

Modelagem de tópicos através de Latent Dirichlet Allocation (LDA)

Adriana Eva Fernandes da Silva

September 23, 2024

1 Introdução

Resumir e organizar grandes volumes de informações pode ser uma tarefa difícil. A Modelagem de Tópicos facilita essa tarefa, uma vez que é projetada para identificar e explorar temas e suas conexões dentro de grandes conjuntos de dados (de Souza, 2019).

A Latent Dirichlet Allocation (LDA) é uma técnica de modelagem de tópicos amplamente utilizada em aprendizado não supervisionado para descobrir padrões não-observáveis em um conjunto de documentos textuais. Ela busca identificar tópicos latentes dentro dos textos sem a necessidade de rótulos pré-estabelecidos. Um tópico é uma distribuição de probabilidade sobre palavras. Ele nos fornece as probabilidades de um conjunto de palavras para o tópico dado. A intuição por trás da LDA é que os documentos são resumidos em vários tópicos. (Blei et al., 2012).

Neste trabalho, propomos uma análise detalhada do modelo LDA e sua aplicação em letras de músicas. O objetivo é demonstrar como essa técnica pode ser aplicada para identificar tópicos em um conjunto de letras de músicas.

2 LDA

O número de palavras pode ser muito grande para que vários documentos sejam comparados diretamente, e o LDA auxilia na identificação de palavras que se referem a tópicos comuns. Esse método foi desenvolvido em 2003 pelos pesquisadores David Blei, Andrew Ng e Michael Jordan. (Blei et al. (2012))

Palavras que frequentemente aparecem juntas podem indicar um tópico comum. Por exemplo, ao encontrar um grupo de palavras associadas, pode-se inferir que elas pertencem a um determinado tópico. Assim, é viável comparar documentos e identificar quais tópicos são mais prevalentes em cada um utilizando um modelo probabilístico generativo junto com distribuições de Dirichlet.

A modelagem de tópicos é um método não supervisionado, o que significa que não requer dados rotulados. Ela pode ser aplicada diretamente a um conjunto de documentos de texto para extrair informações importantes.

Esse método opera de forma exploratória, identificando os temas presentes em um conjunto de dados textuais e descobre tópicos por meio de uma estrutura probabilística que infere os tópicos, que são variáveis latentes, com base nas palavras observadas nos documentos.

Duas suposições importantes feitas pelo LDA são:

- A suposição de bag of words (saco de palavras), significa que a ordem das palavras é ignorada nos documentos. Então, a informação importante para o modelo é o número de vezes que as palavras apareceram no documento, não a ordem das palavras
- A ordem dos documentos também é irrelevante.

2.1 Componentes do LDA

- Tópicos: São representações de grupos de palavras que frequentemente aparecem juntas em documentos e seu uso permite entender os assuntos sobre os quais os documentos se tratam, sem a necessidade de conhecer previamente cada um deles.
- Documentos: Um documento é qualquer texto em que deseja-se analisar para identificar tópicos, como artigos, revistas, músicas e respostas de questionários. Cada documento é visto como uma mistura probabilística de vários tópicos.

2.2 Algoritmo LDA

Para compreender melhor como os tópicos são extraídos e analisados a partir de grandes volumes de texto, vamos explorar o algoritmo Latent Dirichlet Allocation (LDA).

Passo 1

Atribua aleatoriamente um tópico a cada palavra em cada documento. Após essa atribuição aleatória, duas frequências podem ser calculadas:

- A contagem de tópicos em cada documento (distribuição de frequência dos tópicos).
- A contagem de palavras em cada tópico (distribuição de frequência das palavras).

Passo 2

1. Escolha uma palavra em um documento.
2. Cancele a atribuição do tópico previamente designado a essa palavra (ou seja, o tópico atribuído aleatoriamente na inicialização).
3. Reatribua um tópico à palavra com base nas seguintes considerações: A popularidade de cada tópico no documento (frequência do tópico), medida pelas contagens de frequência calculadas na inicialização, e uma distribuição multinomial gerada por Dirichlet sobre os tópicos em cada documento.
4. A popularidade da palavra em cada tópico (frequência da palavra), medida pelas contagens de frequência calculadas na inicialização, e uma distribuição multinomial gerada por Dirichlet sobre as palavras em cada tópico.
5. Multiplique as probabilidades resultantes de 3 e 4 para obter a probabilidade condicional de que a palavra pertença a cada tópico.

6. Reatribua a palavra ao tópico com a maior probabilidade condicional. Repita a Etapa 2 para todas as palavras em todos os documentos.

Passo 3

Durante a Etapa 2, dois Dirichlets são utilizados. Esses Dirichlets desempenham um papel importante na generalização do modelo para novos documentos.

Eles são particularmente úteis quando a contagem de frequência de um tópico em um documento é zero, por exemplo, se um tópico não aparecer em um documento após a inicialização aleatória. O Dirichlet, sendo uma distribuição de probabilidade sobre a distribuição multinomial dos tópicos, gera uma probabilidade diferente de zero para esse tópico, permitindo que ele seja considerado em atualizações subsequentes de atribuição de tópicos.

Essa variabilidade adicional garante que todos os tópicos tenham a chance de serem considerados, o que pode melhorar a capacidade do modelo de representar novos documentos, não vistos anteriormente. Na Etapa 2 do algoritmo LDA, a probabilidade condicional é calculada com base em dois componentes principais: a distribuição de tópicos dentro de um documento e a distribuição de palavras dentro de um tópico.

Dessa forma, a atribuição de um tópico para cada palavra é influenciada pela probabilidade do tópico estar presente no documento e pela probabilidade da palavra aparecer dentro desse tópico. Isso faz com que as palavras se agrupem dentro de um tópico com base na "adequação" da palavra para o tópico e na "adequação" do tópico para o documento, considerando todas as outras atribuições de tópicos para as demais palavras em todos os documentos.

Vale ressaltar que, nesse contexto, a "adequação" é determinada apenas por contagens de frequência e distribuições de Dirichlet, sem levar em conta informações semânticas. Portanto, a estrutura observada dos documentos orienta a descoberta de relações latentes e, conseqüentemente, a identificação da estrutura latente dos tópicos. (ISHI.PRO (2024))

2.3 Parâmetros do modelo

Escolher o número de tópicos K com antecedência é uma etapa fundamental no LDA. Ao definir K , estamos assumindo que acreditamos que o conjunto de documentos que estamos analisando pode ser descrito por K tópicos.

Outro ponto a observar no LDA é que, uma vez identificados os tópicos K , o modelo não nos fornece informações adicionais sobre o conteúdo dos tópicos além da distribuição de palavras associadas a cada um (ou seja, a probabilidade de cada palavra no vocabulário aparecer em um tópico específico).

Portanto, é necessário usar nossa própria interpretação para entender o que cada tópico representa e atribuir um nome significativo a cada um deles.

Cada documento tem uma distribuição de tópicos que é amostrada a partir de uma distribuição de Dirichlet com um parâmetro α , que indica quantos tópicos compõem os documentos, permitindo uma distribuição de tópicos por documento que pode ser mais detalhada ou mais concentrada — quanto maior for o valor, mais tópicos estarão presentes e mais detalhada será a distribuição de tópicos em cada documento.

Cada tópico tem uma distribuição de palavras que é amostrada a partir de uma distribuição de Dirichlet com um parâmetro β que indica quantas palavras compõem os tópicos, permitindo uma distribuição de palavras por tópico que pode ser mais detalhada ou mais concentrada — quanto maior for o valor, mais palavras estarão presentes e mais detalhada será a distribuição de palavras em cada tópico. O número de tópicos (K) pode influenciar a qualidade dos tópicos encontrados.

2.4 Distribuição de Dirichlet

2.4.1 Definição Matemática

A distribuição de Dirichlet é uma distribuição multivariada para variáveis contínuas que são não-negativas e cuja soma é igual a 1.

Dizemos que θ tem distribuição Dirichlet com parâmetro $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$, onde $\alpha_i \geq 0$, e escrevemos $\theta \sim \text{Dirichlet}(\alpha)$ se a função densidade de probabilidade é dada por:

$$f(\theta) = \frac{1}{\beta(\alpha)} \left(\prod_{i=1}^d \theta_i^{\alpha_i-1} \right) I \left(\sum_{i=1}^d \theta_i = 1 \right) \quad (1)$$

onde $\beta(\alpha)$ é definida como:

$$\beta(\alpha) = \frac{\prod_{i=1}^d \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^d \alpha_i\right)} \quad (2)$$

Os parâmetros α determinam a forma da distribuição. Se todos os parâmetros são iguais, a distribuição é simétrica. Se os parâmetros são diferentes, a distribuição é assimétrica.

3 Aplicação em letras de músicas

O modelo LDA foi aplicado em 8662 letras de músicas de diversos gêneros e cantores, para identificar tópicos baseados nas palavras que frequentemente aparecem juntas nas letras. Cada tópico representa um conjunto de palavras que têm alta probabilidade de ocorrer juntas.

3.1 Processamento

Primeiramente, realizamos a **tokenização**, que envolveu a divisão das letras das músicas em palavras. Em seguida, criamos uma **matriz de termos**, onde cada linha representa uma música e cada coluna representa uma palavra. Para a modelagem, foi escolhido $K=15$ tópicos.

Distribuição de Termos por Tópico

Table 1: Tópico 1

| Termo | β |
|---------|---------|
| amor | 0.1718 |
| quero | 0.0599 |
| amar | 0.0433 |
| coração | 0.0363 |
| meu | 0.0253 |
| paixão | 0.0233 |
| faz | 0.0202 |
| corpo | 0.0199 |
| prazer | 0.0187 |
| beijo | 0.0185 |

Table 2: Tópico 2

| Termo | β |
|--------|---------|
| vou | 0.1686 |
| pra | 0.1159 |
| agora | 0.0395 |
| fazer | 0.0314 |
| hora | 0.0286 |
| lugar | 0.0276 |
| dar | 0.0252 |
| aqui | 0.0240 |
| comigo | 0.0201 |
| ficar | 0.0199 |

Table 3: Tópico 3

| Termo | β |
|--------|---------|
| mim | 0.0946 |
| você | 0.0849 |
| assim | 0.0820 |
| sei | 0.0610 |
| faz | 0.0460 |
| por | 0.0406 |
| diz | 0.0363 |
| fim | 0.0309 |
| sempre | 0.0202 |
| bem | 0.0201 |

Table 4: Tópico 4

| Termo | β |
|--------|---------|
| mar | 0.0282 |
| laiá | 0.0240 |
| quando | 0.0202 |
| menina | 0.0166 |
| olha | 0.0156 |
| pai | 0.0147 |
| pro | 0.0145 |
| mãe | 0.0119 |
| água | 0.0117 |
| que | 0.0115 |

Table 5: Tópico 5

| Termo | β |
|--------|---------|
| samba | 0.0952 |
| cantar | 0.0223 |
| povo | 0.0176 |
| mão | 0.0166 |
| sambar | 0.0152 |
| roda | 0.0133 |
| faz | 0.0130 |
| chegou | 0.0127 |
| som | 0.0122 |
| pagode | 0.0118 |

Table 6: Tópico 6

| Termo | β |
|---------|---------|
| vai | 0.1929 |
| ser | 0.0772 |
| bem | 0.0725 |
| pode | 0.0626 |
| ter | 0.0386 |
| vamos | 0.0223 |
| tudo | 0.0192 |
| ninguém | 0.0167 |
| quiser | 0.0150 |
| pra | 0.0148 |

Table 7: Tópico 7

| Termo | β |
|----------|---------|
| cara | 0.0173 |
| porque | 0.0146 |
| ele | 0.0143 |
| malandro | 0.0139 |
| fala | 0.0123 |
| então | 0.0099 |
| bicho | 0.0092 |
| pega | 0.0087 |
| mão | 0.0085 |
| bom | 0.0082 |

Table 8: Tópico 8

| Termo | β |
|---------|---------|
| amor | 0.0645 |
| coração | 0.0538 |
| dor | 0.0299 |
| chorar | 0.0270 |
| fez | 0.0254 |
| saudade | 0.0192 |
| foi | 0.0190 |
| quis | 0.0166 |
| sofrer | 0.0166 |
| fiz | 0.0150 |

Table 9: Tópico 9

| Termo | β |
|--------|---------|
| para | 0.0172 |
| são | 0.0145 |
| brasil | 0.0136 |
| morro | 0.0136 |
| terra | 0.0123 |
| cidade | 0.0122 |
| onde | 0.0113 |
| senhor | 0.0111 |
| rei | 0.0110 |
| rio | 0.0103 |

Table 10: Tópico 10

| Termo | β |
|----------|---------|
| pra | 0.2656 |
| gente | 0.1057 |
| vem | 0.1003 |
| ver | 0.0567 |
| tem | 0.0252 |
| querendo | 0.0170 |
| chega | 0.0161 |
| quer | 0.0113 |
| novo | 0.0112 |
| ficar | 0.0101 |
| olha | 0.0101 |

Table 11: Tópico 11

| Termo | β |
|---------|---------|
| tudo | 0.0444 |
| tempo | 0.0424 |
| nada | 0.0325 |
| dizer | 0.0293 |
| sei | 0.0252 |
| tão | 0.0246 |
| preciso | 0.0242 |
| ainda | 0.0236 |
| tanto | 0.0202 |
| viver | 0.0196 |

Table 12: Tópico 12

| Termo | β |
|-------|---------|
| sol | 0.0265 |
| céu | 0.0261 |
| luz | 0.0217 |
| tão | 0.0212 |
| sonho | 0.0184 |
| olhos | 0.0164 |
| lua | 0.0162 |
| noite | 0.0130 |
| lindo | 0.0124 |
| flor | 0.0122 |

Table 13: Tópico 13

| Termo | β |
|--------|---------|
| não | 0.1400 |
| mas | 0.0901 |
| deixa | 0.0409 |
| nunca | 0.0365 |
| quem | 0.0353 |
| sabe | 0.0352 |
| sempre | 0.0293 |
| quando | 0.0206 |
| nem | 0.0164 |
| sou | 0.0162 |

Table 14: Tópico 14

| Termo | β |
|---------|---------|
| quer | 0.0413 |
| mulher | 0.0386 |
| quero | 0.0358 |
| ela | 0.0314 |
| ninguém | 0.0278 |
| não | 0.0253 |
| casa | 0.0221 |
| noite | 0.0206 |
| toda | 0.0186 |
| coisa | 0.0162 |

Table 15: Tópico 15

| Termo | β |
|-------|---------|
| que | 0.1960 |
| vida | 0.0737 |
| dia | 0.0592 |
| mundo | 0.0521 |
| deus | 0.0481 |
| todo | 0.0429 |
| bom | 0.0324 |
| meu | 0.0273 |
| feliz | 0.0244 |
| paz | 0.0174 |

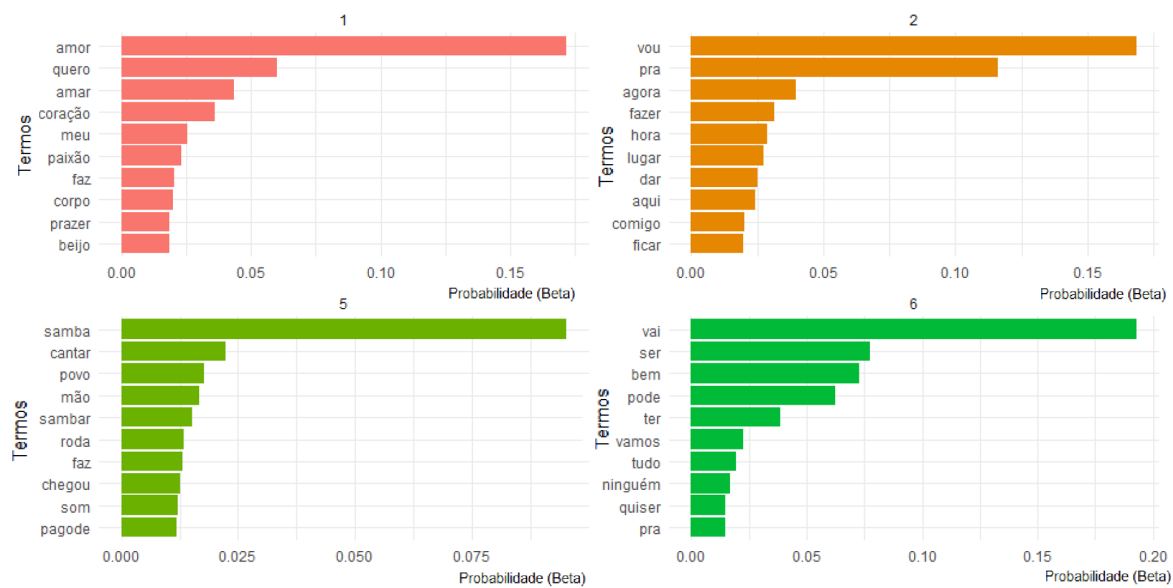


Figure 1: Distribuição dos tópicos 1, 2, 5 e 6 entre os termos

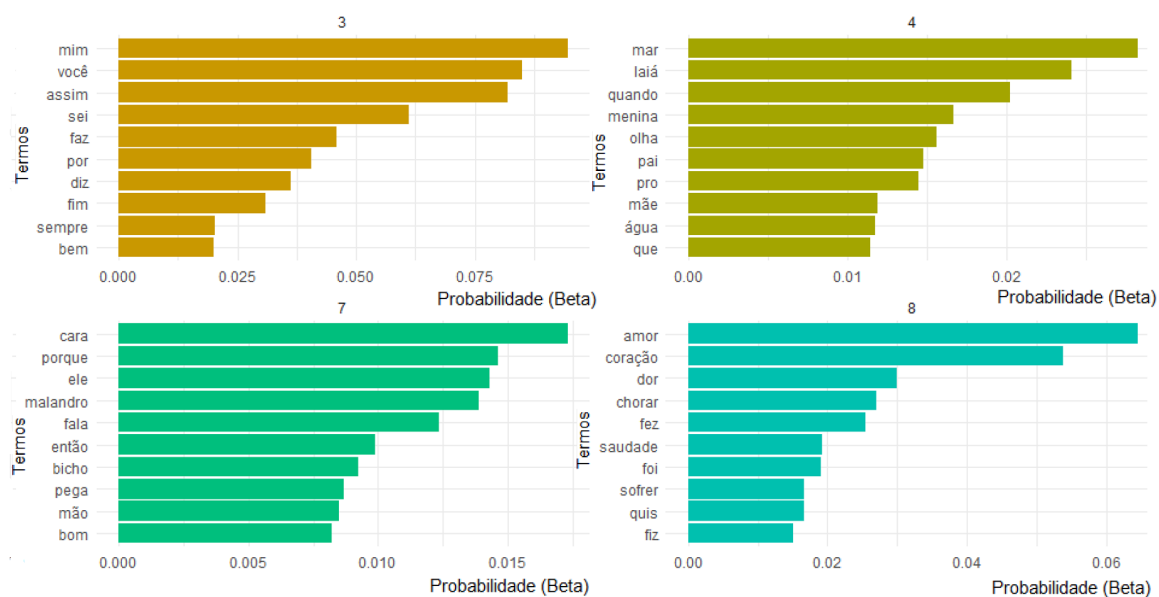


Figure 2: Distribuição dos tópicos 3, 4, 7 e 8 entre os termos

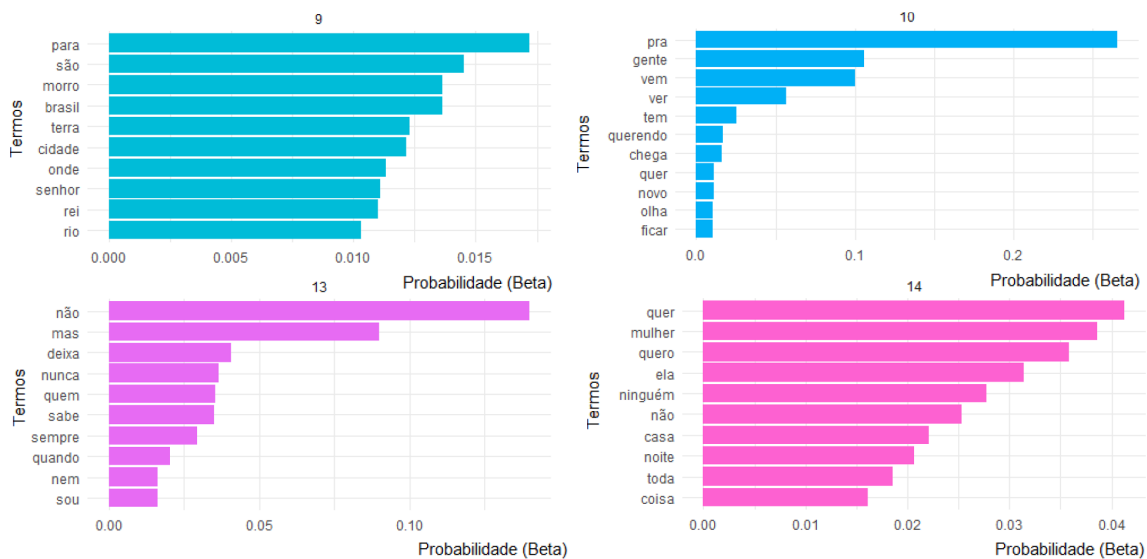


Figure 3: Distribuição dos tópicos 9, 10, 13 e 14 entre os termos

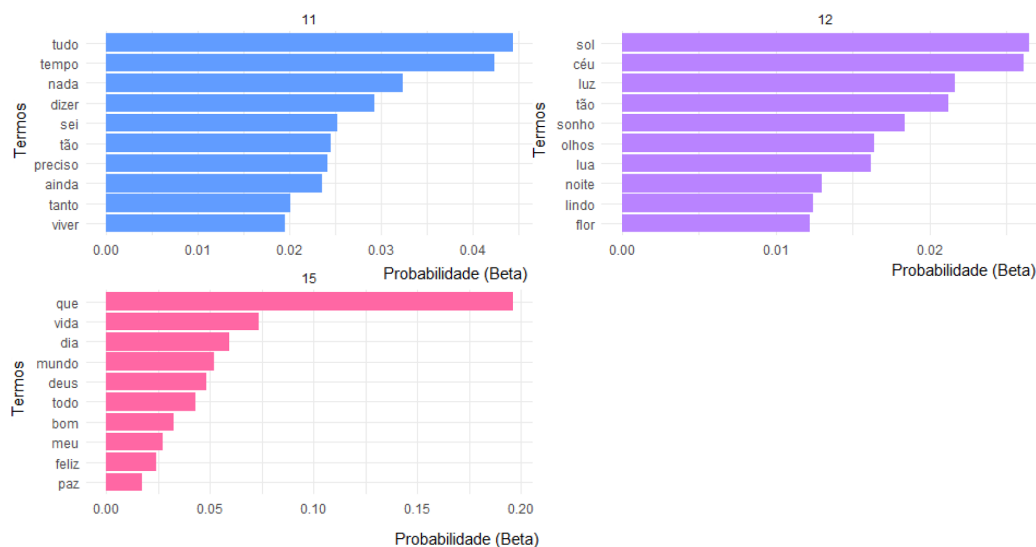


Figure 4: Distribuição dos tópicos 11, 12 e 15 entre os termos

Com base nos resultados dos gráficos e das tabelas, vemos que o **primeiro tópico** é voltado para temas românticos e afetivos, com palavras como “amor” e “coração” predominando, indicando que essas músicas falam sobre romance. O **segundo tópico** gira em torno de ação e mudança, refletido em termos como “vou” e “agora”, indicando que as músicas abordam decisões sobre o que fazer e onde estar. O **terceiro tópico** lida com conversas e relacionamentos interpessoais, com palavras como “mim” e “você”, indicando diálogos e relações pessoais nas músicas. O **quarto tópico** envolve músicas que falam sobre o dia a dia e a natureza, com temas familiares e pessoais. O **quinto tópico** celebra a música e a dança, com destaque para palavras como “samba” e “cantar”. O **sexto tópico** é sobre pensar no futuro e fazer planos, com músicas que falam sobre esperanças e desejos. O **sétimo tópico** é mais informal, apresentando palavras como “cara”, “malandro” e “bicho”. O **oitavo tópico** aborda

sentimentos de dor e saudade. O **nono tópico** trata de temas relacionados ao Brasil, incluindo aspectos geográficos e culturais. O **décimo tópico** reflete músicas que celebram a interação social e experiências compartilhadas. O **décimo primeiro tópico** lida com conceitos de tempo e necessidade, usando palavras como “tempo” e “preciso”, refletindo sobre a vida e o tempo. O **décimo segundo tópico** explora temas de natureza e beleza. O **décimo terceiro tópico** envolve negação e resistência. O **décimo quarto tópico** é focado em termos como “quer” e “mulher”, refletindo relacionamentos e desejos pessoais. E o **décimo quinto tópico** fala sobre temas positivos e universais, celebrando aspectos da vida, como a felicidade, com uma visão otimista.

Essa análise mostrou uma variedade de temas e estilos nas letras das músicas. O modelo LDA ajudou a sintetizar essas informações, destacando os principais tópicos e as emoções predominantes, proporcionando uma visão mais clara do conteúdo e das emoções expressas nas músicas.

Em seguida, são apresentadas as relações entre alguns documentos e tópicos.

| Música | Tópico 1 | Tópico 2 | Tópico 3 | Tópico 4 |
|--------------------------------------|----------|----------|----------|----------|
| Tá no Dna / Música Incidental: o Rap | 0,05 | 0,02 | 0,09 | 0,02 |
| Tchuí, Tchuí | 0,07 | 0,04 | 0,04 | 0,10 |
| Te Desejo (Minha Inha) | 0,20 | 0,11 | 0,08 | 0,08 |
| Tic Nervoso (Part. Anitta) | 0,03 | 0,06 | 0,14 | 0,04 |
| Tira A Mão de Bolso | 0,05 | 0,08 | 0,03 | 0,07 |

Table 16: Distribuição dos tópicos 1 a 4 para cada música.

| Música | Tópico 5 | Tópico 6 | Tópico 7 | Tópico 8 |
|--------------------------------------|----------|----------|----------|----------|
| Tá no Dna / Música Incidental: o Rap | 0,60 | 0,01 | 0,03 | 0,02 |
| Tchuí, Tchuí | 0,13 | 0,04 | 0,07 | 0,03 |
| Te Desejo (Minha Inha) | 0,03 | 0,05 | 0,04 | 0,05 |
| Tic Nervoso (Part. Anitta) | 0,29 | 0,14 | 0,05 | 0,03 |
| Tira A Mão de Bolso | 0,14 | 0,07 | 0,15 | 0,03 |

Table 17: Distribuição dos tópicos 5 a 8 para cada música.

| Música | Tópico 9 | Tópico 10 | Tópico 11 | Tópico 12 |
|--------------------------------------|----------|-----------|-----------|-----------|
| Tá no Dna / Música Incidental: o Rap | 0,03 | 0,02 | 0,01 | 0,02 |
| Tchuí, Tchuí | 0,08 | 0,08 | 0,05 | 0,03 |
| Te Desejo (Minha Inha) | 0,03 | 0,03 | 0,15 | 0,03 |
| Tic Nervoso (Part. Anitta) | 0,03 | 0,03 | 0,03 | 0,02 |
| Tira A Mão de Bolso | 0,04 | 0,08 | 0,08 | 0,05 |

Table 18: Distribuição dos tópicos 9 a 12 para cada música.

| Música | Tópico 13 | Tópico 14 | Tópico 15 |
|--------------------------------------|-----------|-----------|-----------|
| Tá no Dna / Música Incidental: o Rap | 0,02 | 0,02 | 0,03 |
| Tchuí, Tchuí | 0,06 | 0,09 | 0,10 |
| Te Desejo (Minha Inha) | 0,03 | 0,04 | 0,05 |
| Tic Nervoso (Part. Anitta) | 0,02 | 0,06 | 0,03 |
| Tira A Mão de Bolso | 0,04 | 0,04 | 0,04 |

Table 19: Distribuição dos tópicos 13 a 15 para cada música.

Música:

Pois é, quem nasce com a onda não tem jeito. Explode de amor dentro do peito. Toda vez que ouve o Harmonia tocar. Já vem com molejo na cintura. O gingado do corpo é uma loucura. É uma loucura, então. O **pagode** no Dna do povo. Faz a galera **cantar**, faz a galera dançar, faz a galera subir. Faz a galera descer, faz a galera mexer. Eta o **pagode**, o **pagode**, o **pagode** no Dna. Faz a galera **cantar**, faz a galera dançar, faz a galera subir. Faz a galera descer, faz a galera mexer. Não tem jeito a gente gosta, e assume o nosso amor. É tão forte o sentimento. A força da nossa batida que já se espalhou. O **pagode**, o **pagode**, o **pagode**, o **pagode**, o que?. Faz a galera **cantar**, faz a galera dançar, faz a galera subir. Faz a galera descer, faz a galera mexer. Eta o **pagode**, o **pagode**, o **pagode** no Dna. Faz a galera **cantar**, faz a galera dançar, faz a galera subir. Faz a galera descer, faz a galera mexer. . O **samba** duro eu aprendi desde menino. Da Liberdade, há Capelinha eu fui **sambar**. Cá entre nós, era bonito o Gera **samba**. A legião se reunia pra quebrar. Faz a galera descer, faz a galera mexer. Então mãe. O **pagode**, o **pagode**, o **pagode** no Dna. Faz a galera **cantar**, faz a galera dançar. Faz a galera subir (tutututu pá). Faz a galera descer, faz a galera mexer. No meu **pagode** dança pobre, dança rico. Até gringo de arriscando a **sambar**. Chora cavaco e ela vem toda se bulindo. Eu já falei aqui tá no Dna. Faz a galera descer, faz a galera mexer. O **pagode**, o **pagode**, o **pagode** no Dna. Faz a galera **cantar**, faz a galera dançar, faz a galera subir. Faz a galera descer, faz a galera mexer

4 Conclusão

Ao analisarmos as tabelas de Documentos/Tópicos, notamos uma conexão muito clara entre o tópico 5 e a música 'Tá no Dna / Música Incidental: o Rap'. Isso fica evidente quando olhamos para as palavras que aparecem frequentemente na letra da música e como elas se encaixam com os termos do tópico 5.

O LDA demonstrou um bom desempenho ao identificar não apenas as palavras que costumam aparecer juntas, mas também ao capturar temas presentes nas letras das músicas.

References

- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2012). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- de Souza, M. (2019). Resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina. 9(1):3–9.
- ISHI.PRO (2024). Modelagem de Tópico com LDA: Uma Explicação Intuitiva. Accessed: 2024-09-03.

A Apêndice

Códigos utilizados

```
dados <- read.csv("C:/Users/JAVA/Desktop/letras-ptbr-samba-grande.csv")

set.seed(1)
library(data.table)
library(tm)
library(tidyverse)
library(dplyr)
library(stopwords)
library(topicmodels)
library(broom)
library(writexl)
library(tidytext)
library(ggpubr)
library(tidytext)
```

```

library(dplyr)
library(tidyr)

corp<- VCorpus(VectorSource(dados$Lyric))

stopword_pt <- stopwords("pt")
corpus_base <- tm_map(corp, removeWords, stopword_pt)

dtm <- DocumentTermMatrix(corpus_base,
                           control = list(tolower = TRUE,
                                           removeNumbers = TRUE,
                                           removePunctuation = TRUE,
                                           stripWhitespace = TRUE,
                                           bounds = list(global = c(15, Inf))))

dtm
dtm.matrix<- dtm%>% as.matrix()

modelo_Ida <- LDA(dtm, k = 15, method = "Gibbs", control = list(seed = 1234))

beta_topics <- tidy(modelo_Ida, matrix = "beta")

beta_top_terms <- beta_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

grupos <- beta_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered() +
  labs(x = "Probabilidade (Beta)", y = "Termos", title = "Top Termos por Tópicos") +
  theme_minimal()

documento_topico <- tidy(modelo_Ida, matrix = "gamma")

matriz_documento_topico <- gamma_document_topic %>%
  pivot_wider(names_from = topic, values_from = gamma)

head(matriz_documento_topico)

```