UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

NICÁCIO OLIVEIRA DE SOUSA

ANÁLISE E VISUALIZAÇÃO DE DADOS SOBRE A EVOLUÇÃO DAS LETRAS NO FORRÓ

CAMPINA GRANDE - PB

2019

NICÁCIO OLIVEIRA DE SOUSA

ANÁLISE E VISUALIZAÇÃO DE DADOS SOBRE A EVOLUÇÃO DAS LETRAS NO FORRÓ

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Nazareno Andrade.

CAMPINA GRANDE - PB

2019

S725a Sousa, Nicácio Oliveira de.
Análise e visualização de dados sobre a evolução das letras no forró. / Nicácio Oliveira de Sousa. – 2019.

11 f.

Orientador: Prof. Dr. Nazareno Andrade.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Processamento de linguagem natural. 2. Análise de letras. 3. Forró – análise de letras. 4. Processamento de textos. 5. Biblioteca cheerio. 6. Biblioteca Nodejs. 7. Biblioteca biblioteca Sequelize. 8. Biblioteca Puppeteer I. Andrade, Nazareno. II. Título.

CDU:004(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

NICÁCIO OLIVEIRA DE SOUSA

ANÁLISE E VISUALIZAÇÃO DE DADOS SOBRE A EVOLUÇÃO DAS LETRAS NO FORRÓ

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Nazareno Andrade
Orientador – UASC/CEEI/UFCG

Professor Dr. Francisco Vilar Brasileiro
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Examinador – UASC/CEEI/UFCG

Trabalho aprovado em: 25 de novembro 2019.

CAMPINA GRANDE - PB

Análise e Visualização de Dados Sobre a Evolução das Letras do Forró

Trabalho de Conclusão de Curso

Nicácio Oliveira de Sousa
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
nicacio.sousa@ccc.ufcg.edu.br

Nazareno Andrade
Laboratório de Sistemas Distribuídos
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
nazareno@computacao.ufcg.edu.br

PALAVRAS-CHAVE

Análise de Dados, Análise de Letras, Forró, Processamento de Linguagem Natural

1. RESUMO

Entender como a música evolui, nos permite identificar quais os fatores que influenciam na sua concepção e como esses fatores mudam ao longo do tempo. A música é utilizada pela sociedade para externar o sentimento de cada momento que atravessa no decorrer da história, ou seja, está diretamente ligada ao processo de evolução da sociedade. Cada letra de cada música, carrega palavras e contextos que podem ter sido influenciados diretamente pelo momento em que o compositor estava inserido. Cada gênero musical está associado a uma cultura diferente ou a uma mistura de culturas. Por esse motivo, estudar as composições que fazem parte de um gênero específico, ao longo do tempo, pode nos ajudar a entender melhor a evolução dessa cultura em volta do mesmo e permitir o entendimento de termos e contextos utilizados em suas composições. Por tanto, utilizando raspagem de dados, técnicas de análise de dados, em especial técnicas de processamento de texto, e visualizações em conjuntos de letras de músicas organizadas por décadas, este trabalho tem o objetivo de demonstrar quais são as possíveis variáveis que influenciam e como se modificaram, no decorrer do tempo, em relação a letras de músicas do gênero musical forró e também facilitar o estudo de outros gêneros sugerindo uma metodologia para busca de dados de músicas. Este trabalho busca, também, entender e demonstrar quais lacunas existem neste tipo de análise em relação a obtenção dos dados.

2. INTRODUÇÃO

Técnicas de Análise de dados em textos têm sido amplamente utilizadas, nos últimos anos, para buscar informação dentro de textos que parecem ser óbvias. Classificação, análise de sentimentos, mineração em textos etc, são utilizados para avaliar a composição, o vocabulário, possíveis sentimentos advindos das palavras e até mesmo classificação para tentar decifrar como certas letras de músicas fizeram sucesso. A grande maioria dos trabalhos que buscam informações relacionadas a músicas e suas letras, utilizam um conjunto de dados muito abrangente e que envolve diversos gêneros e artistas ou fazem análise apenas em composições de um artista ou listas de maiores sucessos da música. Como exemplo de trabalho mais generalista podemos observar a análise sobre a música brasileira feita por Leo Sales [1]. Neste trabalho, é feito um apanhado com indicadores de quantidades e vocabulário, englobando todos os gêneros musicais que existem no brasil, sem considerar se o gênero é brasileiro de fato ou não, por exemplo. Partindo do ponto de que a análise de dados é feita em um conjunto de vários gêneros e que isso pode não considerar características mais restritas a cada gênero em particular, surgiu a ideia de estudar um gênero específico e brasileiro para tentar, não somente quantificar alguns pontos relativos às letras das músicas do estilo, entender como o gênero evoluiu em termos de como ele foi se modificando ao longo do tempo. Por essa perspectiva de refinar a análise para possibilitar o estudo por gênero, podemos avaliar as letras de uma forma mais concisa e profunda buscando tornar possível interpretar os resultados a partir da data de cada música, ou seja, podemos criar blocos de décadas e permitir que as funções de análise sejam aplicadas baseando-se nas mesmas. Além da análise de dados, é importante destacar que tentaremos entender qual a dificuldade de fazer esse tipo de estudo em relação a como e onde obter dados.

3. FUNDAMENTAÇÃO TEÓRICA

A análise de dados em letras de músicas vem ajudando pesquisadores a entender como e porque as composições estão sendo moldadas para se adequarem às mudanças que ocorrem naturalmente na sociedade. Análises em letras de rap [2], por exemplo, demonstram que existe uma tendência na qual as letras se aproximam do pop cada vez mais. Seja pela quantidade de palavras únicas que estão diminuindo e tornando-se similares a letras da música pop ou pelo contexto em relação a palavras mais utilizadas no decorrer do tempo. Essa aproximação de letras da música pop indica os pontos que levaram as músicas de rap a fazerem sucesso devido a proximidade com letras que possuem um vocabulário mais fácil como no pop, por exemplo.

O contexto indicado nas letras pode indicar muitos fatores sociais ligados à época que elas foram escritas. Um exemplo claro disso pode ser observado na análise do rap [2], onde foi possível criar uma relação entre as letras que fazem referência a localidades

geográficas e, pelo contexto das letras, foi possível notar que todas as localidades citadas estavam ligadas a taxa de criminalidade e esse ponto é constantemente citado em letras de rap tradicionais.

Na análise da música brasileira feita por Sales [1], a distribuição da quantidade de acordes únicos e a quantidade de palavras únicas das composições para a maioria dos gêneros musicais do brasil, permitiu observar a disparidade entre os gêneros assim como, também, a conexão entre alguns gêneros como o punk-rock dos Raimundos que possui certas características do Forró. Em outro ponto da análise feita por Sales, foi possível observar a correlação da quantidade de letras em relação a quantidade de palavras distintas [4]. Nesse ponto é possível observar até mesmo os artistas que possuem um vocabulário repetitivo, configurando, possivelmente, letras de músicas mais "pobres" em relação ao vocabulário. É importante destacar que, nessa análise da música brasileira, a dificuldade na captura dos dados foi notória pelo fato de não existir um lugar, site ou banco de dados que possua informações certificadas e bem estruturadas.

Existem esforços em alguns estudos para tentar classificar músicas que fizeram sucesso e tentar prever se novas composições possuem atributos para fazerem sucesso baseando-se em músicas que fizeram sucesso [5]. Nesse caso, a análise pode ser muito subjetiva pelo fato de que o sucesso de uma música está relacionado a dezenas de fatores que vão desde o artista que endossa a composição até mesmo gosto individual do ouvinte.

4. METODOLOGIA

Para este caso de estudo, foi escolhido o gênero musical forró como objeto de estudo, por ser um gênero em grande ascensão nos últimos anos e principalmente por nos dar a possibilidade de explorar a sua mudança no decorrer dos anos que parte de tradicionalismo nos anos 50 e 60 com o seu precursor Luiz Gonzaga até os dias atuais com sua forma mais moderna e misturada com o sertanejo universitário e o funk carioca.

Baseado em uma análise em relação ao tempo, utilizando as décadas das composições, este trabalho busca explorar e analisar, através de técnicas de NLP[6], os seguintes pontos:

- Variação da quantidade de músicas por década
- Palavras mais utilizadas no contexto geral
- Palavras mais utilizadas (atemporais) no decorrer do tempo
- Tamanho das palavras (no sentido de indicar se as palavras ficaram mais ou menos complexas no decorrer do tempo)
- Diversidade e densidade léxica ao longo do tempo
- Frequência e popularidade de palavras

É importante destacar que a busca pelos dados possa ser considerada como ponto principal deste estudo por ter como proposta incentivar e demonstrar como buscar pelos dados e possibilitar trabalhos futuros que tenham o objetivo de estudar gêneros musicais de forma mais detalhada.

4.1 Desafios

O maior impedimento deste estudo está diretamente ligado a como e onde obter os dados. Diferente de muitos outros gêneros musicais, o forró ainda tem seu meio como "informal" no tocante à formalização das composições em termos de metadados. Ou seja, não existe um lugar comum que contenha uma quantidade razoável de letras e metadados referentes às mesmas. As composições, em sua maioria, são consumidas por blocos regionais do Nordeste do Brasil. Esse fato dificulta ainda mais a obtenção de dados sobre artistas e músicas de forró. A partir do ano de 2010, esse problema sobre falta de dados diminui um pouco pelo fato de que a partir desse ano, as popularidade das músicas de forró tomaram uma proporção maior, atingindo o Brasil inteiro, fazendo com que existam mais letras e meta-dados na internet de forma natural.

A existência do website www.letras.mus.br[1], o qual é nutrido por seus próprios usuários, possibilitou a obtenção do nosso objeto de estudo. Devido a esse ponto, os dados utilizados neste estudo foram coletados no website utilizando raspagem de dados. Apesar de termos uma boa quantidade útil de informações sobre as músicas de forró no website como nome, artista, letra e algumas referências a albúns, uma parte das composições não possui nenhum link referenciando a datas de lançamento. Sendo assim, mais um desafio foi percebido em relação a onde e como buscar mais informações referentes a datas de lançamento para as músicas.

4.2 Conjunto de Dados

O conjunto de dados para a análise que foi descrita, possui informações do nome do artista, a letra da música, nome da música e data de lançamento. É importante destacar que as datas podem não ser exatamente respectivas a cada música, devido a dificuldade de obtenção que foi citado anteriormente e que pode ser observada analisando os scripts que foram criados.

O conjunto de dados possui, aproximadamente, 37 mil letras de músicas de forró. Porém apenas aproximadamente 30% dos dados possuem data de lançamento. As músicas foram separadas por década através dos scripts de análises e esse dado está incorporado ao conjunto para classificação a partir da década de 50 até a década de 2010.

4.3 Coleta dos dados

Todo o processo de raspagem utilizou as bibliotecas: Cheerio, Nodejs, Sequelize e Puppeteer. Todas possuem uma vasta documentação e são de fácil manuseio.

De maneira geral, o processo de raspagem foi pensado para que possa ser utilizado em outras análises e seguiu a seguinte ordem:

- Busca por artistas
- Busca por gêneros musicais e classificação dos gêneros por artista
- Busca por músicas relacionadas aos artistas

[1] Acessar: www.letras.mus.br

- Busca por letras
- Busca por datas de lançamento das músicas

Essa metodologia foi pensada para que outros pesquisadores possam utilizá-la para prosseguir com o estudo nessa temática.

Um ponto importante que devemos levar em consideração é que, sabendo que o website letras não possui todas as datas de lançamento para as músicas, para possibilitar a análise em uma quantidade de músicas razoável, a utilização do browser headless [8] Puppeteer foi necessária para obter dados referentes à datas de lançamento para as músicas através do google, para tentar preencher o máximo possível de composições e suas datas de lançamento e possibilitar um melhor aproveitamento dos dados. Essa busca é feita pesquisando pelo termo "Ano da música" concatenado ao nome da música e do artista e a partir disso, caso o google responda através de um quadro com a data de lançamento da composição, o dado é salvo para a música em questão. Para casos em que não existe resposta fixa do buscador, ocorrem respostas que referenciam vídeos à música que é pesquisada, então a menor data, vinculada ao vídeo possível nessa lista de vídeos é salva como data de lançamento da música. Para os casos em que nenhum dos casos citados anteriormente ocorrem, a pesquisa é descartada e segue para a próxima música.

4.3.1 Processo de raspagem

O website www.letras.mus.br[2] responde a buscas e requisições com páginas estáticas. Isso facilitou o processo de raspagem por termos como resultados, páginas bem estruturadas. Todos os artistas podem ser listados por letras do alfabeto da seguinte forma "https://www.letras.mus.br/letra/A/artistas.html[3]", isso possibilita a criação de requisições para todas as letras do alfabeto, trocando apenas a letra e obtendo o resultado como html estático de toda a página contendo nomes vinculados à links para cada página de cada artista. Isso acontece com todas as páginas do website, seja ela uma página com nomes de músicas ou página do artista.

Devido a essa forma de estrutura do website, que foi citada acima, todos os scripts seguem a mesma lógica e sequência. Cada script busca pelos links para cada página que será processada e guarda-os para possibilitar uma busca direta por cada página. A forma a qual o letras.mus.br está estruturado permite essa estratégia que facilita a raspagem retirando a necessidade do uso de um browser headless por apenas requisições diretas para cada página e também favorece a possibilidade de acompanhar o andamento de cada script para casos de falhas onde precisamos saber o ponto em que o script parou. O detalhamento de cada passo e script para captura dos dados está no repositório [7] do processo assim como um backup do conjunto de dados que foi obtido.

4.3.2 Processo de Análise

Para facilitar a manipulação dos dados, a utilização de um banco de dados relacional foi crucial. Permitiu criar relacionamentos para facilitar a busca de dados relacionados como músicas de um artista, artistas de um gênero etc. Para utilizar os dados na análise, os dados foram convertidos em uma planilha através de um script de conversão no R o qual foi utilizado para esta análise. Esse script assim como os scripts de análise e seu detalhamento estão disponíveis no repositório [9].

5. ANÁLISE E RESULTADOS

Antes de mais nada, é importante ter conhecimento sobre como a forma e quantidade dos dados em cada década vai impactar em nosso entendimento e análise. Logo, tendo em vista que a há uma discrepância considerável em relação a quantidade de músicas por décadas, sempre que fizermos comparações que indicam quantidade ou tenham a quantidade de palavras ou letras como parâmetro, é importante considerar esse ponto. Em alguns pontos, vamos utilizar somente os resultados que tenham uma quantidade de músicas acima de 100 para que isso possibilite uma análise mais concisa dos dados.

5.1 Disposição dos dados

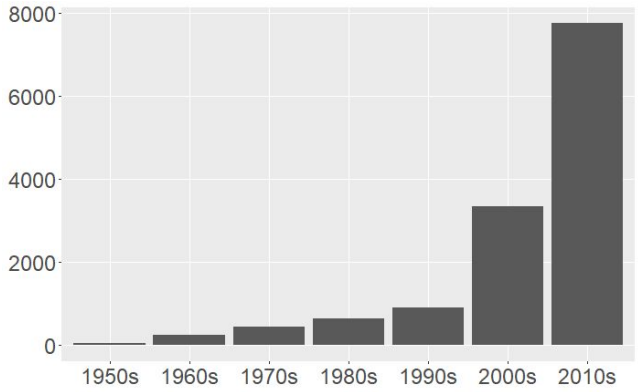

Figura 1: Quantidade de músicas por década

O quantidade de músicas por década, evidencia uma quantidade de músicas expressiva entre os anos 90 e 2010. Esse fato não indica necessariamente que fora desse intervalo existam poucas músicas. Indica que nesse intervalo, com maior quantidade, fatores diversos contribuíram com esse crescimento e podemos ligá-los à ascensão da internet, democratização de ferramentas etc.

[2] Acessar: www.letras.mus.br

[3] Acessar: https://www.letras.mus.br/letra/A/artistas.html

5.2 Palavras mais utilizadas por década

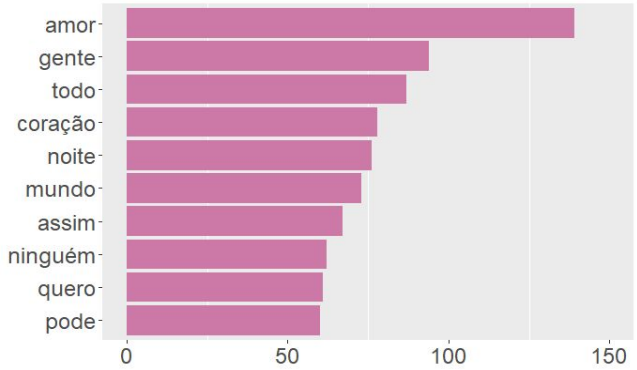

Figura 2: Quantidade de palavras mais utilizadas na década de 70

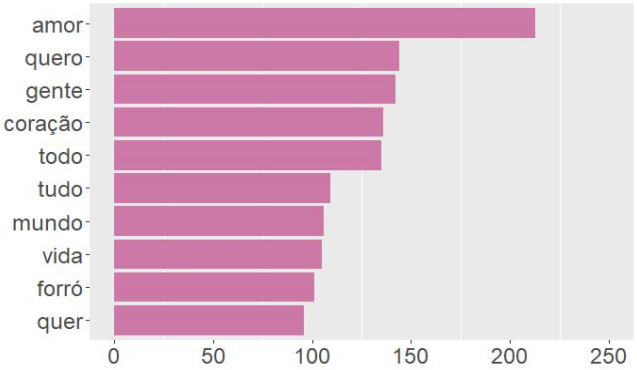

Figura 3: Quantidade de palavras mais utilizadas na década de 80

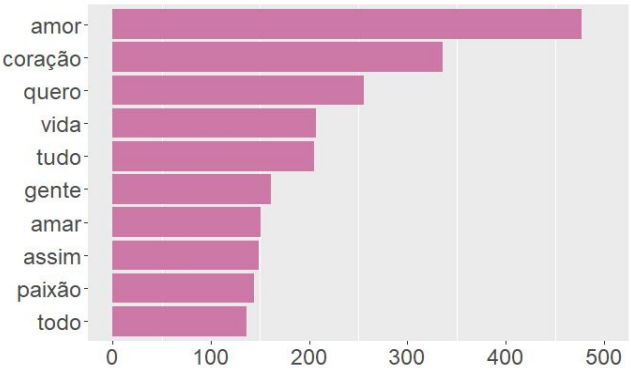

Figura 4: Quantidade de palavras mais utilizadas na década de 90

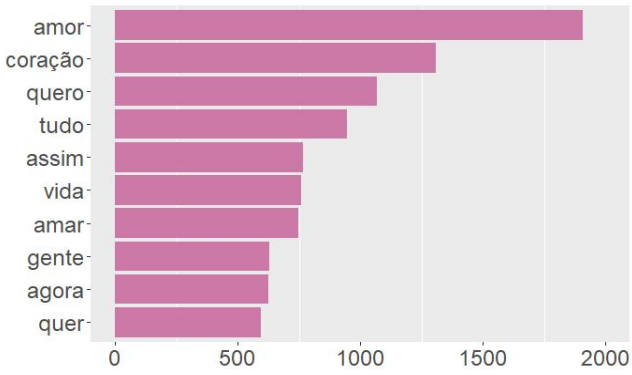

Figura 5: Quantidade de palavras mais utilizadas na década de 2000

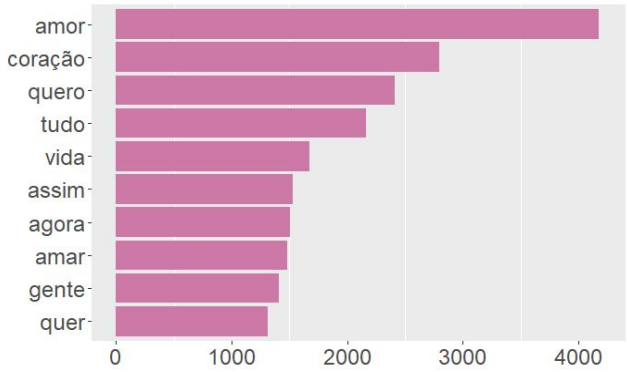

Figura 6: Quantidade de palavras mais utilizadas na década de 2010

A partir da década de 70, podemos observar quais são as palavras mais utilizadas no decorrer do tempo. Tais palavras, como a palavra "amor" ou "coração", ambas se mantêm dentro do conteúdo das letras independente do artista ou década. Essa observação não demonstra uma mudança significativa em relação a quantidade de palavras mais utilizadas no decorrer das décadas a partir dos anos 70.

Por outro lado, ao analisarmos a década de 60, podemos observar o uso da palavra "sertão". Este é um dado interessante ao considerar a época em que essas letras foram escritas. O forró começou a popularizar-se a partir da década de 50 [10] e considerando que o contexto das composições dessa época carregavam temas como a seca, o sertão etc, podemos ligar o uso dessa palavra ao contexto geral da década. Mas, temos que levar em consideração o fato de que temos poucos dados relacionados às décadas de 50 e 60.

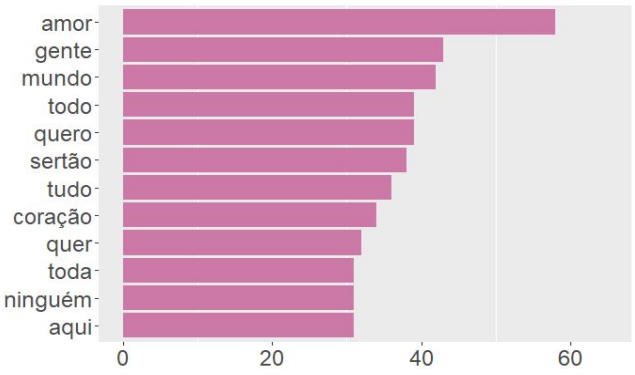

Figura 8: Quantidade de palavras mais utilizadas na década de 60

Aplicando a mesma metodologia tutorada por Debbie Liske [13] em seus métodos para análise lírica, utilizando agrupamento, o gráfico abaixo demonstra as palavras atemporais. O seja, palavras que permanecem em uso independente do decorrer do tempo. Para analisar de forma mais contundente este aspecto do estudo, um estudo mais detalhado sobre o que ocorria em cada década em termos de "assuntos" mais populares, pode auxiliar no entendimento deste ponto. Porém, analisando de forma superficial, seremos conduzidos a crer que os as palavras mais populares nos gráficos anteriores aparecem aqui no topo de cada gráfico.

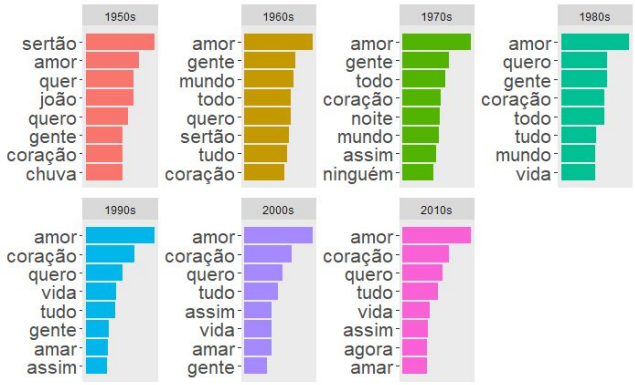

Figura 9: Ranqueamento de palavras atemporais

5.3 Tamanho das palavras e letras ao longo do tempo

Um parâmetro interessante para destacar nesse estudo é a quantidade de palavras levando em consideração o tamanho das palavras. O tamanho das palavras pode indicar a complexidade nas letras, levando em consideração que quanto maior a palavra mais complexa ela é.

Nesse ponto, observamos que a quantidade de palavras é inversamente proporcional ao tamanho das palavras em cada letra de música. Ou seja, quanto maiores são as palavras, menor é a quantidade, ou o contrário. Os gráficos abaixo, demonstram esse ponto por década. É possível perceber que em todas as décadas esse comportamento se repete.

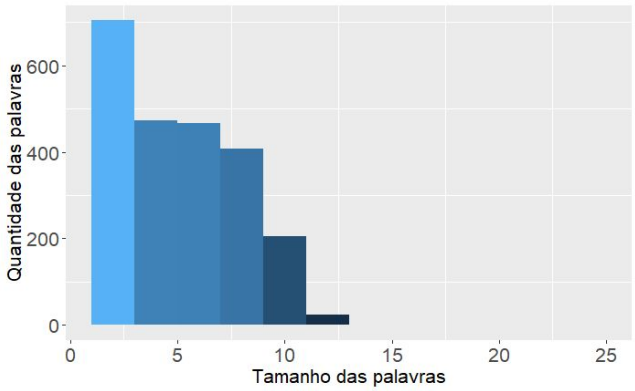

Figura 11: Distribuição tamanho/quantidade das palavras na década de 60

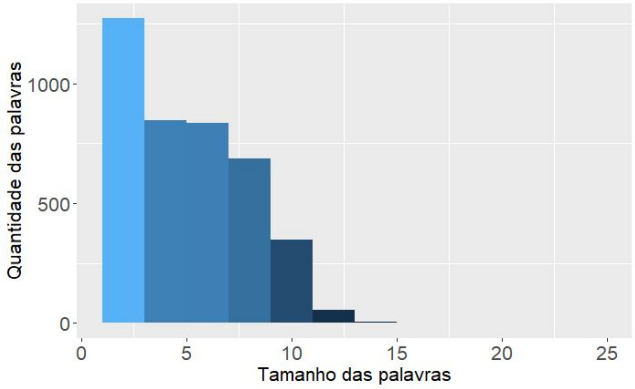

Figura 12: Distribuição tamanho/quantidade das palavras na década de 70

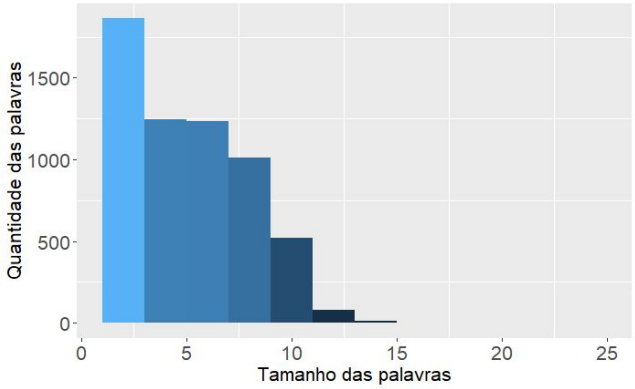

Figura 13: Distribuição tamanho/quantidade das palavras na década de 80

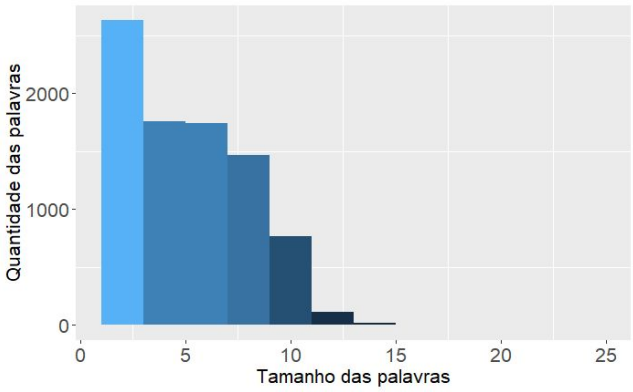

Figura 14: Distribuição tamanho/quantidade das palavras na década de 90

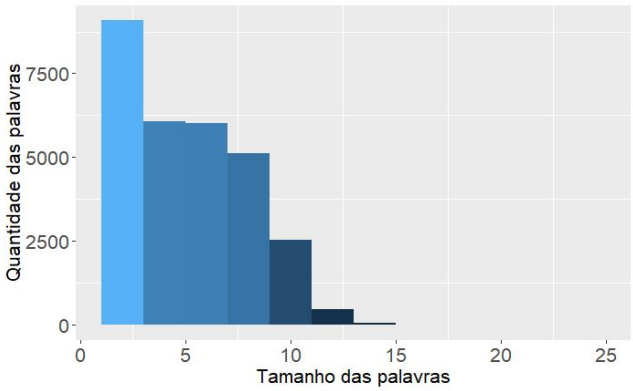

Figura 15: Distribuição tamanho/quantidade das palavras na década de 2000

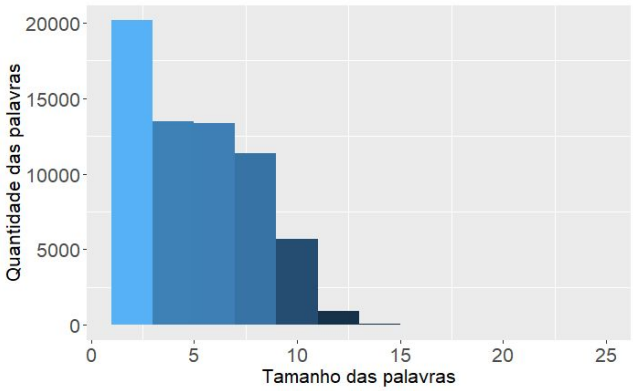

Figura 16: Distribuição tamanho/quantidade das palavras na década de 2010

5.4 Densidade e Diversidade Lexical

Como foi descrito mais acima, a quantidade de letras passa por um crescimento acentuado no decorrer do tempo no corpus em estudo. O gráfico abaixo pode demonstrar o elevado aumento na concentração de letras que é indicado pelas bolhas do gráfico. Podemos observar um leve decaimento na densidade do vocabulário ao longo do tempo, mesmo com o número crescente de letras dos dados.

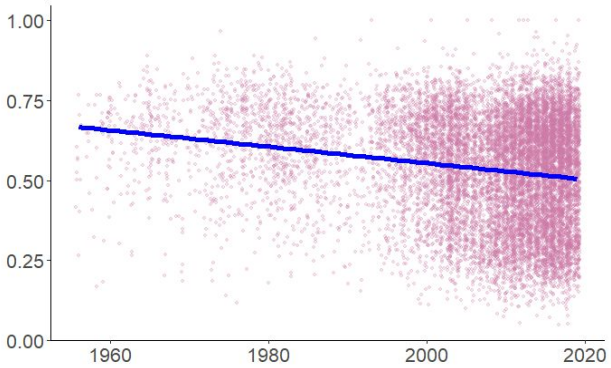

Figura 17: Densidade lexical

Mesmo que essa avaliação da densidade léxica, possa envolver fatores subjetivos, o gráfico nos induz a noção de que, com o passar do tempo, mais letras repetiram palavras e ficaram mais "pobres" em termos de seu vocabulário, sendo assim, tornaram-se mais repetitivas.

Em termos de diversidade de palavras, o gráfico abaixo demonstra a média de palavras únicas ao longo do tempo, reforçando a avaliação sobre a densidade indicada no gráfico da figura 17. Considerando que o conjunto de dados em questão possui um crescimento acentuado de letras no decorrer das décadas e isso pode ser visualizado no gráfico a seguir, apesar disso, podemos perceber que as letras não variaram consideravelmente no decorrer do tempo em relação a diversidade de suas palavras.

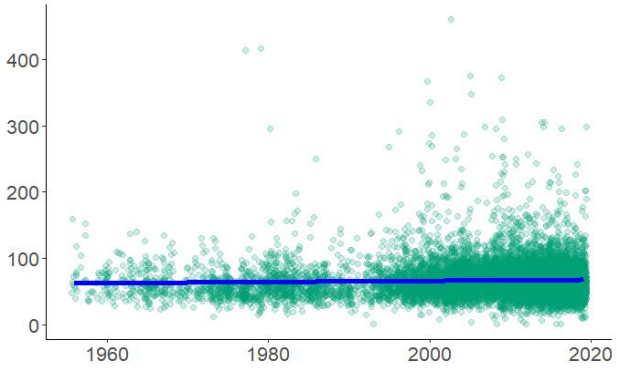

Figura 18: Diversidade lexical

5.5 Popularidade de palavras

Com o intuito de analisar a importância [11] das palavras, foi feita uma análise sobre a frequência com a qual as palavras mais aparecem por letra de cada música em relação ao número de letras de músicas que mais contêm cada uma dessas palavras. Ou seja, a maior quantidade de palavras que aparecem em menos letras de músicas do corpus em questão.

A frequência é calculada por TF - IDF = TF * IDF, onde IDF é a inversa de DF, 1 / DF e sua descrição é a seguinte:

- TF: Frequência de palavras mais repetidas por letra
- DF: Frequência de letras que contém cada palavra
- IDF: Frequência inversa por letra

A partir desse cálculo, as palavras mais comuns devem ter seu IDF e TF * IDF com o valor zero. A tabela a seguir demonstra parte dos resultados dos cálculos.

Década	Palavra	Quantidade	TF	IDF	TF_IDF
2010s	você	4842	0.0152162	0	0
2010s	amor	4176	0.0131232	0	0
2010s	mais	3396	0.0106721	0	0
2010s	coração	2797	0.0087897	0	0
2010s	quero	2408	0.0075672	0	0
2010s	minha	2351	0.0073881	0	0

Figura 19: Tabela de exemplo de cálculo de tf-idf

Para melhor entendermos e visualizarmos a disposição da popularidade das palavras, podemos observar no gráfico abaixo quais as palavras mais populares por década. Esse ponto da análise é expressivamente importante para termos uma noção clara das palavras mais utilizadas por década no sentido de visualizarmos quais foram essas palavras e permitir uma comparação relacionada ao contexto vivido em cada década em estudos mais profundos.

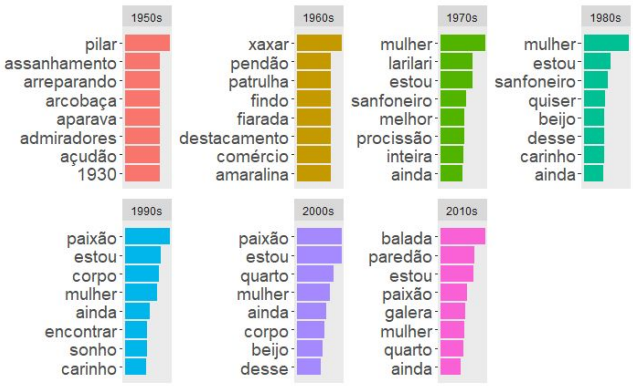

Figura 20: Palavras mais populares por década

6. DISCUSSÃO

O maior esforço neste trabalho foi empregado na busca pelos dados. Como é possível observar, o principal intuito deste trabalho, além de fazer uma análise inicial sobre os dados ao longo do tempo, é possibilitar que mais estudos sejam feitos e que mais pesquisadores tenham um ponto de partida para tomar como base para mais aprofundamentos sobre esta temática.

Um ponto que devemos ressaltar é a escassez de dados para a análise, principalmente nas décadas de 50 e 60. Apesar desse problema, ainda é possível extrair um entendimento razoável em relação ao processamento que foi feito. Podemos perceber que, apesar de entendermos que fazer análise em cultura e arte em geral é um processo subjetivo, a análise feita neste trabalho reforça a importância do uso de técnicas de processamento e mineração de dados para entender melhor como letras de músicas são compostas e quais fatores contribuem para isso.

Esse tipo de análise também pode ser útil para outros interesses além do entendimento da formação das composições. Como em alguns trabalhos que foram citados, existe um esforço para identificar pontos específicos que indiquem quais parâmetros e variáveis contribuem para que músicas cheguem ao sucesso, seja por tipos de palavras utilizadas, seus tamanhos e até mesmo seus significados. Isso pode ser bastante útil para compositores, por exemplo, que baseiam-se no contexto o qual está inserido para buscar palavras e inspirações para compor.

Examinando de forma pessoal a mudança das palavras ao decorrer do tempo, entendo que esse estudo expressa bem a ideia de tentar descrever as mudanças que ocorrem durante o tempo nas músicas através de suas letras. Em especial, a figura 20, demonstra de forma bastante clara o ”contexto” vivido em cada época. Dos anos 50 aos anos 80, palavras como “xaxar” e “sanfoneiro” estão ligadas aos costumes da época e ao forró mais “tradicional” no sentido de raízes do gênero. A partir dos anos 90, o contexto passa a ser um pouco mais relacionado a sexualidade e termos que relacionam ao comportamento mais jovem. Por fim, pude perceber que mesmo com todas as mudanças que ocorreram ao longo do tempo, a palavra mulher aparece com frequência em todas as décadas e isso reforça o entendimento de que as músicas do forró, em sua maioria, desde seu surgimento, possuem uma conotação diretamente ligada ao lado feminino, independente de que sua forma seja apreciativa ou depreciativa.

6.1 Experiência

Utilizar o conhecimento adquirido ao longo da graduação foi crucial para que fosse possível ter sucesso na raspagem dos dados. Em especial, ter conhecimento sobre técnicas e análises de algoritmos e entender o bom funcionamento de um banco de dados tornou possível a construção de algoritmos suficientemente eficientes e que possam ser utilizados por outros pesquisadores para dar continuidade a este trabalho. Também é preciso destacar o conhecimento advindo das disciplinas de probabilidade, estatística e metodologia científica que tornaram possível uma melhor interpretação dos resultados da análise sobre os dados.

6.2 Trabalhos futuros

A escassez de dados observada, pode ser tomada como ponto de partida para ferramentas ou pesquisas que consigam minerar mais informações para gêneros musicais. Este trabalho possibilitou o entendimento de que não existem muitos dados referentes a músicas que tenham forma estruturada relativas a meta-dados e sejam confiáveis. Esse ponto deixa em aberto essa problemática para que novos pesquisadores busquem por soluções para esse problema.

Também, podemos observar que é possível aplicar a mesma metodologia para outros gêneros musicais para não somente entendê-los, mas contribuir com sua história ao longo do tempo.

Principalmente no brasil, a mistura de gêneros musicais, momentos e condições históricas é expressiva e isso pode ser visto como um bom caso de estudo com a mesma forma e metodologia deste trabalho.

A análise que foi feita, focou apenas na parte textual das composições. Uma forma de contribuir com essa análise para acrescentar em sua acurácia é analisar o áudio das composições [12] e unir os resultados a análise textual das mesmas.

Por fim, existe a necessidade de produzir uma análise de sentimentos neste trabalho, porém pelo nível de complexidade este ponto talvez deva ser tratado como um novo estudo. É importante destacar que uma análise de sentimentos relativos às letras pode ser crucial para acrescentar a este estudo. Esse tipo de análise pode contribuir também com os compositores que querem entender melhor que tipo de sentimento suas composições e composições em geral podem causar para tomar como base. Análise de sentimentos pode ser feita também em áudios relacionados a cada composição para somar a análise textual.

7. AGRADECIMENTOS

Primeiro, agradeço a todos os professores e funcionários do curso de Ciência da Computação da Universidade Federal de Campina Grande por proporcionarem uma formação sólida e agradável além de todo o apoio no decorrer da graduação. Agradeço também ao meu colega de curso Helder Machado por todo o apoio e horas de conversa sobre o trabalho em questão e também o apoio acadêmico e pessoal durante a graduação. Agradeço ao meu orientador Nazareno Andrade por ter apoiado toda a ideia deste trabalho e pela orientação. Por fim, agradeço ao professor João Arthur Brunet Monteiro por todas as indicações de leitura na disciplina de Projeto de Trabalho de Conclusão de Curso e também durante a graduação.

8. REFERÊNCIAS

[1] SALES, Leo. ANÁLISE DA MÚSICA BRASILEIRA PARTE 1. Disponível em: https://leosalesblog.wordpress.com/2017/04/21/analise-da-musica-brasileira-parte-1/. Acesso em: 19 de Novembro de 2019.

[2] ABRAHAM, Tony. KOUL, Nikhita. MORALES, Joe. R.A.P - RAP ANALYSIS PROJECT. Disponível em: http://people.ischool.berkeley.edu/~nikhitakoul/capstone/index.html. Acesso em: 19 de Novembro de 2019.

[3] DANIELS, Matt. THE LARGEST VOCABULARY IN HIP HOP. Disponível em: https://pudding.cool/projects/vocabulary/index.html. Acesso em 19 de Novembro de 2019.

[4] SALES, Leo. ANÁLISE DA MÚSICA BRASILEIRA PARTE 3. Disponível em: https://leosalesblog.wordpress.com/2017/04/26/analise-da-musica-brasileira-parte-3/. Acesso em: 19 de Novembro de 2019.

[5] TÓTH, Bence. FROM METALLICA TO ADELE - TEXT ANALYSIS OF SUCCESSFUL SONG LYRICS WITH R. Disponível em: https://towardsdatascience.com/text-analysis-of-successful-song-lyrics-e41a4ccb26f5. Acesso em: 19 de Novembro de 2019.

[6] WIKIPEDIA. NATURAL LANGUAGE PROCESSING. Disponível em: https://en.wikipedia.org/wiki/Natural_language_processing. Acesso em: 19 de Novembro de 2019.

[7] OLIVEIRA. Nicácio. TCC WEB SCRAPER. Disponível em: https://github.com/nicacioliveira/tcc-webscrapper. Acesso em: 19 de Novembro de 2019.

[8] WIKIPEDIA. HEADLESS BROWSER. Disponível em: https://en.wikipedia.org/wiki/Headless_browser. Acesso em: 19 de novembro de 2019

[9] OLIVEIRA. NICÁCIO. ANÁLISE E VISUALIZAÇÃO DE DADOS SOBRE A EVOLUÇÃO DAS LETRAS DO FORRÓ. Disponível em: https://github.com/nicacioliveira/tcc-analise-forro. Acesso em: 19 de Novembro de 2019.

[10] WIKIPEDIA. FORRÓ. Disponível em: https://pt.wikipedia.org/wiki/Forr%C3%B3. Acesso em 19 de Novembro de 2019.

[11] WIKIPEDIA. TF-IDF. Disponível em: https://pt.wikipedia.org/wiki/Tf%E2%80%93idf. Acesso em 19 de Novembro de 2019.

[12] CHINOY, Sahil. MA. Jessia. WHY SONGS OF THE SUMMER SOUND THE SAME. Disponível em: https://www.nytimes.com/interactive/2018/08/09/opinion/do-songs-of-the-summer-sound-the-same.html. Acesso em: 19 de Novembro de 2019.

[13] LISKE, Debbie. LYRIC ANALYSIS WITH NLP & MACHINE LEARNING WITH R. Disponível em: https://www.datacamp.com/community/tutorials/R-nlp-machine-learning. Acesso em: 19 de Novembro de 2019.