Implementação para deployment remoto do script

1.Obtenção da lista de artistas

No último documento deste projeto, foi utilizada uma lista de 100 artistas obtida da revista Rolling Stones Brasil. Para essa etapa de implementação é necessário encontrar uma forma de obter o nome de mais artistas brasileiros relevantes.

Não há uma forma viável de obter uma lista de artistas com número de ouvintes pré determinado, porque o endpoint do spotify não retorna número de ouvintes mensais nem ouvintes, então seria necessário realizar um web-scraping massivo. A API do last-fm retorna o número de usuários ouvintes total, porém, da mesma forma, seria necessário exaustivamente passar pelos artistas indexados na API. Logo a melhor forma é utilizar o método tag.getTopArtists da API do last-fm que retorna um número fixo de artistas com determinada tag, será utilizada a tag “brazil”, “brasil”, “brazilian” e “musica brasileira”.

A técnica descrita no parágrafo anterior,foi aplicada com êxito. Porém deve-se levar em consideração que a tag getTopArtists retorna artistas em ordem aparentemente não somente proporcional ao número total de ouvintes e sim de acordo com critérios internos não dispostos em sua documentação. Além disso é bom sempre lembrar do viés intrínseco da utilização da API do [last.fm](http://last.fm), que é uma rede social com público predominantemente jovem, a maior parte dos artistas retornados será de gêneros ouvidos em sua maioria por esse público. Porém é seguro assumir que uma Query com um número grande de artistas retornará o nome não somente desses artistas, mas de muitos outros que têm notoriedade no Brasil.

2. Obtenção das músicas e letras das músicas dos artistas.

A lista de músicas foi obtida, assim como descrito no último documento, com o método artist.getTopTracks da API [last.fm](http://last.fm), satisfatoriamente.

Por sua vez, a API lyrics.ovh obteve resultados decepcionantes. Para sanar essa deficiência e obter as letras das músicas, utilizou-se uma combinação de requisições HTTP à API do Genius e web scraping das páginas das músicas. A API foi utilizada para localizar a URL da música a partir do nome do artista e da faixa, e em seguida a letra foi extraída diretamente da página da música com uso da biblioteca *BeautifulSoup*, já que a API não fornece o conteúdo completo por padrão. Nota-se que a biblioteca *lyricsgenius* facilita o acesso à API do Genius para buscar músicas e letras, mas atualmente não está funcionando.

Os resultados obtidos foram satisfatórios por enquanto, porém não consegui isolar alguns comentários de usuários sobre as músicas da letra. Já que às vezes nesses comentários são citados outros artistas, há uma distorção no resultado final.

Deve se registrar que, atualmente, o fator limitante para a implementação com uma quantidade massiva de músicas está na etapa de obtenção das letras, visto que para cada música o método está demorando cerca de 1-2s. Tempo que rapidamente figura o tempo necessário para obtenção do conjunto em dias. Com o intuito de viabilizar a execução do processo por longos períodos de tempo, a cada artista processado (obtido o conjunto de letras de suas músicas), o script armazena as letras num arquivo de formato ndjson. Para que, caso haja interrupção, o script passará rapidamente pelos artistas já processados e continuará no primeiro artista cujas letras ainda não estão armazenadas.

3. Retirada do modelo de NLP

Previamente, havia sido utilizado um modelo NER (Named Entity Recognition) numa tentativa de reduzir a captação de falsas referências. Com fins de viabilizar a implementação com muitos artistas e músicas, essa etapa foi retirada, visto que há um custo computacional significativo e ainda não foi realizada uma análise de custo-benefício da técnica. Logo, o código implementado capta referências simplesmente iterando sobre as letras de um artista e registrando quando a letra contém nomes dos outros artistas do conjunto.