ANÁLISE DA SIMILARIDADE DE GÊNEROS MUSICAIS BRASILEIROS UTILIZANDO WEB SCRAPING E MINERAÇÃO DE TEXTOS NO R

Luiz Fernando Guilhem Nassif Maia[1], Alinne de Carvalho Veiga[2] e Renata Souza Bueno[3]

Introdução

A música é uma das formas de arte mais importantes para o ser humano, em especial para o povo brasileiro. A música brasileira é muito diversificada visto que sofre influências de diversas culturas como a europeia, a indígena, a africana e a americana.

Este trabalho se propõe a verificar quais as principais palavras que diferenciam cada estilo musical brasileiro e quais os gêneros que mais se assemelham em relação às letras de suas músicas.

Para isso, foi montado um banco de dados utilizando *Web Scraping*, uma técnica de extração de dados da internet pela leitura de códigos *HTML*. Inicialmente, foram considerados treze estilos musicais diferentes e coletou-se dados como nome da música, dos cantores e letra da música para mil músicas de cada estilo.

Técnicas de processamento de texto comumente utilizadas em mineração de texto como remoção de *stop words* e stemização foram utilizadas. Além disso, para cada palavra em cada estilo musical foi calculada a medida *tf-idf* (JONES, 1972). Finalmente, foi computada a medida de similaridade do cosseno (ARAÚJO NETO e NEGREIROS, 2017) para cada gênero musical levando em conta os valores do *tf-idf* e foi realizada uma análise de agrupamentos com essa similaridade.

Objetivos

Construir um banco de dados de músicas, seus cantores, estilos e suas letras utilizando *Web Scraping*.

Verificar quais palavras caracterizam um determinado gênero musical.

Agrupar os diferentes estilos musicais empregando técnicas de análise multivariada e características das letras.

[1] Escola Nacional de Ciências Estatísticas (ENCE), e-mail: luiz_fgnm@hotmail.com

[2] Escola Nacional de Ciências Estatísticas (ENCE), e-mail: alinne.veiga@ibge.gov.br

[3] Escola Nacional de Ciências Estatísticas (ENCE), renata.bueno@ibge.gov.br

Material e Método

Todo o trabalho foi realizado usando o software *R* (R DEVELOPMENT CORE TEAM, 2011). A coleta de dados foi realizada no site Letras[4] utilizando o pacote *rvest* (WICKHAM, 2016). Foram coletadas mil músicas dos estilos sertanejo, funk, gospel, axé, mpb, samba, pagode, hip hop/rap, eletrônica, pop, rock, heavy metal e soul, os cinco últimos foram ignorados visto que apresentavam musicais internacionais em sua grande maioria.

O pré-processamento das letras musicais ocorreu em algumas etapas. Primeiramente, foram excluídas músicas repetidas em um mesmo gênero pois existiam músicas semelhantes de cantores distintos. Posteriormente, palavras com menos de duas letras, números e *stop words* (palavras que não acrescentam muita informação ao texto como pronomes e artigos) foram removidas assim como palavras repetidas dentro de uma mesma música. Por último, todas as palavras foram reduzidas ao seu radical, utilizando um processo de stemização, todas as músicas de um mesmo estilo foram agregadas e foi calculado a medida *tf-idf* para cada palavra em cada estilo musical. Essa medida reflete o quão importante uma palavra é para um documento, nesse caso para um estilo, em um conjunto de documentos e nos possibilita averiguar quais palavras caracterizam cada estilo musical. Os principais pacotes utilizados durante o pré-processamento foram *tidytext* (SILGE e ROBINSON, 2016), *tm* (FEINERERER, HORNIK e MEYER, 2008) e *ptstem* (FALBEL, 2017).

Podemos visualizar cada gênero musical como sendo um vetor n-dimensional em que o valor computado para a estatística *tf-idf* para cada palavra é uma dimensão desse vetor. Desse modo, podemos calcular a semelhança entre os estilos utilizando alguma medida de similaridade. Normalmente, em mineração de textos, esses vetores são muitos esparsos e uma medida de similaridade que não leve em consideração as correspondências entre zeros se torna bastante interessante. Assim sendo, foi utilizada a medida de similaridade do cosseno para esse trabalho dado que ela tem essa propriedade.

De posse das similaridades, foi realizada uma análise de agrupamentos hierárquica para visualizar as semelhanças entre os gêneros musicais, o método utilizado foi o de Ward (WARD JR, 1963), mas vale a pena ressaltar que outros métodos foram empregados e possuíram resultados semelhantes.

[4] Disponível em https://www.letras.mus.br/, acessado em 7 de março de 2018.

Resultados e Discussão

Os maiores valores para a medida *tf-idf* ocorreram nos gêneros gospel, axé, funk e hip hop/rap, o que sugere que esses estilos possuem palavras que realmente os caracterizam. Palavras como aleluia, ressureição e Israel estão entre as que mais definem o estilo gospel enquanto que os estilos funk e hip hop/rap distinguem-se dos demais pelo uso excessivo de palavrões.

Considerando a similaridade do cosseno, os estilos mais semelhantes foram funk e hip hop/rap enquanto a menor similaridade ocorreu entre os gêneros gospel e funk. Em relação à análise de agrupamentos, podemos separar esses gêneros em quatro grupos: axé, samba e mpb; funk e hip hop/rap; pagode e sertanejo e um grupo formado apenas pelo gênero gospel.

Conclusão

Neste trabalho foi realizada uma análise de oito diferente estilos musicais brasileiros no que se refere às letras de suas músicas, foi possível agrupar esses diferentes gêneros e explorar as palavras que mais particularizam cada estilo.

Futuramente, estuda-se utilizar outras variáveis dessas músicas como batidas por minuto, duração e volume em conjunto com as variáveis textuais.

Referências

ARAÚJO NETO, A.; NEGREIROS, M. Avaliação da performance de índices de similaridade aplicados ao agrupamento de objetos textuais. **Revista Brasileira de Computação Aplicada**, v. 9, n. 4, p. 43-59, 2017.

FALBEL, D. ptstem: Stemming Algorithms for the Portuguese Language. **Comprehensive R Archive Network**, 2017. Disponível em <https://CRAN.R-project.org/package=ptstem>. Acesso em 4 de abril de 2018.

FEINERER, I.; HORNIK K.; MEYER D. Text Mining Infrastructure in R. **Journal of Statistical Software**, 2008. Disponível em <https://www.jstatsoft.org/article/view/v025i05>. Acesso em 4 de abril de 2018.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, v. 28, n. 1, p. 11-21, 1972.

R CORE TEAM. **R - A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing. Vienna, 2017.

SILGE, J.; ROBINSON, D. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. **The Journal of Open Source Software**, v. 1 n. 3, 2016. Disponível em <http://joss.theoj.org/papers/10.21105/joss.00037>. Acesso em 4 de abril de 2018.

WARD JR, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association**, v. 58, n. 301, p. 236-244, 1963.
WICKHAM, H. rvest: Easily Harvest (Scrape) Web Pages. **Comprehensive R Archive Network**, 2016. Disponível em <https://cran.r-project.org/web/packages/rvest/>. Acesso em 4 abril de 2018.