no momento. Um exemplo desse tipo de tweet é "Flux by Ellie Goulding (espaço contendo o link da música executado na plataforma de streaming) #NowPlaying", tweets como esse além de contar com a hashtag previamente definida como caso de estudo, contém o nome da música e do artista.

### 2.4.4 Bibliotecas do Twitter

Além das APIs que o Twitter disponibiliza na área de desenvolvedores, existem bibliotecas que embora não tenham sido criadas e testadas pelo próprio Twitter, são utilizadas e funcionam para ter conectar a API à linguagens de programação permitindo assim ter acesso às informações almejadas. Essas bibliotecas apresentadas foram elaboradas em diferentes linguagens de desenvolvimento, como Clojure, Go, Java, Javascript, PHP, Python e etc.

Por se tratar de linguagens diferentes basicamente utilizadas para o mesmo propósito, a questão de escolha da linguagem a ser trabalhada acaba sendo por detalhes, como a complexidade do desenvolvimento e a familiaridade que o desenvolvedor possui com a mesma. De acordo com as características descritas, Python foi a linguagem escolhida para a execução do trabalho.

## 2.5 Perfil de usuário

Com o auxílio da internet, o mercado fonográfico se transformou ao longo dos anos, o que antes era dominado por sites e softwares de download se transformou em plataformas de streaming trazendo o acesso a um acervo musical de forma rápida e cômoda para os usuários. Essas mudanças trouxeram uma nova forma de interação com a música digital para melhor entendimento por qualquer usuário, até mesmo para as gravadoras,

para orientar seus produtos para pessoas que consomem música dessa forma. Portanto, é preciso criar um perfil de usuários consumidores.

Qualquer sistema que deseja oferecer serviços personalizados para seus usuários precisa criar um perfil de usuário. Esse perfil é o que distingue o usuário de outros e o classifica baseado nos interesses e no que não é interessante. O perfil de usuário geralmente é criado apresentando um peso em palavras-chave que podem, possivelmente representar os interesses desses usuários.

As formas para obtenção de dados, como a criação do perfil do usuário, envolvem processamentos de dados por algoritmos. Neste caso, os dados que serão obtidos dos repositórios do Twitter utilizando a hashtag #NowPlaying até que se obtenha dados suficiente para iniciar o trabalho. Após os dados serem processados, será possível ter uma noção dos dados úteis para o estudo realizando o processo de KDD, tal como mostrado na Figura 1, afim de ter o conhecimento necessário sobre esses dados e para classificá-los de acordo com suas preferências musicais.

## 2.6 Trabalhos relacionados

A análise de perfil de usuário tem sido elaborada de diversas formas, ultimamente com o uso de redes sociais, o Twitter não é o único que tem sido estudado para estes fins. Independente da rede social ou do banco de dados utilizado para criação do perfil, diferentes métodos de classificação são aplicados para tal fim, como métodos probabilísticos, estatísticos ou de aprendizagem de máquina.

Silva Richardson Ribeiro (2014) apresentam uma proposta para identificar perfis de usuários através dos posts circulando na rede social Facebook. Nesses perfis são analisadas palavras baseadas no Anew-br, que é uma base de dados contendo algumas palavras que fazem com que se possa identificar o conteúdo delas. Esse conjunto de palavras foi traduzido para o português, e todo o trabalho de identificação é feito em cima dessa base de dados. O objetivo, então, é identificar o perfil dos usuários através de emoção. Foi utilizado como suporte de aprendizagem SVM (Support Vector Machine) para classificação binária desses sentimentos.

Junior Eric Rommel (2008) utilizam a importância do KDD e da descoberta de padrões para trabalhar com um banco de dados de uma loja que atua no setor de varejo, atacado, consórcio, empréstimo e garantia estendida afim de descobrir o perfil dos usuários em base de suas compras. Essa estratégia auxilia às empresas em planejamentos estratégicos e no marketing direcionado ao cliente. O software WEKA foi utilizado para encontrar os padrões desejados e mediante critérios previamente estipulados. Dois algoritmos de classificação foram utilizados, sendo que um deles, o Tertius, apresentou resultados em menor tempo de processamento do que o segundo, Apriori.

A ideia do trabalho proposto por [Xu Zhiheng; Ru \(2011\)](#) é a criação de um framework para descobrir o interesse dos usuários baseados no que eles postam em suas contas. Filtrando os dados baseados nos critérios desejados, os posts são processados através de tokens utilizados para excluir pontuações, termos que não foram encontrados em mais de 10 tweets, realizar a remoção de stop-words e transformar nomes compostos de entidades em uma só palavra.

[Makki et al. \(2016\)](#) propõe um framework chamado Twimer para indicar tweets relevantes para os usuários baseado em seu perfil. Utilizando palavras-chave e modelos probabilísticas de linguagem é feita uma lista de tweets rankeados de acordo com a relevância, os mesmos são submetidos a outro processo de avaliação confirmando a relevância e para finalizar, são aplicadas técnicas de agrupamento com a intenção de encontrar tweets repetidos para assim, recomendar o tweet que foi postado primeiro.

O trabalho de [Lim e Datta \(2013\)](#) cria um framework para classificar o interesse dos usuários utilizando informações da Wikipedia. Analisando as celebridades que um usuário segue, o framework proposto busca a profissão dessas celebridades na página da Wikipédia para classificá-las dentro de categorias como música, publicidade, esportes, moda e etc. Com isso, utilizando uma biblioteca com palavras-chave e associando essas palavras às celebridades, é possível classificar se os tweets postados pelo usuário estão de acordo com interesse analisado e detectado ou se são sobre qualquer outro assunto.

[Ashktorab et al. \(2014\)](#) introduz uma ferramenta de mineração para o Twitter chamada Tweedr, criada com a intenção de auxiliar correspondentes de primeiros socorros ao chegarem em situações de riscos. O trabalho tem por função dar visibilidade ao que está acontecendo em regiões de desastre, em tempo real, para que os primeiros socorros possam chegar ao local informados sobre o que houve e preparados para lidar com a situação. Para classificar se os tweets informam um desastre ou informações casuais, são utilizados métodos de classificação como sLDA, SVM e regressão. São utilizados regras de agrupamento para encontrar tweets que são similares e por fim, ocorre a fase de extração onde são trabalhados com tokens e frases para identificar os tipos de danos ocorridos.

## 2.7 Discussão das técnicas

Algumas técnicas utilizadas nos trabalhos relacionados foram escolhidas para serem aplicadas no desenvolvimento deste trabalho, como por exemplo a classificação que foi estudada nos trabalhos de [Junior Eric Rommel \(2008\)](#), [TwMessage2016](#), [Lim e Datta \(2013\)](#) e [Ashktorab et al. \(2014\)](#). Assim como as regras de associação citadas no trabalho de [Lim e Datta \(2013\)](#), uma biblioteca de palavras-chave é associada à lista de celebridades para classificar o interesse do usuário.

Outras técnicas de agrupamento e extração, que foram aplicadas nos trabalhos an-

teriores para atingir seus respectivos objetivos, se encaixam na proposta do que queremos com este trabalho.

E por fim, utilizar o SVM como classificador de texto, assim como visto no trabalho de [Silva Richardson Ribeiro \(2014\)](#).

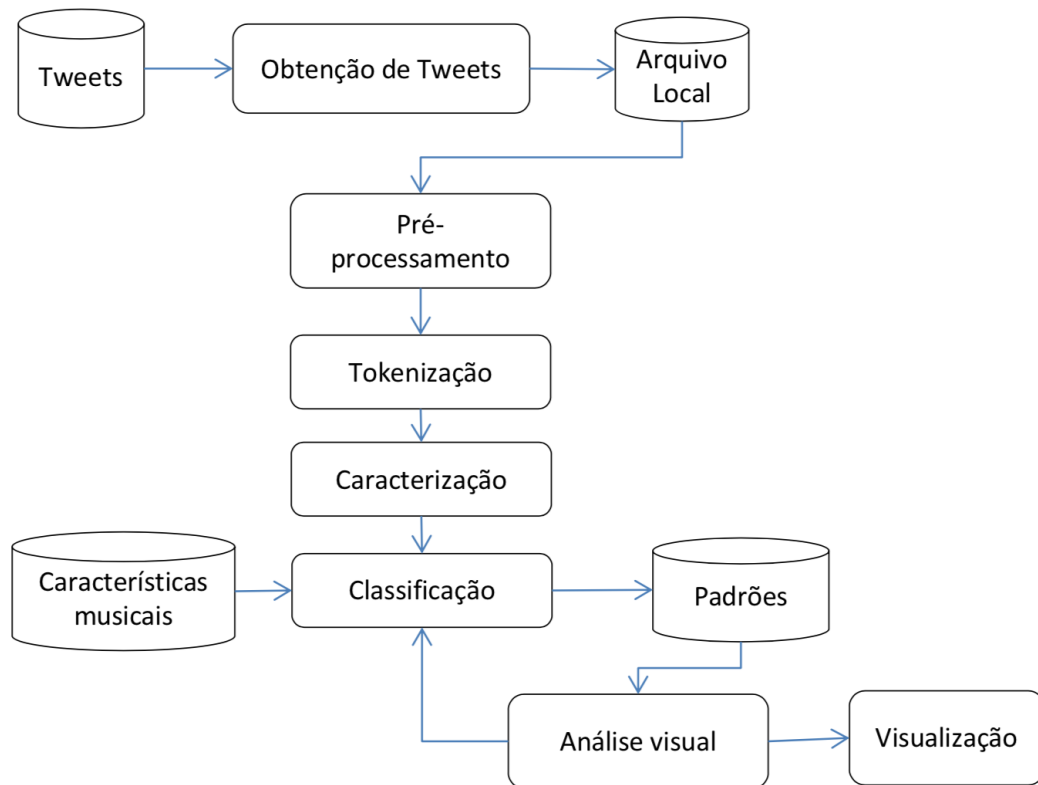
### 3 Modelo de Análise de Perfil

O Twitter é um dos maiores geradores de conteúdo nos dias de hoje, a cada segundo um número grandioso de tweets é enviado por seus usuários. Os tweets são de temas diversificados e permitem usuários se conectarem em uma escala global, pois não existem limitações dentro da rede social quando se trata do conteúdo enviado. Todos os usuários têm acesso a tweets de qualquer lugar do mundo, a não ser que o tweet seja proveniente de uma conta de perfil privado. As informações postadas permitem analisar uma série de elementos que podem auxiliar na definição de um perfil de usuário e mais que isso, trabalhar com os dados de forma ampla, permitindo por exemplo, realizar análise de sentimento baseada no que foi dito pelos usuários sobre um determinado tema, marca ou em algum momento específico da história. Além disso, o conteúdo dos tweets também pode ser estudado para descobrir, por exemplo, desastres naturais ([ASHKTORAB et al., 2014](#)) ou descobrir a revelância de tweets para cada usuário baseada no próprio conteúdo que o mesmo postou em seu perfil [Makki et al. \(2016\)](#).

O Twitter não só autoriza o acesso a API como também permite que a busca seja realizada através de dados diferentes disponibilizados pelo próprio Twitter em sua área exclusiva para desenvolvedores. Além das mensagens propriamente enviadas pelos usuários, esses tweets geram outros dados como localização, informações sobre o usuário, horário de criação e etc. Esses dados são importantes e podem ser utilizados de diversas formas, possibilitando o melhor entendimento dentro do objetivo desejado e o estudo da criação de padrões dentro dessa plataforma.

A arquitetura proposta neste trabalho e ilustrada na Figura 3, apresenta as diversas etapas do processo para análise de perfil. A começar pela obtenção dos tweets, ou seja, dos dados que são fundamentais para execução deste trabalho. Esses dados obtidos através da API do Twitter, são armazenados em um banco de dados local para serem tratados. O pré-processamento é executado de diferentes maneiras e em etapas diferentes. A extração dos links que fazem parte dos tweets, pois geralmente quando se compartilha um tweet contendo essa hashtag #NowPlaying, esses tweets vêm com o link do serviço de streaming onde o usuário estava ouvindo a música, como esses links não são úteis para o objetivo desejado, são removidos. Além de links, outros dados a serem removidos pois não fazem diferença no contexto geral do trabalho são pontuação e stop-words. Após os processos de pré-processamento de dados e baseando-se em um banco de características musicais previamente definido, pode-se então agrupar os artistas de acordo com o gênero musical estabelecido neste banco. Com os dados agrupados e reconhecimento de padrões, é possível ter então uma análise visual dos tweets.

Figura 3 – Processo de análise de perfil de usuário proposto



Fonte: Autor

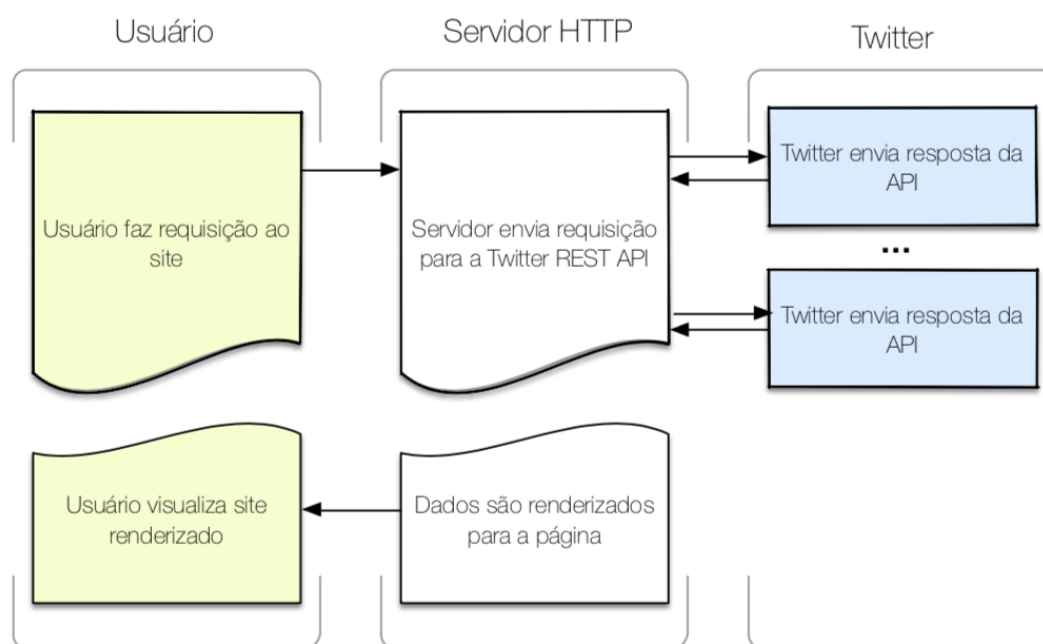
### 3.1 Obtenção de dados

A obtenção dos dados no Twitter pode ser realizada através de APIs disponibilizadas pelo mesmo que são a REST API e a Streaming API. Através do acesso a API é possível fazer a busca dos dados, com a REST API utilizando a Search API, que é uma parte da mesma, podemos realizar buscas por hashtags, assuntos, usuário, dentre outros. Além das palavras-chave que podem ser utilizadas na busca, também conta com operadores para auxiliar e especificar melhor a busca a ser realizada, como por exemplo, os casos de operadores como "e" e "ou" para retornar resultados que contém uma palavra e outra, ou então no caso de ter uma ou outra. Enquanto isso, através da Streaming API podemos realizar buscas em tempo real e sem atraso na fila da REST API, o que difere esses dois métodos é que a Streaming API requer uma conexão ativa e contínua entre cliente e servidor.

Segundo Zangrandi (2014), uma das formas de utilizar a REST API constitui em criar um servidor HTTP mediador das ações entre o usuário e o Twitter. O servidor recebe as ações do usuário em seu website, analisa a requisição e então a repassa para o Twitter para realizar qualquer ação necessária para enviar a resposta para o usuário. O website então atualiza sua página e o usuário vê as informações que pediu, como segue na Figura

4:

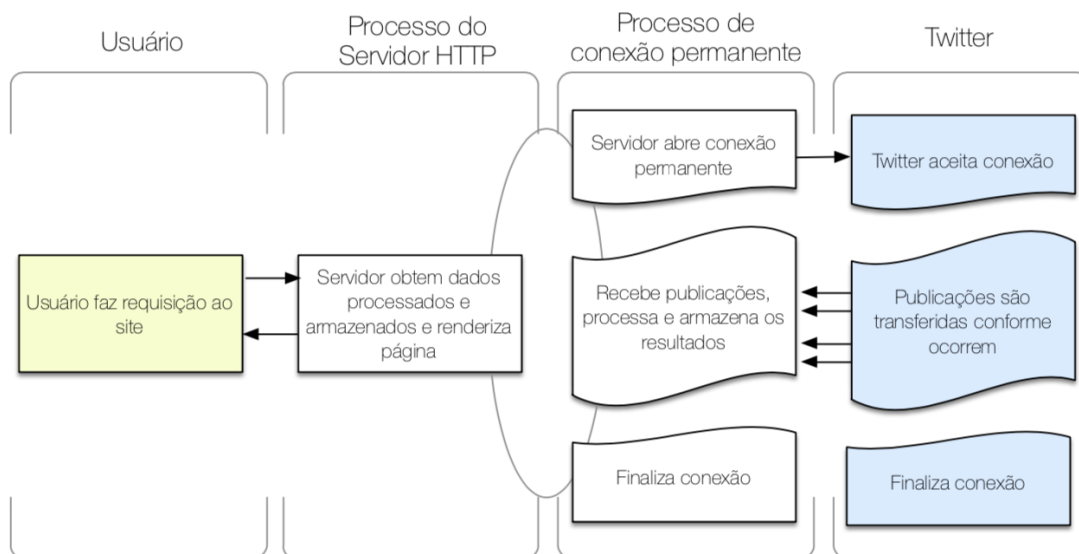
Figura 4 – Funcionamento da REST API do Twitter



Fonte: Zangrandi (2014)

Em relação a Streaming API, segundo Zangrandi (2014), o primeiro passo é solicitar uma conexão permanente (streaming) com o servidor do Twitter, que então aceita a sua solicitação e abre a conexão. O Twitter começa, então, a enviar as publicações de forma automática, conforme elas são criadas. O serviço que solicitou se encarrega de processá-las e guardá-las como for conveniente. O usuário então, ao acessar o serviço, faz a requisição e o serviço obtém os dados processados de seu banco de dados (não diretamente do Twitter) e renderiza a página para que o usuário possa visualizar os resultados, como explicitado na Figura 5

Figura 5 – Funcionamento da Streaming API do Twitter



Fonte: Zangrandi (2014)

O usuário deve estar cadastrado no Twitter para obter credenciais de acesso aos serviços que ele fornece. O usuário necessita registrar um aplicativo no site e com isso o Twitter fornece um conjunto de chaves e tokens a serem utilizados no script da aplicação. As credenciais são utilizadas para criar a conexão entre usuário e servidor de forma que nenhum login ou senha precise ser compartilhada, essa autenticação ocorre através do protocolo OAuth. As credenciais são as seguintes:

```
access_token = "Access Token"
```

```
access_token_secret = "Secret Token"
```

```
consumer_key = "API Key"
```

```
consumer_secret = "API Secret"
```

Assim que esse procedimento de credenciamento é realizado, o acesso às APIs é liberado e a troca de dados entre o Twitter e a aplicação pode ser realizado. Os métodos utilizados para realizar essa troca de informações dentro da aplicação são GET e POST.

Enquanto o método GET é utilizado para recuperar dados do servidor, o método POST é utilizado para realizar alterações no servidor do Twitter, por isso a manipulação de diversos conteúdos é feita através desses dois métodos. Os dados solicitados, são retornados em formato JSON (Javascript Object Notation) que é um formato leve e que permite que os dados possam ser transferidos na rede da internet, além de ser de fácil leitura e escrita tanto para o homem quanto para a máquina.



Para a implementação do modelo de análise perfil, pretende-se buscar por uma hashtag em específico, que é a hashtag #NowPlaying, mas como pode ser observado no Código 3.1, que mostra um exemplo de tweet buscado com a hashtag #NowPlaying, o resultado dessa busca gera uma série de informações que para o objetivo deste trabalho não serão utilizadas. Com isso grande parte do que é recebido como informação pode ser descartado, podendo assim melhorar os resultados desses tweets utilizando a estruturação dos tweets filtrando os dados que realmente importam.

Exemplo da estrutura de tweet buscado na API do Twitter

```
1 {
2   "created_at": "Mon Jul 01 14:31:49 +0000 2019",
3   "id": 1145701545624965121,
4   "id_str": "1145701545624965121",
5   "text": "#NP Live: Britney Spears - Heaven on Earth - Now @
6     https://t.co/qinNCgs5SD #NowPlaying #Music #Radio #MusicYouLove\u2026
7   "source": "\u003ca href=\"http://live.rcdradio.com\"
8   "rel=\"nofollow\" \u003eRCD Radio IRC - Now Playing App\u003c/a\u003e",
9   "truncated": true,
10  "in_reply_to_status_id": null,
11  "in_reply_to_status_id_str": null,
12  "in_reply_to_user_id": null,
13  "in_reply_to_user_id_str": null,
14  "in_reply_to_screen_name": null,
15  "user": {
16    "id": 445590277,
17    "id_str": "445590277",
18    "name": "RCD Radio - Now Play",
19    "screen_name": "RCDRadioNP",
20    "location": "United Kingdom | England",
21    "url": "http://live.rcdradio.com",
22    "description": "Live Twitter Feed of what's broadcasting on RCD Radio.
23    "translator_type": "none",
24    "protected": false,
25    "verified": false,
26    "followers_count": 570,
27    "friends_count": 3,
28    "listed_count": 162,
29    "favourites_count": 2,
30    "statuses_count": 617104,
31    "created_at": "Sat Dec 24 16:01:13 +0000 2011",
32    "utc_offset": null,
33    "time_zone": null,
34    "geo_enabled": false,
35    "lang": null,
36  },
37  "geo": null,
```

```
38 "coordinates": null,
39 "place": null,
40 "contributors": null,
41 "is_quote_status": false,
42 "extended_tweet": {
43   "full_text": "#NP Live: Britney Spears - Heaven on Earth - Now @ https://t.co/
qinNCgs5SD #NowPlaying #Music #Radio #MusicYouLove #MusicYouNeed #MusicIsLife",
44   "display_text_range": [0, 141],
45   "entities": {
46     "hashtags": [
47       { "text": "NP", "indices": [0, 3] },
48       { "text": "NowPlaying", "indices": [75, 86] },
49       { "text": "Music", "indices": [87, 93] },
50       { "text": "Radio", "indices": [94, 100] },
51       { "text": "MusicYouLove", "indices": [101, 114] },
52       { "text": "MusicYouNeed", "indices": [115, 128] },
53       { "text": "MusicIsLife", "indices": [129, 141] }
54     ],
55     "entities": {
56       "hashtags": [
57         { "text": "NP", "indices": [0, 3] },
58         { "text": "NowPlaying", "indices": [75, 86] },
59         { "text": "Music", "indices": [87, 93] },
60         { "text": "Radio", "indices": [94, 100] },
61         { "text": "MusicYouLove", "indices": [101, 114] }
62       ],
63       "timestamp_ms": "1561991509239"
64     }
65   }
```

Código 3.1 – API do Twitter

## 3.2 Tweets como fonte de dados

Conjunto de Tweets armazenados em um banco local para devidos processos com as informações afim de alcançar o objetivo. Apesar de conter poucos caracteres, quando um tweet é buscado na API, ele fornece informações relevantes e que podem ser de uso para diferentes situações dependendo do objetivo. Os dados relevantes de um tweet são:

- Data de criação
- Geolocalização
- Idioma
- Informação sobre links e hashtags

- Dados dos usuários (nome, localização, descrição e informações de perfil)
- Informação sobre o alcance da mensagem (likes, retweets e reply)

Inúmeras buscas devem ser feitas para que se tenha uma quantidade suficiente de dados para realizar o estudo de caso necessário. Os dados obtidos são salvos em formato JSON. Como o número de informações presentes em cada tweet é grande, é preciso limitar quais informações serão úteis para o desenvolvimento do trabalho proposto, neste caso os seguintes dados serão selecionados:

- Id dos Tweets
- Texto dos Tweets
- Localização dos Tweets
- Data de criação dos Tweets
- Linguagem dos Tweets
- Fonte dos Tweets (por qual meio o tweet foi enviado)

Os dados foram escolhidos com base no objetivo proposto. A data de criação e id dos tweets foram selecionadas para auxiliar na identificação de tweets duplicados de acordo com os tweets que estão sendo buscados. O texto do tweet é parte principal deste trabalho, pois nele contém as informações necessárias para realizar grande parte deste trabalho. A localização auxilia a ter melhor ideia de quais locais os tweets estão sendo enviados e com a ajuda do idioma do tweet poder criar algo mais específico para o tipo de local e idioma. A fonte dos tweets contém o dispositivo de onde o usuário envia o tweet, outro fator decisivo na personalização de conteúdo destinado ao usuário.

### 3.3 Limpeza de Dados

Grande numero de informações, algumas inúteis para o propósito do trabalho, contém cada tweet. Então, deve-se limpar todo tipo de informação que não se encaixa nos objetivos deste trabalho além de todas as informações que não serão selecionadas para dar continuidade no processo. É possível encontrar tweets compartilhados com a hashtag #NowPlaying feitos manualmente pelos usuários, mas em grande parte uma característica presente nesse tipo de tweet é que são feitos diretamente de uma plataforma de streaming de música, o que faz com que um dos atributos dele seja ter um link redirecionando o usuário para a plataforma de streaming específica onde a música pode ser ouvida, esses

links não são úteis para o trabalho que estamos propondo, então todos os termos que começam com "http://" serão descartados.

Os tweets mencionados, com a devida hashtag, normalmente seguem um padrão de estrutura independente se o usuário fez algum comentário ou inseriu algum caractere ou emoji. Isso torna mais fácil visualizar como o texto do tweet vai ser tratado para obter o fim desejado. A Tabela 1 mostra, como exemplos, como esses tweets são estruturados.

Tabela 1 – Estrutura de Tweets com a hashtag #NowPlaying

Tweets
#NowPlaying Bohemian Rhapsody by Queen
#NowPlaying Nickelback - Photograph
#NowPlaying Vestígios de Jorge e Mateus
James Brown - Get up on - youtube #NowPlaying
Evanescence / Going Under #NowPlaying

Fonte: Autor

Nessa tabela 1, os tweets apresentam estruturas semelhantes, normalmente contendo o nome do artista e a música separados por alguma palavra ou símbolo.

Seguindo o processo, com os dados filtrados, ou seja, contendo apenas as informações necessárias que seriam: a identificação do Tweet, o texto do Tweet, fonte do tweet, linguagem, a localização e o horário de criação, esses tweets passarão por processos de mineração de textos utilizando técnicas de agrupamento e classificação baseadas em um banco de dados com características musicais que foram previamente definidas e serão comparadas com as mensagens que foram filtradas, o que geraria assim dados agrupados e classificados de acordo com o gênero musical. Com a realização dessa etapa de mineração, espera-se encontrar padrões entre essas mensagens, o que por consequência geraria uma análise visual desses resultados.

### 3.4 Tokenização

Quando se trabalha com textos é preciso identificar as palavras presentes de forma separadas. Através delas será realizada a transformação dos nomes dos artistas em tokens que serão convertidos em termos para criar uma lista dos artistas presentes nos tweets. Vale ressaltar que os artistas que possuem nome composto devem ter seus nomes reconhecidos como apenas uma palavra. Outro fator relevante nessa fase de tokenização é que os termos devem ser normalizados, ou seja, independente do tipo de grafia, se escrito maiúsculo ou minúsculo, eles devem se comportar como um único termo independente desse tipo de diferença.

Os tokens são definidos como cada palavra de uma sentença, podendo conter caracteres especiais, emojis e etc. Um dos exemplos de tokenização baseado no texto de um tweet com estrutura da hashtag NowPlaying, como apresentado na Tabela 1, é como mostra a Figura 6.

Figura 6 – Tokenização de Tweets utilizando NLTK



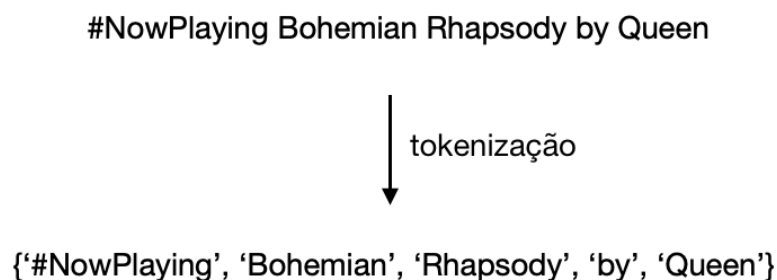
Fonte: Autor

O mesmo aconteceria caso tivesse outro tipo de delimitador entre a música e o nome do artista, como mostra na Tabela 1. Sendo assim, as palavras como "de" ou os símbolos "/" e "-" entrariam no mesmo esquema de tokenização.

Como visto no exemplo de tokenização apresentado na Figura 6, a hashtag é considerada um token separado da palavra que a prossegue; isso acaba sendo um problema quando se está trabalhando com dados do Twitter, pois é necessário que o termo seja apresentado apenas como um token e não como dois. O mesmo ocorre quando em algum tweet é encontrada alguma menção a usuários. No Twitter quando um usuário é mencionado, é utilizado o símbolo de "@" para anteceder essa menção, o que vem depois do @ é então o nome do usuário. Sendo assim, ocorreria o mesmo problema da hashtag e essa menção ficaria separada gerando assim dois tokens. Esse é um dos principais problemas encontrados ao aplicar a tokenização em dados do Twitter, pois mencionar usuários e utilizar hashtags é algo comum, por isso, com a utilização de NLTK tem-se um pacote chamado TweetTokenizer, que é bastante útil para lidar com casos como esses.

O TweetTokenizer foi criado justamente para resolver esse tipo de problema, através dele pode-se identificar as hashtags, menções de usuários e até mesmo emojis. Com isso a tokenização apresenta uma saída diferente do que foi previamente mostrado.

Figura 7 – Tokenização de Tweets utilizando NLTK



Fonte: Autor

Outro fator interessante sobre a tokenização é o reconhecimento de nomes, companhias e afins. Conhecido como Reconhecimento de Entidades Nomeadas (NER), sigla que em inglês significa Named-Entity Recognition, esse passo contribui de maneira imprescindível para auxiliar na criação de uma lista de artistas presentes nos tweets que foram buscados.

### 3.5 Banco de características musicais

As características musicais dos artistas que estão presentes nos tweets buscados deste trabalho e a forma como essas informações serão classificadas afim de poder trabalhar com os dados coletados de maneira completa. As características podem ser encontradas de diversas maneiras, um dos fatores a ser levado em conta é de que os serviços de streaming que estão cada vez mais populares e têm dominado o mercado fonográfico, vêm se consolidando como uma das principais formas de se consumir música no mundo, o que mostra o quanto a era digital tem crescido e evoluído. Essas plataformas têm facilitado a maneira de como os dados têm sido tratados e a forma de acessá-los.

Com os dados significativos, deixados pelo processo de filtro, cabe intuir que haverá uma lista de cantores ou bandas, com esses dados será então necessário criar comparação com o banco de dados que contém características sobre o estilo musical do artista em questão, descobrindo então qual gênero por exemplo eles fazem parte. O Spotify, que é um dos maiores aplicativos nesse meio de streaming, disponibiliza a API para trabalhar com dados de cantores, álbuns, gêneros musicais e etc. Por isso, esses dados serão utilizados para comparar com os dados obtidos do Twitter, podendo assim ter informações sobre os gêneros musicais e em quais categorias os mesmos se encontram.

Em serviços como esses, pode-se encontrar uma lista completa de informações sobre os artistas, podendo ter acesso as suas discografias, história do artista, gênero musical e

mais algumas informações que formam a caracterização musical como visto na Figura 3. Esse banco com as características é essencial para agrupar esses artistas baseado no gênero musical que os mesmos compartilham, esse processo é realizado através da clusterização

## 3.6 Clusterização

Uma das técnicas a ser considerada, que pode se utilizar neste trabalho é a clusterização, que segundo (BEZERRA, 2010), é utilizada para separar os documentos de uma base de textos em subconjuntos ou clusters, de tal forma que os documentos de um cluster compartilhem de propriedades comuns que os distingam de documentos em outros clusters. O objetivo nesta tarefa é maximizar similaridade intracluster e minimizar similaridade intercluster. Diferente da tarefa de classificação, que tem rótulos pré-definidos, a clusterização precisa automaticamente identificar os grupos de documentos aos quais o usuário deverá atribuir rótulos.

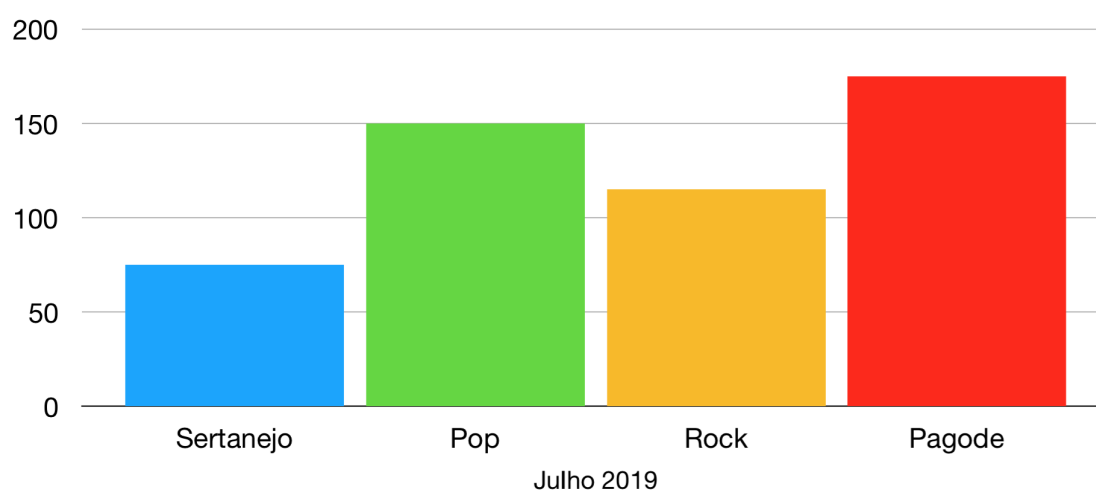
Através da clusterização será possível agrupar artistas baseados nas características musicais que os mesmos apresentam, como por exemplo o estilo musical de cada um deles, formando assim clusters. Esses clusters serão criados de forma iterativa e estarão em constante mudança, pois se em uma nova busca de dados são encontradas características similares às características presentes nos clusters existentes, esse dados serão então integrados a esse cluster.

## 3.7 Análise Visual

A visualização dos dados é concebida através da exposição dos mesmos em forma de gráfico, o objetivo de visualizar os dados dessa forma é para trazer a informação ao usuário da maneira mais clara e objetiva possível, promovendo a compreensão de forma intuitiva e para que até mesmo usuários que não tenham conhecimentos técnicos aprofundados sobre o assunto, possam compreender e absorver o que foi representado de forma gráfica.

No caso deste trabalho, os resultados serão apresentados através de gráficos contendo as informações que foram previamente estabelecidas como critério de análise e que foram transformadas em um produto final baseado no que o trabalho propõe. Um dos exemplos de resultado e visualização desses dados seria um gráfico de gêneros musicais mais escutados em um determinado período de tempo baseado na quantidade de Tweets daquele mês. Além de existir a possibilidade de realizar comparação com outros meses do ano e apresentar os diferentes gêneros musicais e cantores mais ouvidos do ano em questão, ou seja, os que apareceram mais vezes nos tweets que foram buscados. A Figura 8 ilustra um exemplo de barras de frequência, baseado na estatística, de gêneros musicais.

Figura 8 – Gráfico de gêneros musicais mais escutados baseado na quantidade de Tweets



Fonte: Autor



## 4 Modelo implementado

A arquitetura ilustrada pela Figura 3 é uma referencia lógica para a implementação do modelo desejado. Ela está composta por módulos procedurais, informações de conhecimentos e relações de fluxos. Os módulos são: obtenção de tweets, pré-processamento, tokenização, caracterização, classificação, interpretação e visualizações. As informações de conhecimentos são os bancos de dados dos tweets, repositórios locais, bancos de atributos e banco de padrões.

As informações são obtidas e colocadas no arquivo local. Essas informações devem ser preparadas (purificadas e padronizadas) para a identificação dos elementos a serem tratadas - os tokens e colocadas como entidades representativas- com operações numéricas na classificação. São classificadas as informações em função dos atributos musicais de referencia, consideradas como válidas. Os resultados dessa classificação são registrados como tipos musicais identificados (padrões). Esses padrões são analisados por tendências, frequências e localidades para sua visualização.

As informações são obtidas dos bancos de dados de Twitter, por tanto são respeitadas as regras exigidas este. Também são usadas as ferramentas disponibilizadas para desenvolvedores com os elementos de Twitter.

### 4.1 Twitter e exigências

Para que o processo de análise musical tenha início, o primeiro passo é registrar a aplicação na área de desenvolvedor do Twitter, criando um nome único e obtendo assim as credenciais necessárias para acesso aos dados do Twitter. Essas credenciais são geradas pelo próprio Twitter juntamente com um id único para a aplicação registrada. O Código 4.1 ilustra as variáveis com as credenciais de autorização geradas pela API do Twitter.

Exemplo de credenciamento na API do Twitter

```
1 import tweepy
2 from tweepy.streaming import StreamListener
3 from tweepy import OAuthHandler
4 from tweepy import Stream
5
6 # Variaveis contendo as credenciais de autorizacao a API do Twitter
7 access_token = "30966760-10kJVJs8zk1K0qXKHPJDorSzfjykvN1PJFZ4M8z0I"
8 access_token_secret = "EiFg65hJEPYLPRD5nyVPomZ2LAH0cbHMGWgZW07qBHWi5"
9 consumer_key = "JKx8yzc75VMqDdJqyqlSLqAFA"
```

```
10 consumer_secret = "ANKPAfPOabW9YMNRXhJ8X5EEjvtJK3kA1enzDpPL0jlAZKefpi"
```

#### Código 4.1 – Credenciamento Twitter

## 4.2 Obtenção de Tweets

Com as credenciais corretas, são realizadas as buscas de tweets com a palavra-chave `#NowPlaying`. Hashtag escolhida para a obtenção dos dados, pois é a palavra-chave presente em grande parte dos tweets de usuários que compartilham o que estão ouvindo no Twitter. Esses tweets podem ser feitos "a mão"; ou seja, com o próprio usuário digitando a música que deseja ou então compartilhando através de alguma plataforma de streaming musical. Essas mensagens são importantes para iniciar o banco de dados que será preenchido com as informações que serão retornadas.

Como o Twitter limita bastante as buscas através da REST API e Search API. Por exemplo, limita receber tweets com mais de uma semana e também por retornar um número limitado de Tweets. Nesse sentido tem que realizar várias requisições. O Streaming API do Twitter é mais apropriado nesse caso, pois obtém esses dados requeridos e o que estão sendo postados no momento em que a busca está sendo realizada. Essas buscas foram realizadas em duas semanas diferentes, a primeira leva de tweets foi obtida entre os dias 29 e 31 de outubro de 2019 e a segunda leva de 4 e 6 de novembro, com um total de 56.215 tweets. As publicações obtidas através da streaming API foram salvas localmente em arquivos JSON.

## Seleção

Os tweets salvos em formato JSON, como mostrado no Código 4.2, contém uma série de informações inúteis para o propósito do trabalho. Devem ser purificados realizando uma limpeza buscando apenas pelos campos que foram definidos como necessários para a implementação deste trabalho.

#### Exemplo de tweet salvo em arquivo JSON

```
1 {"created_at":"Thu Nov 07 17:52:10 +0000 2019","id":1192500001471180800,"id_str":"1192500001471180800","text":"#NowPlaying Aretha Franklin - Bridge Over Troubled Water https://t.co/OZ1fr61pIz #KWAYDB","source":"\u003ca href=\"https://fastcast4u.com\" rel=\"nofollow\" \u003eTwitterCast by FastCast4u\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":800113308107821056,"id_str":"800113308107821056","name":"KWAY-DB","screen_name":"KwayDb","location":"Oklahoma City, OK","url":"http://www.kwaygospel.com/","description":"KWAY-DB Internet Radio Station is God's Way! Good Ole Home Town Gospel Music 24/7","
```

```

translator_type":"none","protected":false,"verified":false,"followers_count"
:308,"friends_count":331,"listed_count":1,"favourites_count":154,"
statuses_count":107780,"created_at":"Sat Nov 19 23:07:32 +0000 2016","
utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":null,"
contributors_enabled":false,"is_translator":false,"profile_background_color":"
000000","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":"FF691F","profile_sidebar_border_color":"000000","
profile_sidebar_fill_color":"000000","profile_text_color":"000000","
profile_use_background_image":false,"profile_image_url":"http://pbs.twimg.com
/profile_images/800121293139939328/06B9YEwj_normal.jpg","
profile_image_url_https":"https://pbs.twimg.com/profile_images
/800121293139939328/06B9YEwj_normal.jpg","profile_banner_url":"https://pbs.
twimg.com/profile_banners/800113308107821056/1500825636","default_profile":
false,"default_profile_image":false,"following":null,"follow_request_sent":null
,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors
":null,"is_quote_status":false,"quote_count":0,"reply_count":0,"retweet_count"
:0,"favorite_count":0,"entities":{"hashtags":[{"text":"NowPlaying","indices"
:[0,11]},{"text":"KWAYDB","indices":[81,88]}],"urls":[{"url":"https://t.co/0
Z1fr61pIz","expanded_url":"http://kwaygospel.com/","display_url":"kwaygospel
.com","indices":[57,80]}],"user_mentions":[],"symbols":[]},"favorited":false,"
retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en","
timestamp_ms":"1573149130586"}

```

Código 4.2 – Tweet salvo no formato JSON

Seis campos foram escolhidos, como mostra o Código 4.3, para serem trabalhados, entre eles teremos: data de criação, id dos tweets, texto, fonte dos tweets, linguagem e localização. Uma forma rápida que se encontrou de obter esses campos foi utilizando o método "map" de Javascript para localizar essas informações dentro do JSON e assim extraí-las.

Exemplo do código para buscar os campos dentro do JSON

```

1 const array = require("../UENF/Monografia/Code/novALL.json");
2 const routes = express.Router();
3
4 routes.get("/filter", (req, res) => {
5   const newArray = array.map(item => {
6     const newObject = {
7       created_at: item.created_at,
8       id: item.id,
9       text: item.text,
10      source: item.source,
11      lang: item.lang,
12      location: item.user.location
13    };

```

```
14
15     return newObject;
16   });
17
18   return res.json(newArray);
19 });
20 module.exports = routes;
```

Código 4.3 – Código de busca dos campos no JSON

Os dados selecionados para o andamento do trabalho, são salvos e estruturados de maneira que fiquem organizados e de forma que seja fácil o entendimento para qualquer pessoa. O Código 4.4 mostra dois exemplos de como os dados são apresentados depois de estarem forma limpa e organizada.

Tweets estruturados após extração dos campos a serem utilizados

```
1 {
2   "created_at": "Wed Nov 06 21:57:43 +0000 2019",
3   "id": 1192199407426588700,
4   "text": "Run the World (Girls) - Homecoming Live by Beyonc #NowPlaying",
5   "source": "https://mobile.twitter.com Twitter Web App",
6   "lang": "en",
7   "location": "United States"
8 },
9 {
10  "created_at": "Wed Nov 06 21:57:43 +0000 2019",
11  "id": 1192199408588329000,
12  "text": "#nowplaying: Towkio - Forever feat. Vic Mensa",
13  "source": "https://radio.co Radio.co now playing",
14  "lang": "da",
15  "location": null
16 },
```

Código 4.4 – JSON organizado

## Extração

Com os campos selecionados, fica mais fácil de trabalhar com os dados desejados pois muitas informações já foram descartadas. O próximo passo é focar no campo "text", que é basicamente o tweet em si, pois será através desse campo que algumas informações serão obtidas e transformadas em um resultado final após aplicação de mineração de textos.

Os dados do formato JSON devem ser transformados para formato CSV para permitir a disposicao dos dados por tipos de colunas, o que facilita a visualização das in-

formações existentes. Um exemplo de dados JSON colocados em formato CSV é ilustrada pela Tabela 2.

Tabela 2 – Tabelas de Tweets no arquivo CSV

created_at	id	text	source	lang	location
<b>Tue Oct 29 02:28:23 +0000 2019</b>	1189006030794973200	#NowPlaying Coldplay - A Sky Full of Stars  https://t.co/gPoWhWjMQb	TweetDeck	en	Indonesia
<b>Tue Oct 29 02:28:28 +0000 2019</b>	1189006054908158000	#NowPlaying: "LOYALTY. feat. Rihanna"by Kendrick Lamar , DAMN. (2017). https://t.co/whNzHE4Bk4	Now Playing	en	Suckerville, USA
<b>Tue Oct 29 02:33:12 +0000 2019</b>	1189007244081401900	#NowPlaying Circles by Post Malone #listen at https://t.co/RGN0Gpaxia	@101kdrs	en	null

Fonte: Autor

### 4.3 Pré-processamento de Dados

O trabalho com os tweets começa a ser feito diretamente com a coluna text do arquivo CSV. Para que os dados estejam prontos para serem tratados, deve-se realizar alguns processos que precisam ser realizados de formas individuais, ou seja, em várias etapas, entre elas:

- Remoção de URLs, pontuação e stop-words
- Reconhecimento de entidades

#### Remoção de URLs, pontuação e stop-words

A remoção dos URLs é necessária porque, neste caso, não representa um texto significativo dentro do contexto do que quer dizer a mensagem textual. Os URLs sempre

iniciam com "http". Os símbolos especiais, como pontuações, dos textos também são elementos excedentes dos textos, a não ser que se considere a parte de significados, o que não é o caso neste trabalho. Portanto, também devem ser removidos. Na Figura 9 temos exemplos do que são considerados como pontuação segundo normas. O comando "punctuation" de "string" em python permite detectar esses elementos em um texto.

Figura 9 – Exemplos de pontuação

```
[1]: import string
      string.punctuation

[1]: '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

Fonte: Autor

Os elementos numéricos são também considerados não representativos no contexto de textos, que também devem ser removidos. A remoção de pontuação e números que será executada através do código apresentado na Figura 10.

Figura 10 – Remoção de pontuação e números

```
In [10]: def remove_punct(text):
          text = "".join([char for char in text if char not in string.punctuation])
          text = re.sub('[0-9]+', '', text)
          return text

          df['Tweet_punct'] = df['Tweet'].apply(lambda x: remove_punct(x))
          df.head(10)
```

Fonte: Autor

A partir dessa primeira limpeza feita nos tweets removendo a pontuação, números e URLs teremos os dados apresentados de forma um pouco diferente da que foi originalmente apresentada, como podemos ver na Tabela 3.

Tabela 3 – Primeira limpeza dos textos

id	text	tweet sem pontuação, números e URL
1189006030794973200	#NowPlaying Coldplay - A Sky Full of Stars <a href="https://t.co/gPoWhWjMQb">https://t.co/gPoWhWjMQb</a>	NowPlaying Coldplay A Sky Full of Stars
1189006054908158000	#NowPlaying: "LOYALTY. feat. Rihanna"by Kendrick Lamar , DAMN. (2017) <a href="https://t.co/whNzHE4Bk4">https://t.co/whNzHE4Bk4</a>	NowPlaying Loyalty feat Rihanna by Kendrick Lamar DAMN
1189007244081401900	#NowPlaying Circles by Post Malone \#listen at <a href="https://t.co/RGN0Gpaxia">https://t.co/RGN0Gpaxia</a>	NowPlaying Circles by Post Malone listen at

Fonte: Autor

As stop-words, como exemplificado na Figura 11, são palavras que aparecem com frequência e que podem ser ignoradas, pois não fazem tanta diferença quando se trata do contexto inteiro da frase. A biblioteca NLTK do Python tem uma lista de stop-words em diferentes línguas e com isso, facilita a remoção desse tipo de palavra utilizando uma simples linha de código.

Figura 11 – Exemplo de stop-words em inglês e português

{ a, about, across, after, all, also, an, and, any, are, as, at, be, but, by, can, do, does, either, else, ever, every, for, from, get, he, her, hers, him, his, how, i, if, in, into, is, it, just, like, may, me, might, most, must, my, no, nor, not, of, off, often, on, only, or, rather, said, say, she, so, some, than, that, the, them, then, there, they, this, tis, to, too, we, were, what, when, where, which, while, who, whom, why, will, with, yet, you, your},

{a, alguém, algum, amplo, antes, ao, aos, após, aquela, as, até, através, cada, coisa, coisas, com, como, contra, contudo, da, daquele, daqueles, das, de, dela, delas, dele, deles, depois, dessa, dessas, este, estes, estou, eu, fazendo, fazer, feita, feitas, feito, feitos, foi, for, foram, fosse, fossem, grande, grandes, há, isso, isto, já, la, lá, lhe, lhes, lo, mas, me, meu, pelo, pequena, pequenas, pequeno, pequenos, per, perante, pode, quais, qual, quando, quanto, que, quem, são, se, seja, sejam, seu, só, sob, sobre, sua, talvez, também, tem, ter, teu, teus, ti, tinha, toda, tu, última, um, ver, vez, vós },

Fonte: Autor

Os exemplos de remoção de stop-words podem ser encontrados na Tabela 4. No caso dos tweets que estão sendo utilizados como exemplos, as stop-words removidas foram: a, of, by e at.

## Reconhecimento de entidades

Após as etapas realizadas, faz-se necessário trabalhar com o reconhecimento de entidades. Processo esse que foi possível utilizando o software, que faz uso de processamento de linguagem natural, criado por Stanford e que foi disponibilizado para ser utilizado em diversas linguagens de programação diferentes, entre elas o Python através de NLTK. Além disso, foi utilizado o reconhecimento de pessoas presente no site Monkey Learn, onde é possível realizar buscas e ao submeter o texto para análise, ele retorna os dados em formato de lista ou formato JSON. Este segundo tem informações sobre quantas vezes o nome apareceu no texto. O texto submetido contém apenas os dados do campo text e por já ter sido tratado anteriormente de formas para que o mesmo se encontre sem hashtags, URL e etc, então as informações estão limpas e todo o texto foi contado como fator único.

Foi preciso realizar as buscas de formas diferentes para que tivesse melhor aproveitamento possível dos nomes dos artistas e para que se perdesse o mínimo possível entre os tweets pois através dos testes realizados em ambos sites, foi percebido que os programas de reconhecimento de entidades não reconheceram certos nomes.

Uma versão online do software pode ser encontrada no próprio site de Stanford (citar o site aqui <https://nlp.stanford.edu/software/CRF-NER.shtml>) e o mesmo acontece com o site Monkey Learn (citar o site <https://app.monkeylearn.com/main/dashboard/>). Alguns testes foram realizados antes de decidir seguir com os softwares para a realização deste trabalho.



Figura 12 – Teste na versão online do software de Stanford

**Stanford Named Entity Tagger**

Classifier: english.conll.4class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

```
'NowPlaying','Loyalty','feat','Rihanna','Kendrick','Lamar','DAMN']
['NowPlaying','Circles','Post','Malone','listen']
['NowPlaying','Coldplay','Sky','Full','Stars']
```

Submit Clear

'NowPlaying','Loyalty','feat','**Rihanna**','**Kendrick**','**Lamar**','DAMN'] ['NowPlaying','Circles','Post','**Malone**','listen'] ['NowPlaying','**Coldplay**','Sky','Full','Stars']

Potential tags:

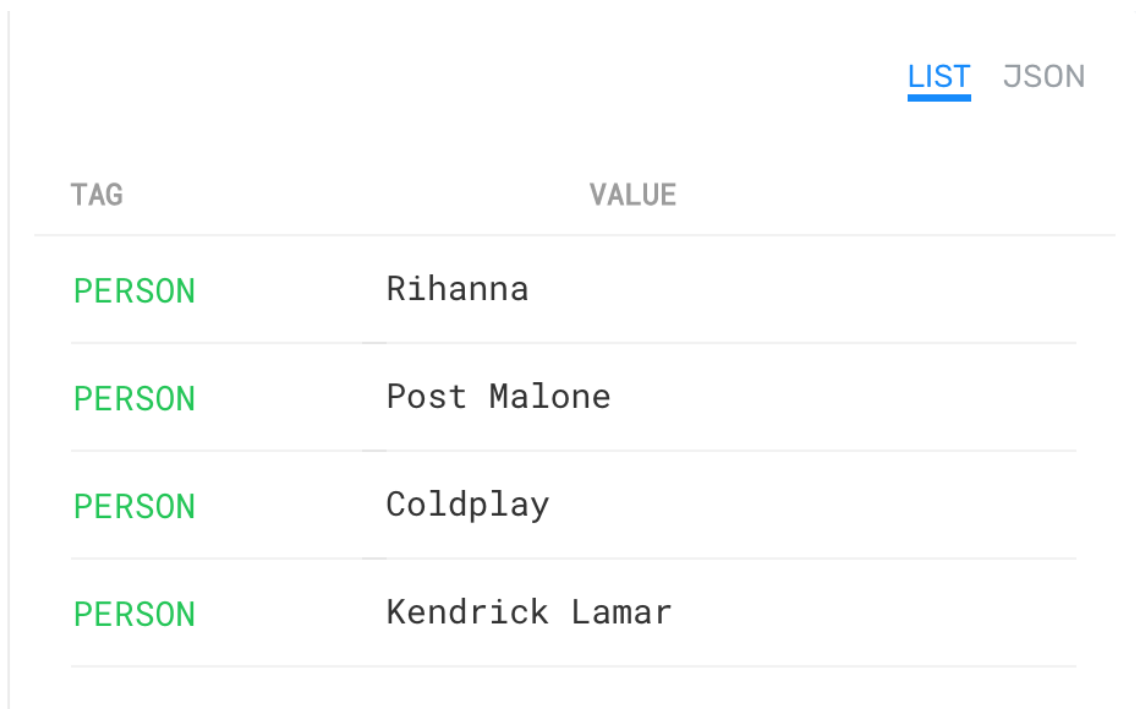
**ORGANIZATION**  
**LOCATION**  
**PERSON**  
**MISC**

Fonte: <https://nlp.stanford.edu/software/CRF-NER.shtml>

Utilizando os tweets que foi mostrado na primeira limpeza exemplificada na tabela 3 e sendo tokenizado como foi mostrado na tabela 4, o resultado do software de Stanford em relação aos nomes de entidades pode ser visto na Figura 12. Ele reconheceu os nomes Rihanna, Kendrick, Lamar, Malone e Coldplay como nomes de pessoas. Apesar de não reconhecer nomes compostos como foi o caso de Kendrick Lamar, e isso se dá por motivos de cada palavra estar sendo vista como única por causa da tokenização, essa foi ainda a melhor forma de reconhecer os nomes dos artistas. O mesmo teste foi executado antes da tokenização e alguns nomes como o da banda Coldplay não foram reconhecidos, o que não aconteceu após a tokenização, por esse motivo decidimos seguir com o reconhecimento de entidades realizado com os tweets tokenizados.

Os mesmos testes foram realizados utilizando o Monkey Learn e como exemplificado na Figura 13, podemos ver que o cantor Post Malone teve seu nome devidamente reconhecido, diferente do acontecido com o software de Stanford.

Figura 13 – Teste na versão online do software do site Monkey Learn



TAG	VALUE
PERSON	Rihanna
PERSON	Post Malone
PERSON	Coldplay
PERSON	Kendrick Lamar

Fonte: <https://app.monkeylearn.com/main/extractors/ex<sub>5mwSdZ3C</sub>/>

Os dois apresentam semelhanças, mas se tratando de algumas palavras eles identificam de formas diferentes. O software de Stanford identifica também localidades e organizações, alguns nomes de banda foram identificados como organização e no Monkey Learn foi reconhecido como pessoa.

Em contrapartida, apesar de ter sido testado em ambos softwares, alguns nomes não foram identificados e assim descartados, mesmo assim foi possível ter um grande aproveitamento de dados. Alguns dos artistas que contém nome composto, como por exemplo, a banda The White Stripes foi identificada como organização e não como pessoa pelo software de Stanford. Vale ressaltar que após a tokenização a palavra "the" que dá início ao nome da banda não está mais presente, pois é uma stop-word e foi eliminada. Como é de conhecimento algumas dessas bandas, após analisar os dados que foram retornados, foi preciso fazer uma análise mais próxima para checar casos como esses.

## 4.4 Tokenização

Através da tokenização é possível identificar cada termo presente nos tweets de maneira correta, sem afetar o fluxo de informações, pois muitas vezes os tweets contam, por exemplo, com hashtags e emojis. A tokenização separa cada palavra como um termo individual e como não temos mais símbolos e pontuação, o trabalho é simplificado.

Um exemplo de tokenização utilizando a biblioteca do Python chamada NLTK a

estrutura de um tweet contendo a hashtag #NowPlaying pode ser encontrada na Tabela 4.

Tabela 4 – tweets tokenizados e sem stop-words

text	tweet sem pontuação e URL	tweet tokenizado e sem stop-words
\#NowPlaying Coldplay - A Sky Full of Stars <a href="https://t.co/gPoWhWjMQb">https://t.co/gPoWhWjMQb</a>	NowPlaying Coldplay A Sky Full of Stars	['NowPlaying', 'Coldplay', 'Sky', 'Full', 'Stars']
#NowPlaying: "LOYALTY. feat. Rihanna"by Kendrick Lamar , DAMN. (2017) <a href="https://t.co/whNzHE4Bk4">https://t.co/whNzHE4Bk4</a>	NowPlaying Loyalty feat Rihanna by Kendrick Lamar DAMN	['NowPlaying', 'Loyalty', 'feat', 'Rihanna', 'Kendrick', 'Lamar', 'DAMN']
#NowPlaying Circles by Post Malone \#listen at <a href="https://t.co/RGN0Gpaxia">https://t.co/RGN0Gpaxia</a>	NowPlaying Circles by Post Malone listen at	['NowPlaying', 'Circles', 'Post', 'Malone', 'listen']

Fonte: Autor

## 4.5 Caracterização

As músicas se encaixam em um nicho que é estabelecido de acordo com os instrumentos e ritmo de uma música. Segundo o site Every Noise, (citar aqui <http://everynoise.com/everynoise1d.cgi?scope=all>) existem 3756 gêneros musicais registrados até hoje. Uma música pode conter gêneros diferentes, assim como o artista. O artista, normalmente tem um gênero musical principal, mas acaba se encaixando em outros gêneros devido ao estilo da música que o mesmo criou. Além disso, dentro dos próprios gêneros principais, existem variações de nichos e que podem incluir ou não o artista ou música. Um exemplo sobre isso é o gênero musical Pop e que tem uma das variações como Pop rock.

A caracterização tanto do artista quanto da música se dá por padrões que são seguidos quanto a esse tipo de quesito. Então esse processo não é algo que precisa ser implementado, pois todas essas informações podem ser buscadas em outros lugares como Spotify e Lastfm.

## 4.6 Classificação

A classificação tem por objetivo categorizar um item dentro de uma ou mais classes. No caso deste trabalho ela foi utilizada, pois foram definidos gêneros musicais para serem trabalhados e o objetivo foi classificar os artistas perante os gêneros indicados. Assim, os artistas foram classificados entre os gêneros: pop, rock, rap, hip-hop, reggae e nacional. Como os artistas já têm seu gênero musical atribuído, a classificação fica mais fácil de ser realizada. Com o conhecimento do gênero musical do artista, então é preciso realizar uma contagem para descobrir quais artistas mais apareceram nos tweets e conseqüentemente os gêneros foram mais populares, etc.

Como os algoritmos de aprendizado de máquina não conseguem lidar com textos, é preciso convertê-los em números. Se tratando do processo de linguagem natural, uma técnica utilizada para tal feito é ter todas as palavras que aparecem no texto em uma espécie de "bolsa". O nome desse tipo de técnica é bag of words.

Através de bag of words, criamos um dicionário com os termos que foram extraídos através dos sites de Stanford e Monkey Learn. Filtrando o processo mais ainda, foi criada uma tabela com uns artistas de alguns estilos musicais e com essa lista, fizemos essa contagem dentro dos tweets. Basicamente, ao percorrer o arquivo com os tweets, cada vez que o termo aparece, é acrescentado 1 ao termo e caso não apareça, acrescenta 0. Com isso, tem-se a quantidade de vezes que os artistas apareceram nos tweets e quais foram os mais citados.

### 4.6.1 Clusterização

teoria e implementação

## 4.7 Atributos e padrões

Os artistas possuem atributos que são comuns a outros artistas, o que no caso deste trabalho, se dá pelo cluster em que o mesmo se encontra baseado no gênero musical em que o mesmo faz parte. Existem diversas formas de ter acesso ao gênero musical de um artista, com as plataformas de streaming tem-se de maneira mais fácil em qual gênero o artista se encaixa, pois ao acessar a página do mesmo podemos encontrar hashtags que o identificam em relação a isso.

Outra forma foi utilizar o Lastfm que é um serviço de rádio onde os scrobbles das músicas são enviadas para a plataforma e ficam registrados em seu perfil. O Lastfm conta com diferentes tipos de abastecimento de dados, podendo receber streaming do Spotify ou do próprio site deles. Assim como o Spotify, os artistas tem suas próprias páginas com uma breve biografia e suas características. Além disso, o Lastfm conta com páginas

específicas voltadas justamente para os gêneros musicais. São páginas individuais e que ao acessar o gênero desejado, é retornado uma lista de artistas que fazem parte dele.

Ambas plataformas foram utilizadas para pesquisar e abastecer um banco de dados próprio com características musicais contendo gêneros musicais. Com base nas informações cruzadas pelos dois sites, foi criado um banco de informações atribuindo os artistas a seu gênero musical. Inicialmente foi pensado em utilizar diretamente a api das plataformas citadas, mas por motivos de limitações decidiu-se criar esse banco de informações local para que essa etapa seja executada.

#### 4.7.1 Características Musicais

As características musicais fazem com que seja possível identificar a qual gênero uma música ou artista pertencem. Elas são identificadas através dos instrumentos utilizados nas músicas ou a forma como a música é feita. Essas características são importantes pois segmentam tanto os artistas quanto o público consumidor do tipo de música específico.

#### 4.7.2 Padrões

A geração de padrões busca informações de dados até então desconhecidos. Sem nenhuma informação prévia conhecida dos dados e baseado no objetivo desejado, é possível ter conhecimento sobre os dados que estão sendo trabalhados.

Os padrões são buscados com a intenção de trazer conteúdo novo, útil e que possa ser lido por qualquer tipo de usuário. Mas caso os padrões encontrados não sejam úteis, o usuário precisa ter entendimento o suficiente para decidir que o mesmo não é interessante para o objetivo estabelecido.

### 4.8 Análise Visual

Os padrões que foram gerados e descobertos são analisados, documentados e ilustrados para facilitar o processo de conhecimento sobre esses dados. Como esse produto final precisa ser feito de forma clara e que facilite o entendimento, foi escolhido então utilizar gráficos e tabelas para tal feito. Através desse conhecimento final, além de conhecer, de fato, os dados, é possível ter total compreensão do que foi definido.

Com essas informações em mãos, interpretação dos dados pode ser aplicada na tomada de decisões ou em outras definições estratégicas dependendo do segmento traçado.

## 5 Resultados

Apesar do objetivo deste trabalho ser trabalhar com o conteúdo musical extraído dos tweets, outros campos nessa busca de tweets foram utilizados para auxiliar na análise de dados geradas pelos tweets e algumas técnicas de mineração de textos aplicadas para tal finalidade.

Utilizando o software Anaconda, que é conhecido por ser voltado para o trabalho com dados, aprendizado de máquina e por ser uma ferramenta de fácil instalação, que tem distribuição gratuita e código aberto das linguagens de programação Python e R, temos resultados baseado nos dados que foram escolhidos para análise. Uma vantagem da utilização desse software é a instalação de bibliotecas complexas para trabalhar com ciência de dados e aprendizado de máquina, que são instaladas juntamente com o gerenciador de pacotes chamado conda. Após a criação do ambiente virtual, instalou-se duas ferramentas essenciais para desenvolvimento deste trabalho, que são Jupyter Notebook e o Orange. O Jupyter Notebook é uma aplicação web popular no ambiente no Data Science e permite mesclar o código e a visualização de dados. E o Orange, facilita a entrada e manipulação de dados e possui diferentes tipos de ferramentas para trabalhar com aprendizado de máquina e mineração de dados.

### Idioma dos Tweets

A coluna "lang" do arquivo CSV foi um dos atributos escolhidos para ser trabalhado e é um campo fácil de ser analisado, pois o Twitter tem suporte para 34 linguagens diferentes e a resposta deste campo é composta por uma sigla do idioma, o que não é mais que 3 caracteres. Assim, é criado um dicionário com os termos e é realizada a contagem para saber quantos vezes esses termos aparecem. Dentre os tweets armazenados no banco de dados local temos vários idiomas presentes. O inglês, que é um idioma presente em grande parte do mundo, foi predominante nos tweets, mas foi encontrado também número significativo de tweets em japonês, por exemplo. O português apareceu em nono lugar com apenas 415 tweets. Como pode ser observado na Figura 14, esse é o top 10 idiomas com a quantidade de tweets de cada um deles.

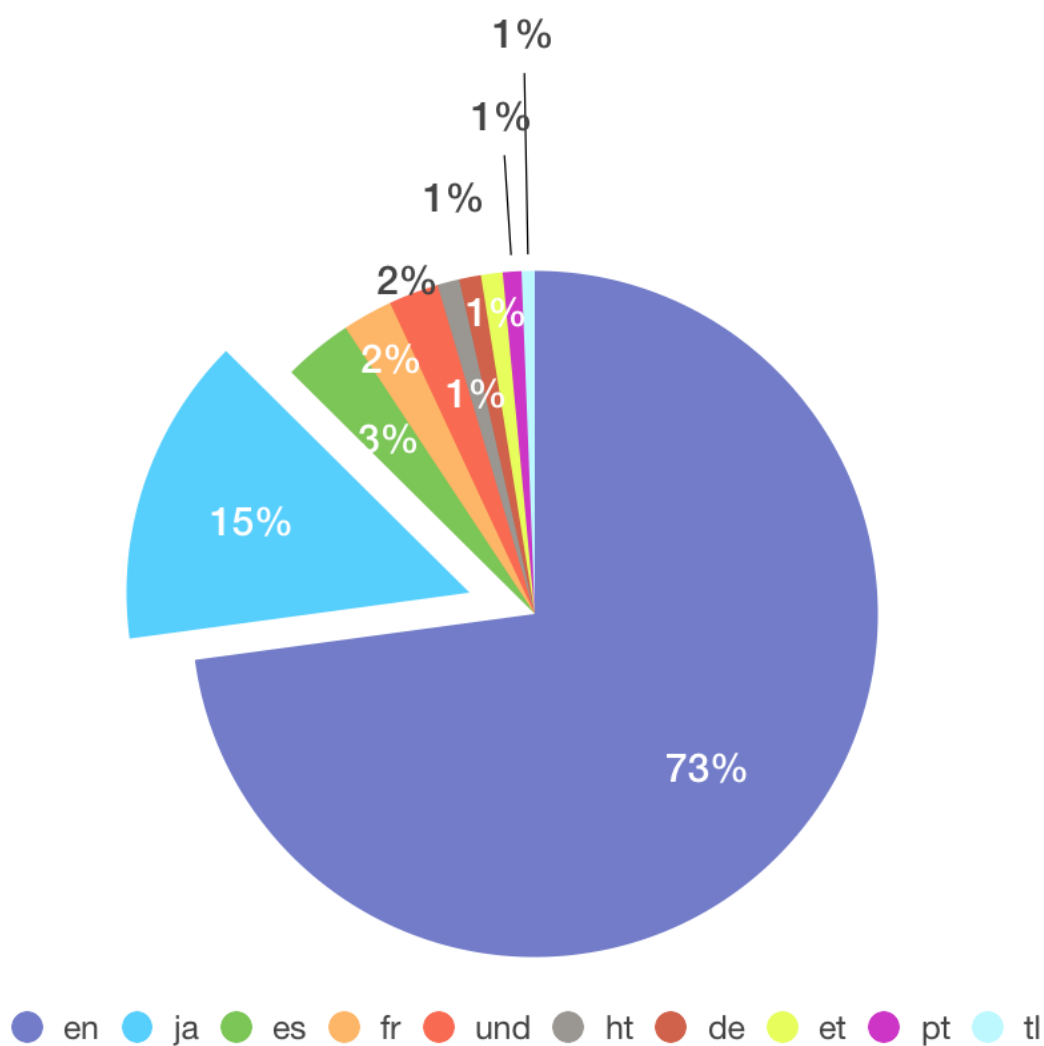
Figura 14 – Tabela com top 10 de idiomas dos tweets

Idioma dos Tweets	
IDIOMAS	QUANTIDADE DE TWEETS
en	34,558
ja	6,915
es	1,546
fr	1,123
und	1,118
ht	491
de	488
et	485
pt	415
tl	293

Fonte: Autor

Em relação ao top 10 de idiomas, o banco de dados local, que tem um total de 56.125 tweets, em termos de porcentagem os tweets representam os números apresentados no gráfico mostrado na Figura 15. O idioma inglês que recebeu 34.558 tweets teve 73% dos tweets, seguido do japonês que representa 3% dos tweets, enquanto o português aparece apenas com 1%, como vários outros.

Figura 15 – Gráfico com top 10 de idiomas dos tweets



## Fonte dos Tweets

Outro fator interessante e que foi escolhido para ser trabalhado é a fonte dos tweets, ou seja, de onde os usuários estão enviando os tweets. Apesar de nem todos os tweets apresentarem esse campo e muitos conterem "null" como resposta, obtivemos dados significativos em relação a essa fonte e os mesmos estão apresentados nas figuras 16 e 17.



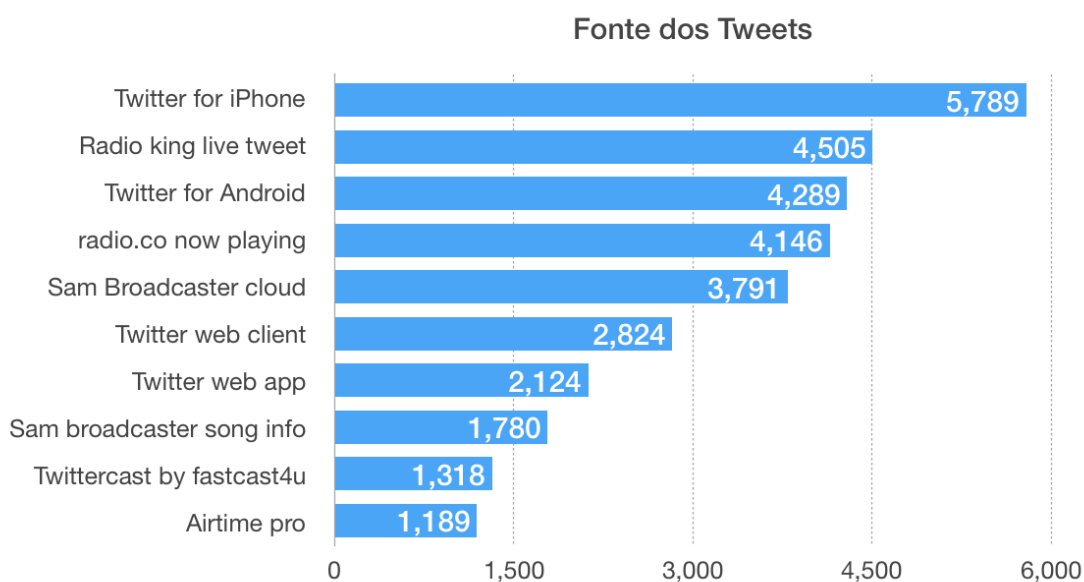
Figura 16 – Tabela com top 10 fonte dos tweets

## Fonte dos Tweets

SOURCE	TWEETS
Twitter for iPhone	5,789
Radio king live tweet	4,505
Twitter for Android	4,289
<u>radio.co</u> now playing	4,146
Sam Broadcaster cloud	3,791
Twitter web client	2,824
Twitter web app	2,124
Sam broadcaster song info	1,780
Twittercast by fastcast4u	1,318
Airtime pro	1,189

Fonte: Autor

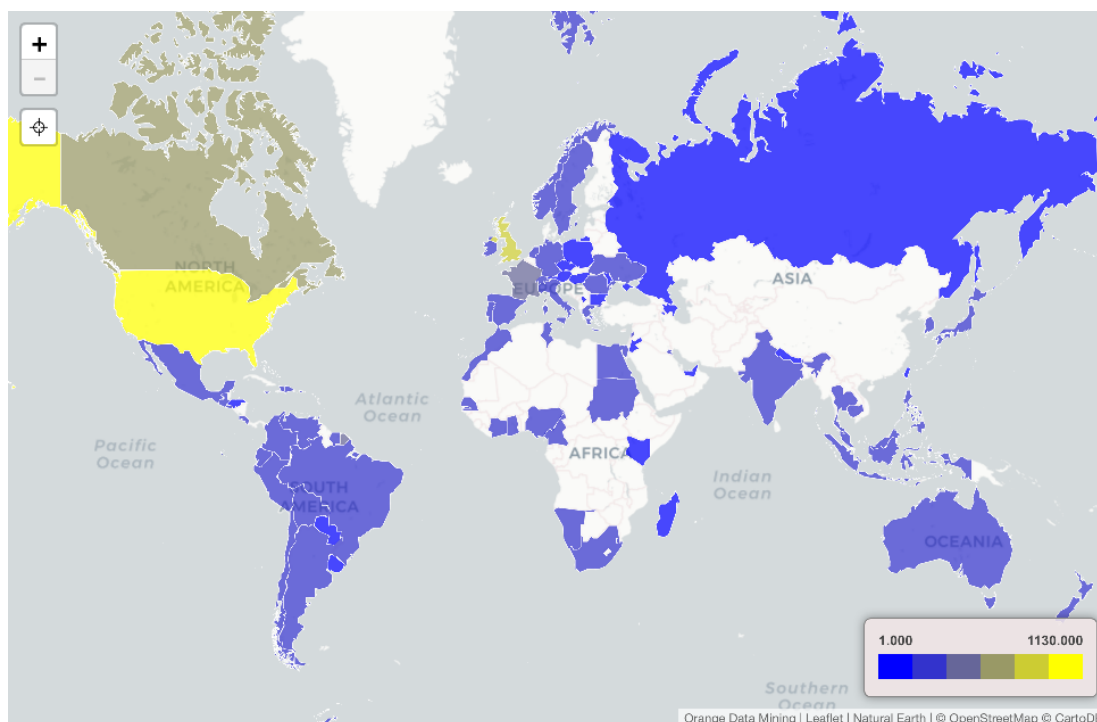
Figura 17 – Gráfico com top 10 fontes dos tweets



Fonte: Autor



Figura 19 – Mapa dos locais dos tweets



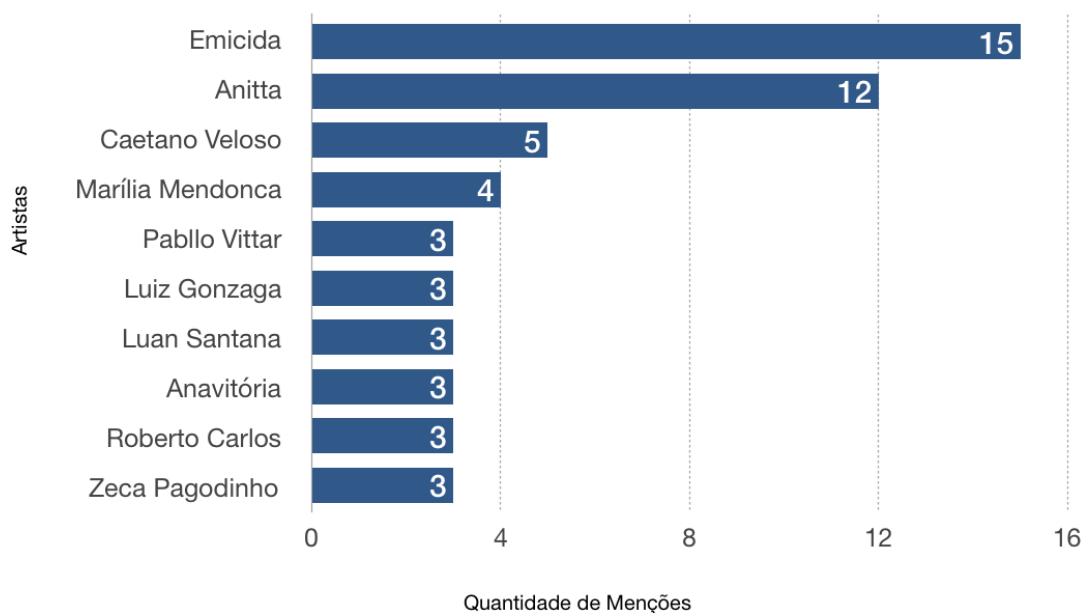
Fonte: Autor

## Lista de Artistas e gêneros musicais

Apesar da perda de nomes por falta de reconhecimento, utilizando o software de Standford e Monkey Learn teve-se como resultado um número total de 1013 artistas, o que foi um resultado satisfatório.

Foi criado um cluster chamado nacional, que engloba todos os artistas nacionais presentes independente do gênero musical que cada um apresenta. Inicialmente realizado alguns testes apenas com esses que estão nessa categoria para saber quanto desses artistas estavam presentes na lista de Tweets. Como são apenas 25 nomes no dicionário de artistas, utilizando o arquivo total de Tweets obteve-se a frequência de menções que esses artistas tiveram. Os nomes nacionais que mais apareceram nos Tweets totais, pode ser encontrado na Figura 20.

Figura 20 – Artistas nacionais mais ouvidos na busca dos meses de outubro e novembro

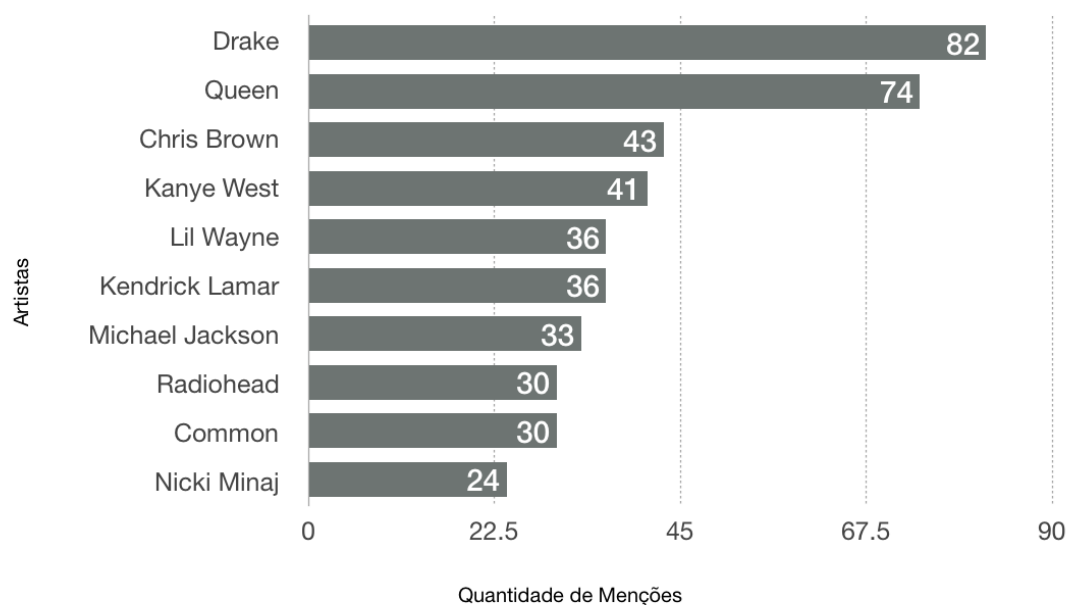


Fonte: Autor

O teste foi realizado para conseguir identificar os artistas, independente dos gêneros musicais. No caso dos artistas da Figura 20, todos estão na mesma categoria, que é nacional, então não tivemos outras categorias para serem testadas.

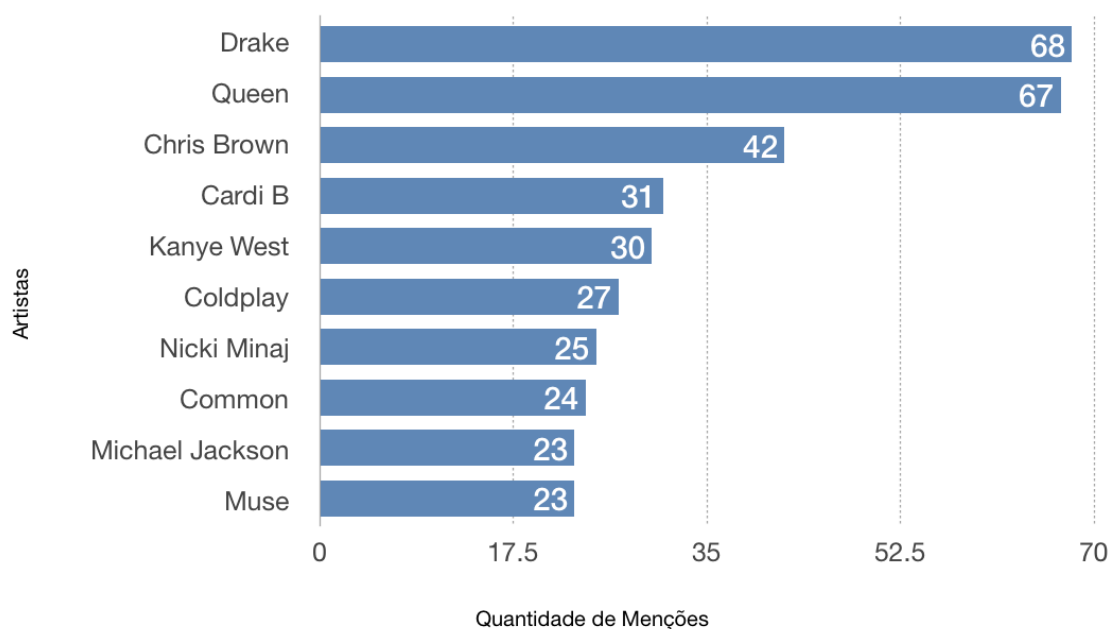
Com o reconhecimento de artistas testado utilizando os artistas nacionais, o próximo passo foi, além de conseguir pegar os nomes, conseguir identificar também o gênero musical que está a ele atribuído segundo a tabela de artistas que foi salva localmente baseado nos dados do Spotify e Lastfm. Após os testes com os artistas nacionais, o próximo passo foi criar clusters com outros gêneros musicais. Devido a grande quantidade de artistas e como muitos deles não aparecem tantas vezes, foram definidos 5 gêneros musicais principais. Os gêneros escolhidos foram pop, rock, rap, hip hop e reggae. Os arquivos dos tweets foram separados em dois arquivos, um com os tweets de outubro e outro com os de novembro. Desta forma, obteve-se os artistas mais mencionados nas buscas que ocorreram no mês de outubro e no mês de novembro independente do gênero musical que os mesmos estão caracterizados, como mostra as Figuras 21 e 22.

Figura 21 – Top 10 Artistas mais mencionados em outubro



Fonte: Autor

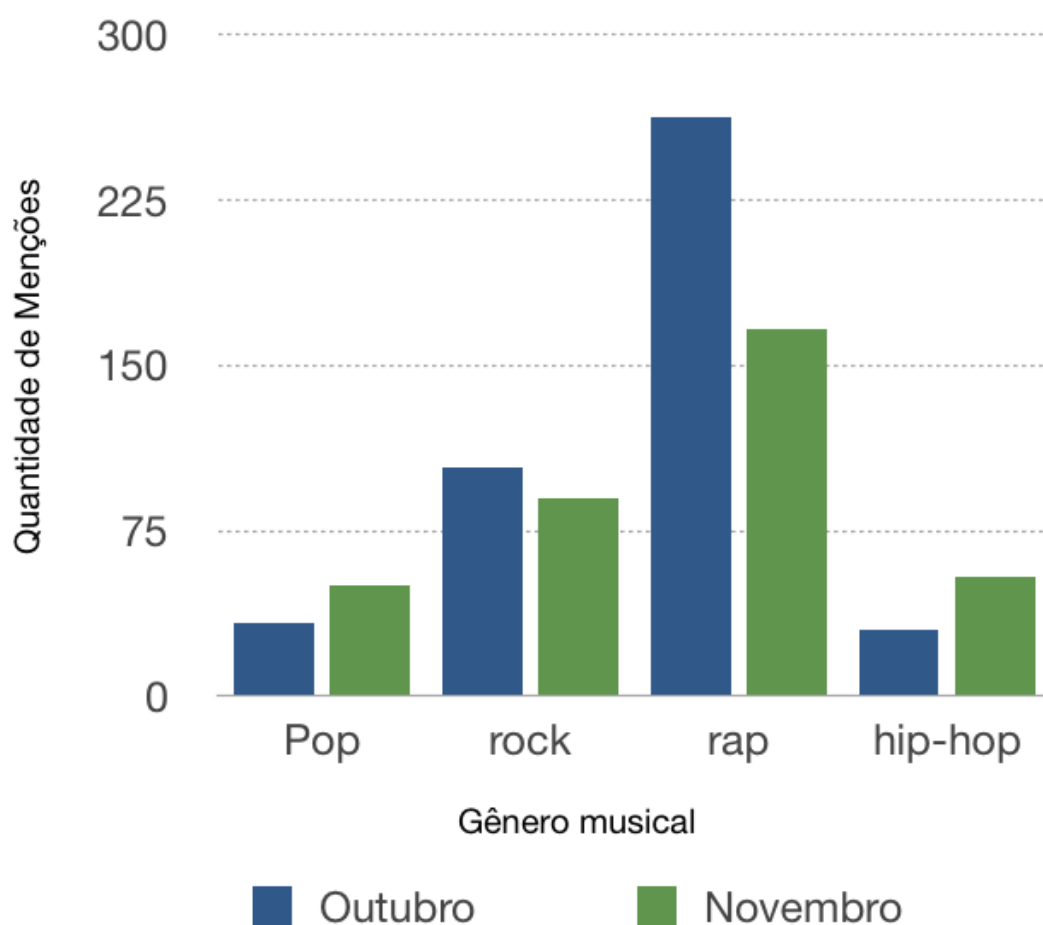
Figura 22 – Top 10 Artistas mais mencionados em novembro



Fonte: Autor

Como foi possível observar nas Figuras 21 e 22, alguns nomes em comum apareceram no top 10 de artistas mais mencionados. Comparando com o gênero musical de cada artista que foi mais mencionado, foi possível descobrir quais gêneros musicais foram os mais ouvidos, como mostra a Figura 23. Como foi notado, o reggae não figura entre os gêneros mais ouvidos no top 10 de outubro e novembro.

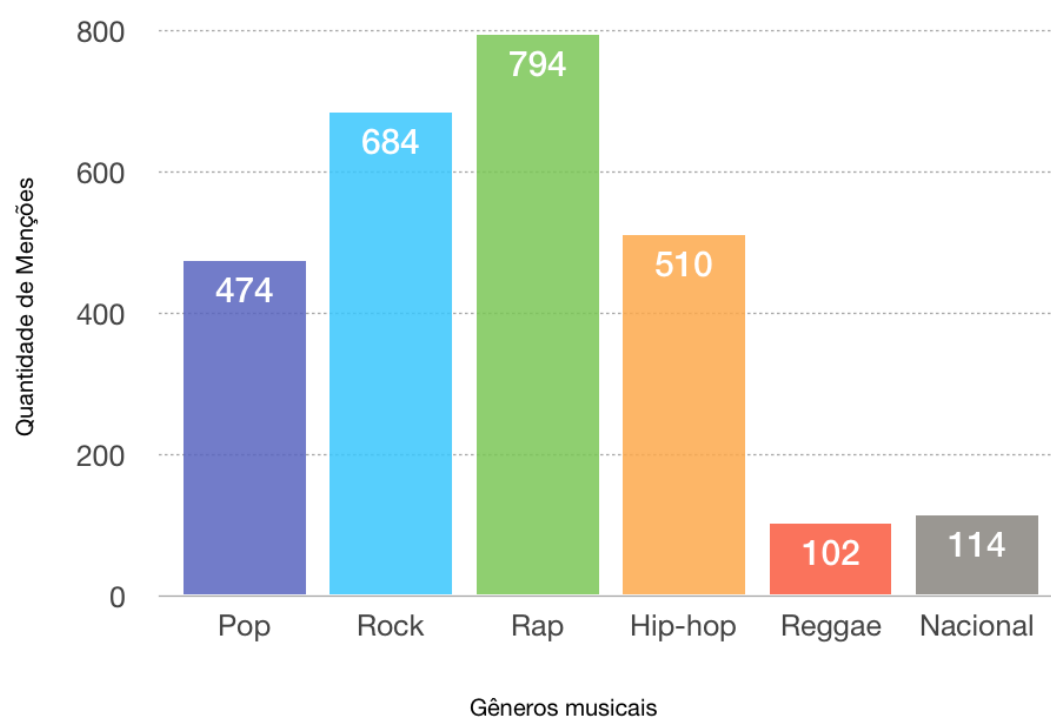
Figura 23 – Gêneros musicais mais ouvidos baseado no top 10 de artistas de outubro e novembro



Fonte: Autor

E por fim, baseado não só no top 10, mas sim no conjunto inteiro de menções a cada artista que está presente na lista dos 150 artistas, incluindo os nacionais, que estão presentes no banco de dados criado para atribuir seus nomes aos seus gêneros musicais, temos os gêneros musicais mais ouvidos nos dias buscados de outubro e novembro, como mostra a figura 24.

Figura 24 – Gêneros musicais mais ouvidos baseado nos 150 artistas e nas menções que receberam nos meses de outubro e novembro



Fonte: Autor

# Conclusão

O trabalho apresentado lida com o excesso de informações presentes na rede social Twitter trabalhando com tweets e tem intenção de apresentar uma abordagem sobre o assunto buscando conhecimento sobre os dados em questão. Todo o processo foi realizado utilizando Python e suas bibliotecas que facilitaram incalculavelmente o trabalho com os dados. Obteve-se resultados positivos em relação aos campos selecionados e foi possível realizar a limpeza, extraindo informações que podem servir para uma personalização futura de conteúdo para os usuários combinando o idioma, localização e estilo musical que os mesmos escutam.

## 5.1 Dificuldades

Um dos fatores cruciais e que gerou uma mudança de planos em relação ao conteúdo do trabalho é a limitação que o Twitter implantou ao longo dos anos. Em épocas de pesquisa para iniciar esse trabalho, teve-se acesso a diferentes trabalhos onde as buscas por tweets se fazia de maneira mais completa e sem precisar realizar tantas requisições. Um dos fatores que teve maior mudança foi em relação a data, pois a busca por tweets antigos era possível a qualquer usuário e não era limitado a tweets com data de até 7 dias. Atualmente o Twitter ainda permite fazer buscas por datas mais antigas, mas só é permitido para usuários ou empresas que comprarem direito para utilizar ferramentas da conta premium.

## 5.2 Trabalhos futuros

Devido a quantidade de etapas a serem realizadas e de como os dados precisavam ser tratados de diversas formas até que se tivesse um certo conhecimento sobre os mesmos, não foi possível tratar de todos os detalhes que foram previamente estabelecidos, portanto é necessário que a pesquisa continue em trabalhos futuros para conseguir melhor resultados, dentre eles:

- Melhoramento do reconhecimento de entidades, pois não foi capaz identificar todos os nomes de artistas, principalmente os compostos. Outro problema encontrado no reconhecimento de entidade, foi em relação ao idioma, pois apesar de conter a segunda maior quantidade de tweets, os conteúdos em japonês não foram reconhecidos como nomes.



- O modelo implementado contou com um banco local baseado em informações do Lastfm e Spotify, e com isso teve número limitado de artistas pois os dados foram abastecidos manualmente. Um caso para próximos trabalhos seria ter maior quantidade de artistas e seus respectivos gêneros musicais.
- O modelo apresentado para captura e reconhecimento de informações torna possível trabalhar com os dados de maneira mais completa e pode ser usado para criar personalização para os usuários que enviaram mensagens contendo determinado artista e com isso haver o reconhecimento de qual nicho o mesmo se encontra.

# Referências

- ASHKTORAB, Z. et al. Tweedr: Mining twitter to inform disaster response. In: *ISCRAM*. [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 26 e 28.
- BEZERRA, R. G. E. A tarefa de classificação em text mining. *Revista de Sistemas de Informação da FSMA*, v. 5, p. 42–62, 2010. Citado 2 vezes nas páginas 21 e 38.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. 03 2000. Citado 3 vezes nas páginas 17, 18 e 19.
- GUPTA, V.; LEHAL, G. A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, v. 1, 08 2009. Citado na página 21.
- JUNIOR ERIC ROMMEL, D. S. R. R. d. A. W. C. V. J. C. Uma análise comparativa entre algoritmos estatísticos de mineração de dados. Natal/RN,, 2008. Citado 4 vezes nas páginas 14, 20, 25 e 26.
- LIM, K. H.; DATTA, A. Interest classification of twitter users using wikipedia. In: . [S.l.: s.n.], 2013. Citado na página 26.
- Makki, R. et al. Twitter message recommendation based on user interest profiles. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.: s.n.], 2016. p. 406–410. ISSN null. Citado 2 vezes nas páginas 26 e 28.
- SILVA RICHARDSON RIBEIRO, M. T. F. E. W. Identificando emocoões em redes sociais: Um estudo de caso no facebook. 2014. Citado 2 vezes nas páginas 25 e 27.
- XU ZHIHENG; RU, L. X. L. Y. Q. Discovering user interest on twitter with a modified author-topic model. In: *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. [S.l.: s.n.], 2011. v. 1, p. 422–429. Citado na página 26.
- ZANGRANDI, A. Detecção de eventos com dados do twitter. v. 5, 2014. Citado 3 vezes nas páginas 29, 30 e 31.