PROFESSUR FÜR
ANGEWANDTE STATISTIK
DER FREIEN UNIVERSITÄT BERLIN

ADVANCED STATISTICAL MODELING
SEMINAR PAPER

Application of LDA topic model to song lyrics

Silvia Ventoruzzo

Semester: Summer Term 2019
Responsible: Prof. Dr. Timo Schmid, Dr. Natalia Rojas-Perilla
Supervisor: Nora Ulrike Würz

Submission: August 10, 2019

Contents

1. Introduction

Text mining is the specific data mining area which deals with text data. Because of the big amount of text documents coming from newspapers as well as social media (Silge & Robinson 2017), text mining is seen as increasingly important for "knowledge discovery" (A.-H. Tan et al. 1999).

In trying to understand the large collection of texts that are present nowadays, one would like to divide them into groups in order to separately comprehend their message (Silge & Robinson 2017). Topic modelling serves this purpose, being an unsupervised classification method which searches for natural groups of words in the documents, called topics (Silge & Robinson 2017).

A specific topic modelling algorithm is called Latent Dirichlet Allocation (LDA), which "treats each document as a mixture of topics, and each topic as a mixture of words" (Silge & Robinson 2017). This can be considered as overlapping clustering, meaning that "an object can simultaneously belong to more than one group" (P.-N. Tan 2018). This is appropriate to text data because it betters represents natural language (Silge & Robinson 2017). In particular, LDA is based on the bag-of-words exchangeability assumption, meaning that the order of the words in the documents is not important (Blei et al. 2003).

In this project LDA was applied to a corpus of 200 song texts from multiple artists of different music genres to search for the underlying topics and to analyze if there are various topics within genres or artists. It was supposed at first that topics would be quite distinct, since the songs are from different artists and genres, and that the songs for an artist could nonetheless involve numerous topics. However, it becomes clear already from the initial exploratory data analysis that the songs for different artists and genres have quite a lot in common. In fact, it has been discovered that the songs do contain different topics, but that these are quite similar among each other. In fact, only Hip-Hop songs are related to almost only one topic, while the other genres and artists have more contrasting song texts.

The paper starts with chapter 2, which describes the theory behind LDA and the preparation steps. Subsequently, in chapter 3 the dataset used and the implementation of LDA in `R` is described. Afterwards, chapter 4 displays the results of the application of LDA on the dataset and the findings from those. Finally, chapter 5 outlines the conclusions.

2. Theory

Topic models are statistical models used to discover the intrinsic "topics", i.e. "semantic structure", of a collection of documents (Blei & Lafferty 2009).

In this paper the focus is on the topic model called Latent Dirichlet Allocation (LDA). Latent Dirichlet Allocation derives its name from the fact that "characteristics of topics and documents are drawn from Dirichlet distributions" (Ponweiser 2012) and the presence of hidden elements, i.e. the topics. An explaination of the Dirichlet distribution can be found in the appendix A. Specifically, LDA is a "generative probabilistic model for collections of discrete data such as text corpora" (Blei et al. 2003).

2.1. Definitions

The algorithm can indeed be applied to other types of data, but it is mostly applied to text analysis, reason why explanations will be done with text-related terms. Blei et al. (2003) defines in fact the following elements:

- The corpus D is a collection of M documents: $D = \{d_1, ..., d_M\}$
- A document of the corpus d_j is a sequence of N_j words: $d_j = (w_1, ..., w_{N_j})$
- A word present in the corpus w_i is an item from the vocabulary, that is the set of all words present in the corpus (Ponweiser 2012): $w_i \in \{1, ..., V\}$

Therefore, the documents from the considered corpus can be seen as "random mixtures over latent topics", each of which has a distribution over words (Blei et al. 2003).

A list of the used symbols with relative explanation can be found in the table C.1 in the appendix.

2.2. Model definition

2.2.1. Generative process

LDA is a hidden variable model, since the observed words in the documents interact with the hidden topics and their distribution across documents (Blei & Lafferty 2009). This interaction is shown in the generative process for the corpus assumed by LDA (Blei et al. 2003; Blei & Lafferty 2009; Ponweiser 2012; Hornik & Grün 2011):

1. For all topics $k \in \{1, ..., K\}$:
   a) Choose a word distribution: $\beta_k \sim Dir(\eta)$
2. For all documents d_j where $j \in \{1, ..., M\}$:
   a) Choose a topic distribution: $\theta_j \sim Dir(\alpha)$
   b) For all words w_i where $i \in \{1, ..., N_j\}$:
      i. Assign topics to word: $z_{j,i} \sim Mult(\theta_j)$
      ii. Draw a word w_i: $w_{j,i} \sim Mult(\beta_{z_{j,i}})$

The Dirichlet distribution, explained in the appendix A, is assumed for the document-topic proportions θ and the topic-word distributions β since "each component in the random vector is the probability of drawing the item associated with that component" (Blei & Lafferty 2009), that is topics and words in the vocabulary respectively.

2.2.2. Assumptions

The number of topics K will be assigned by the user and therefore considered to be fixed (Blei et al. 2003). It will be explained in section 2.5 how this number can be determined.

Moreover, LDA depends on the bag-of-words exchangeability assumption, i.e. words and therefore topic are exchangeable within a document (Blei et al. 2003).

2.3. Parameter estimation

LDA tries to backtrack the generative process, i.e. find the topic-word and document-topic associations from the documents in the corpus. This is mathematically described by the posterior distribution of the hidden variables given a document (Blei et al. 2003):

$$p(\theta, \mathbf{z}|d_j, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, d_j|\alpha, \beta)}{p(d_j|\alpha, \beta)} \tag{2.1}$$

For specific α and β, the coefficient for the topic distribution in a document and the word distribution for a topic respectively, the marginal distribution of a document d_j is (Blei et al. 2003):

$$p(d_j|\alpha, \beta) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_i)} \int \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right) \left(\prod_{i=1}^{N_j} \sum_{k=1}^{K} \prod_{l=1}^{V} (\theta_k \beta_{j,l})^{w_i^l} \right) \tag{2.2}$$

However, this function is intractable since it involves the sum over all latent topic assignments (Ponweiser 2012), consequently it is not possible to use maximum likelihood (ML) to maximize the log-likelihood with respect to the parameters α and β.

2.3.1. Inference

Because of the intractability of the log-likelihood function different approximate inference algorithms have been proposed. In this paper we will focus on convexity-based variational Bayesian inference, as in the original paper from Blei et al. (2003), which takes advantage of Jensen's inequality to calculate a tractable lower bound on the log-likelihood function (Blei et al. 2003).

One can procure a family of distributions on the latent variables α and β by introducing variational parameters γ and ϕ which are document specific and therefore free to vary over documents (Hornik & Grün 2011; Blei et al. 2003). This family is identified by the following variational distribution (Blei et al. 2003):

$$q(\theta, \mathbf{z}|\gamma, \phi) = q_1(\theta|\gamma) \prod_{i=1}^{N_j} q_2(z_{j,i}|\phi_{j,i}) \tag{2.3}$$

where $q_1()$ is a Dirichlet distribution with parameters γ and $q_2()$ is a multinomial distribution with parameters ϕ_i (Blei et al. 2003).

The variational parameters γ and ϕ are then recovered by minimizing the Kullback-Leibler (KL) divergence between the variational and the the posterior distribution (Blei et al. 2003):

$$(\gamma^*, \phi^*) = \underset{(\gamma,\phi)}{\arg\min}\, D_{KL}(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|d_j, \alpha, \beta)) \tag{2.4}$$

2.3.2. Estimation

One can now apply variational expectation-maximization (VEM) to find approximative estimates of the model's parameters α and β.

As usual in EM, the algorithms alternates between two steps, the expectation (E-step) and the maximization step (M-step), until convergence. In the E-step one derives the expectation of the log-likelihood with respect to the latent variables using the present parameters' estimates, while in the M-step one calculates new estimates for the parameters (Dempster et al. 1977; Moon 1996).

In the LDA case the two steps of the VEM algorithm are as follows (Blei et al. 2003; Hornik & Grün 2011):

- **E-step**: Find $\{\gamma_j^*, \phi_j^*\}$, where $d_j \in D$, i.e. the optimal values of the variational parameters γ and ϕ for each document
- **M-step**: Maximize the resulting lower bound on the log-likelihood with respect to the latent parameters α and β

2.4. Creation of corpus

In order to apply the LDA algorithm one needs to transform the data into a Document-Term-Matrix (DTM), i.e. a representation in which rows correspond to the documents, the column to the words and the matrix values are the frequencies of the words in the documents (Ponweiser 2012).

Before this is possible, texts need firstly to be split into the different words, a process called tokenization. One could also tokenize texts in groups of multiple words, often pairs, but in this paper we will focus on singletons. Additionaly, one needs to perform some pre-processing to prepare the data for text analysis. Some important steps are the following (Welbers et al. 2017):

- lowercasing: making all text lower case
- stemming: reducing the words to their base form, e.g. 'love', 'lover' and 'lovingly' become 'love'
- removal of stopwords: filtering out words that are too common and non-informative, e.g. 'the' in the English language

However, one needs to pay attention to the order in which these operations are carried out, because they might work against each other. For example, a stemmed word might not be found in the list of stopwords and therefore not be removed.

2.5. Number of topics

The number of topics K is taken as fixed and known, but actually this also needs to be estimated.

To this purpose different measures have been developed. In particular, when applying the VEM inference process with LDA, one can look at the perplexity, as explained by Blei et al. (2003), and the metrics developed by Cao et al. (2009), Arun et al. (2010) and Deveaud et al. (2014).

Firstly, Blei et al. (2003) looks at the perplexity of a held-out set to evaluate LDA in comparison to other models, but one can use this metric to compare LDA with different values of K. This is calculated for a hold-out set of T as follows (Blei et al. 2003):

$$Perplexity(D_{test}) = exp\left\{ - \frac{\sum_{j=1}^{T} log\ p(d_j)}{\sum_{j=1}^{T} N_j} \right\} \tag{2.5}$$

and the lower it is, the better the model.

Beside that, Cao et al. (2009) proposed a method for finding the appropriate number of topics based on density. Topics are seen as semantic clusters, thus a "smaller similarity between intra-clusters shows that the topic structure is more stable" (Cao et al. 2009). The approach uses a sample set from the word distribution and calculates α and β for an LDA model with an initial value of topics K_0, which is then updated in the following way (Cao et al. 2009):

$$K_{n+1} = K_n + f(\mathbf{r})(K_n - C_n) \tag{2.6}$$

where $f(\mathbf{r})$ is the changing direction of the distance $\mathbf{r}$ and C_n is the cardinality, i.e. the number of topics of the LDA model whose densities are less than the positive integer n (Cao et al. 2009). One calculates for each step the average cosine distance

$$ave\_dis(\beta) = \frac{\sum_{i=1}^{K} \sum_{l=i+1}^{K} cosine\_dist(T_i, T_l)}{K(K-1)/2} \tag{2.7}$$

and repeats the process until the models's average cosine distance $ave\_dis(\beta)$ and cardinality C_n both convergence (Cao et al. 2009).

Differently, Arun et al. (2010) regards LDA as a factorization method of the document-word frequency matrix and produces a measure that is small when the 'correct' number of topic is considered. In particularly, this matrix M is split into two stochastic matrices, a topic-word matrix $M1$ and a document-topic matrix $M2$, and the distribution of singular values of $M1$, C_{M1}, and the distribution of the normalization of $M2$ multiplied by the vector of document lengths D, C_{M2}. Finally, one obtains the following metric (Arun et al. 2010):

$$Metric(M1, M2) = KL(C_{M1}||C_{M2}) + KL(C_{M2}||C_{M1}) \tag{2.8}$$

Finally, the metric developed by Deveaud et al. (2014) tries to maximize the "information divergence D between all pairs (k_i, k_j) of LDA's topics" (Deveaud et al. 2014), in fact it derives the optimal number of topics from the following function (Deveaud et al. 2014):

$$\hat{K} = \arg\max_{K} \frac{1}{K(K-1)} \sum_{(k,k')\in\mathbb{T}_K} D(k||k') \tag{2.9}$$

where k and k' correspond to each pair of topics in the set of K topics modeled by LDA, $\mathbb{T}_K$. To avoid issues when computing the information divergence between pairs of topics, Deveaud et al. (2014) uses the Jensen-Shannon divergence, thus:

$$D(k||k') = \frac{1}{2} \sum_{w\in\mathbb{W}_k\cap\mathbb{W}_{k'}} \beta_{j,k} log \frac{\beta_{j,k}}{\beta_{j,k'}} + \frac{1}{2} \sum_{w\in\mathbb{W}_k\cap\mathbb{W}_{k'}} \beta_{j,k'} log \frac{\beta_{i,k'}}{\beta_{i,k}} \tag{2.10}$$

where $\mathbb{W}_k$ is the set of the n words with the highest $\beta_{i,k}$ in topic k.

However, one should also consider their own assessment in this part, especially when using the model for descriptive and not predictive purposes. In fact, Chang et al. (2009) show that traditional metrics, such as the ones explained so far, are often uncorrelated, or even negatively correlated, with human judgment. Therefore, Chang et al. (2009) suggest to concentrate on people's performance evaluations instead of metrics to optimize.

3. Analysis

3.1. Data

The dataset used for this project is the union of information provided by Sergey Kuznetsov and Gyanendra Mishra on kaggle.com.

The dataset was then restricted to 200 songs randomly selected among the ones from 2010 onward, of the most common music genres and of a limited number of artists. The complete list of the songs can be found in table C.2 in the appendix.

3.1.1. Exploratory Data Analysis

We can see from figure 3.1 that only songs from 17 artists were kept, almost 50% of which being from only 3 artists. Moreover, it is evident how some music genres are more predominant, in fact country and jazz together account for less than 10%. Finally, the songs are quite spread out across the 7 years taken into consideration.

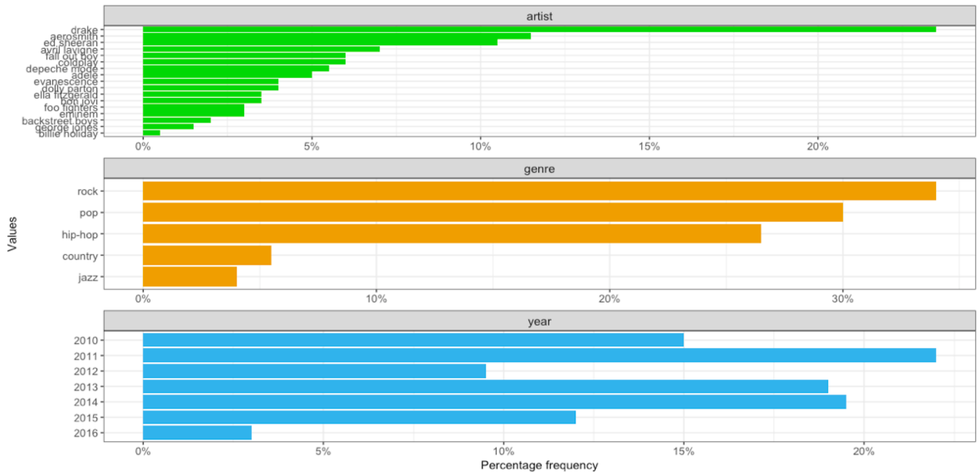


Figure 3.1.: Frequency plots

Particularly intriguing is the fact that all song from each artist are assigned the same music genre. This might depend on a few factors, like the songs randomly selected or the way the information about the music genre was acquired.

Furthermore, it is interesting to look at the distribution of the word counts, especially for different music genres. In figure 3.2 we observe how Hip-Hop is in general the genre with the longest texts, while Jazz and Country the ones with the shortest lyrics.

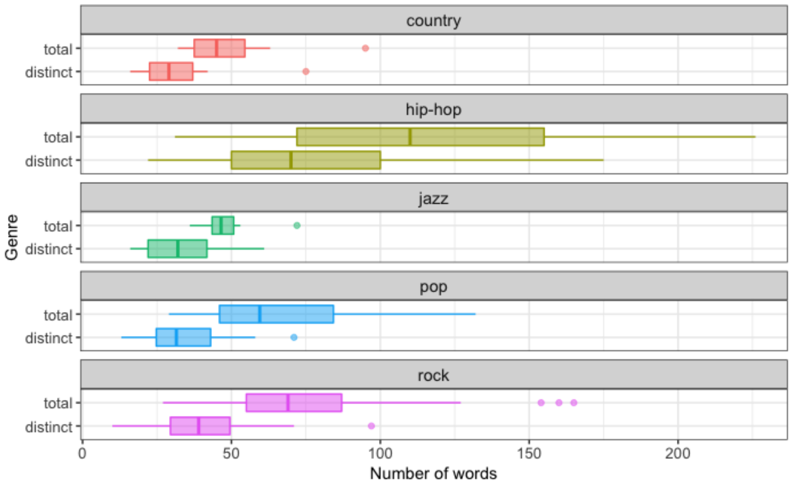


Figure 3.2.: Distribution of total and distinct words for genres

Since we are interested in finding the topics per artist and per genre, we should firstly have a look at the most used words for each genre, which are displayed in figure 3.3. One can see how some words are present for each genre, but how the general spirit of the words revolve around different themes. Pop and Rock seem to talk more about relationships, while Country seems to focus on everyday life and hip-hop is more rude in its text. Jazz, contrarily, has more words that have the same count and appears also to talk about love, but in a more romanticized way.

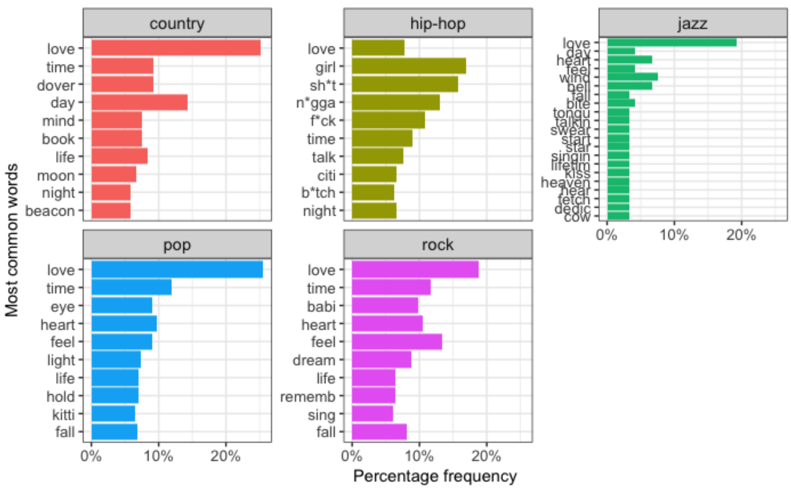
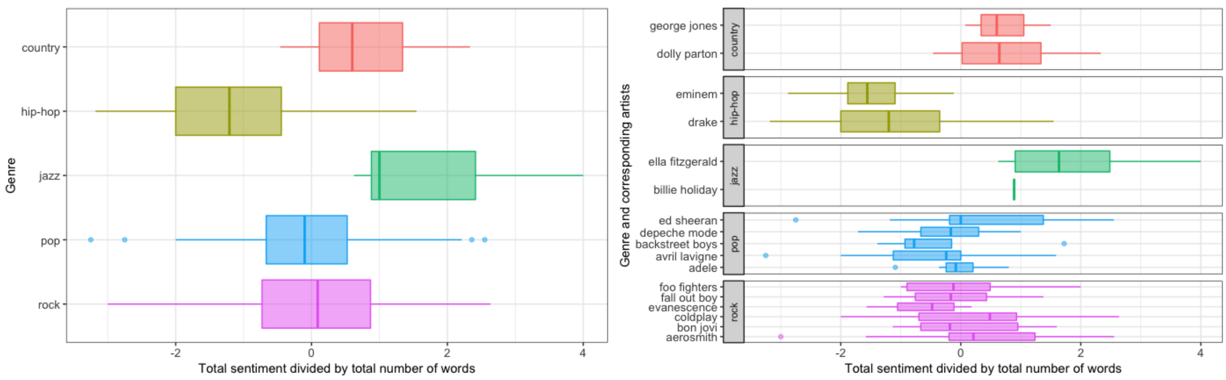


Figure 3.3.: Ten most common words in each genre

3.1.2. Sentiment Analysis

As part of the general analysis of the dataset simple sentiment analysis was performed. As explained in the appendix section B, various lexicons are available for sentiment analysis. We used *afinn* by Nielsen (2011) because it makes the comparison among genres and artists easier, since it gives integer values to the words, which can then be summed for the different song texts.



(a) Genre (b) Artist

Figure 3.4.: Weighted sentiments with respect to genre and artist

One can see from figure 3.4 that Hip-Hop is the most negative genre in the dataset. Both artists have predominantly negative songs, but Drake has a wider range of sentiments in his songs than Eminem, having even a positive outlier. However, some Pop and Rock songs are also quite negative, especially the ones from Avril Lavigne

On the opposite side, Jazz and Country and all their artists have almost only positive songs. One could therefore expect Jazz and Country to mainly have positive topics, while Hip-Hop to have a tendency towards negative topics.

Finally, Rock and Pop songs are more split in positive and negative sentiments, also within the same artist. Beside Avril Lavigne, also Evanescence and Backstreet Boys have a tendency towards negative texts.

3.2. Implementation in R

3.2.1. Corpus creation

To create the object to apply the function `LDA`, the function `dfm` from the package `quanteda` was used, since it automatically carries out the pre-processing tasks described in section 2.4

in the correct order while simultaneously creating the Document-Term-Matrix (Welbers et al. 2017).

Specifically, the stopwords removed are all the ones for the English language present in the function the function `stopwords` from the package `stopwords`. Furthermore, after initial application of the LDA algorithm, it became clear that in the song lyrics many more stop words were present than the standard ones present in the function `stopwords`. Examples of this are 'chorus' and 'verse', which are just written in the text to let the reader know when these parts of the song start. A complete list of extra stopwords can be red in the table C.3 in the appendix. These were therefore added to the list of stopwords to be removed in creating the DTM.

3.2.2. Determination of the number of topics

As explained in subsection 2.5, the number of topics will be taken as given from the algorithm. Therefore, we firstly need to determine which value of K best fits the corpus.

The function `FindTopicsNumber` from the package `ldatuning` calculates the metrics developed by Cao et al. (2009), Arun et al. (2010) and Deveaud et al. (2014). To compute the perplexity according to Blei et al. (2003), the function `perplexity` from the package `topicmodels` will be used.

In figure 3.5 the scaled values for all the metrics are displayed for values of K between 2 and 200, i.e. the total number of documents, where the metrics are divided according to their objective, maximization or minimization

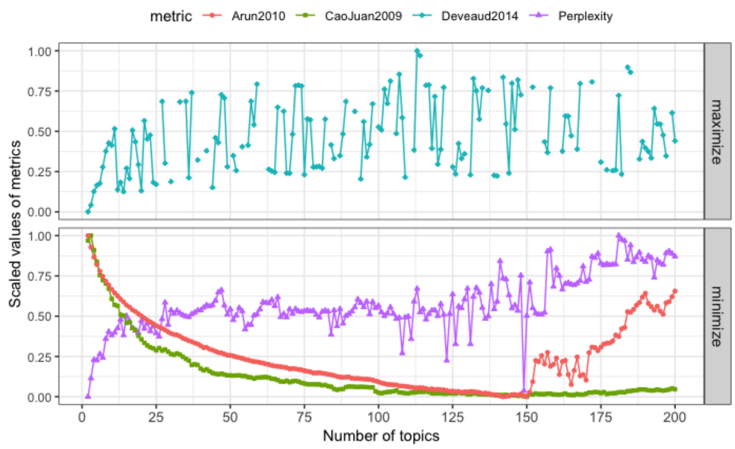


Figure 3.5.: Scaled distribution of metrics' values with respect to the number of topics

Metrics	Number of topics
Perplexity	2
CaoJuan2009	150
Arun2010	142
Deveaud2014	113

Table 3.1.: Optimal number of topics for different metrics

Some values of the metric *Deveaud2014* are removed because they were ∞, which is not an acceptable value for this metric. Moreover, this metric does not convey a clear pattern, since there are many peaks. Secondly, both metrics *CaoJuan2009* and *Arun2010* are decreasing until values of K between 140 and 150, but the former keeps being low while the latter increases afterwards. Finally, the values for *Perplexity* have a rising tendency, with a local minimum also between 140 and 150. The optimal values for each metric are shown in table 3.1.

Unfortunately, different metrics deliver different optimal number of topics. We will therefore make use of our human judgment, as disclosed in subsection 2.5. Indeed, we are looking to make a descriptive analysis of our dataset, more than predicting topics for other documents, therefore a smaller number of topics will be more easily interpretable. Since we have previous information about the dataset, we will look at the results with values of K ranging from 2, the optimal value according to *Perplexity*, to 5, the number of music genres.

3.2.3. LDA application

There are different packages in `R` that perform LDA, but for this project `topicmodels` was selected, since the other do not offer an implementation of the VEM algorithm, which is the one applied by the original paper on LDA, Blei et al. (2003), and the one selected for the research.

After running the algorithm `LDA` provided by this package there are two main matrices with coefficients of interest (Silge & Robinson 2017):

- *beta*: per-topic-per-word probability, i.e. the "probability of that term being generated from that topic" (Silge & Robinson 2017)
- *gamma*: per-document-per-topic probability, i.e. the "estimated proportion of words from that document that are generated from that topic" (Silge & Robinson 2017)

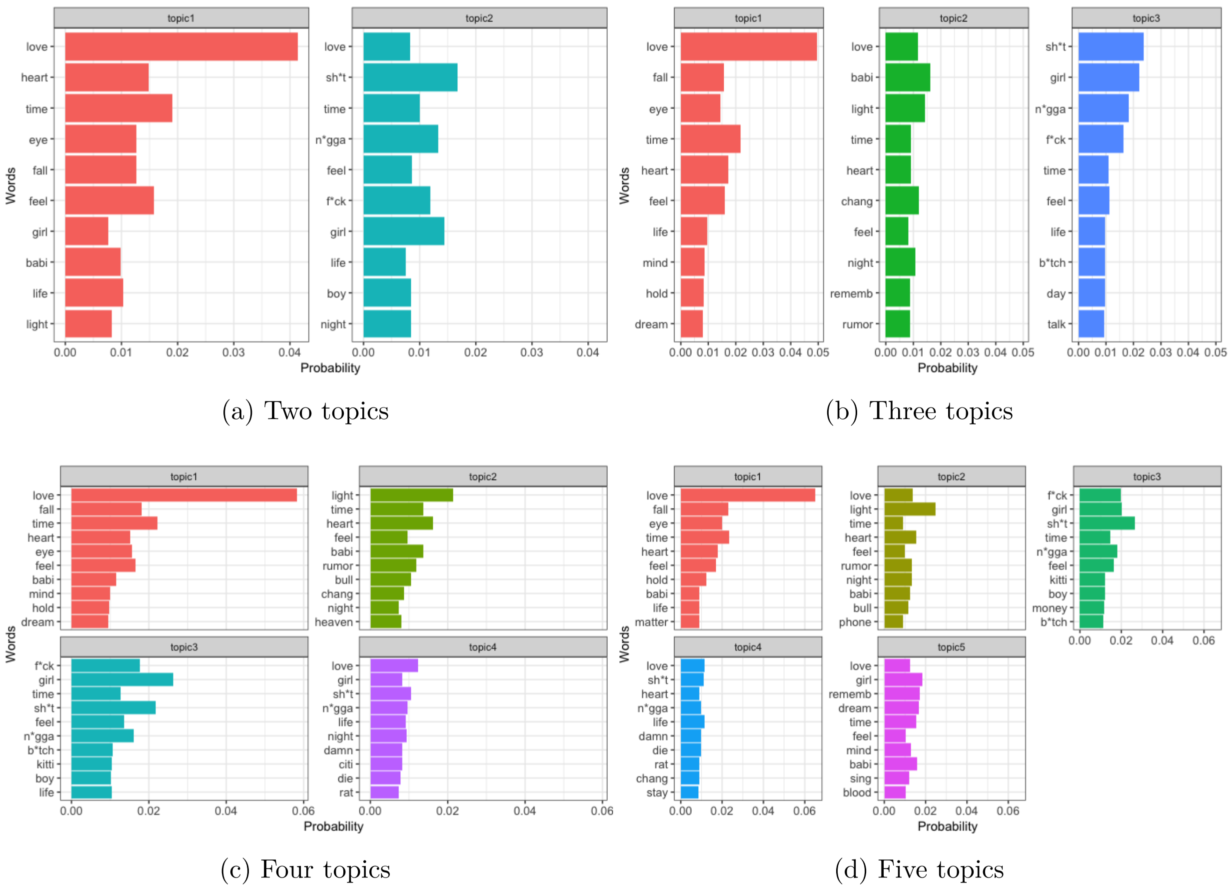
This can be accessed using the implementation for LDA of the function `tidy` from the package `tidytext` giving the values *beta* or *gamma* for the argument *matrix*. This returns a dataframe with either topic and term or document and topic and their respective probabilities.

4. Results

4.1. LDA outcomes

4.1.1. Per-topic-per-word probability

Firstly, we look at the top ten words by *beta*-probability for each topic for values of K from 2 to 5, which are displayed in figure 4.2.



(a) Two topics

(b) Three topics

(c) Four topics

(d) Five topics

Figure 4.1.: Per-topic-per-word probabilities of the top 10 words per topic

One can immediately see that there are some words common to multiple topics, such as 'love' and 'life'. Furthermore, one cannot see a clear distinction among topics, especially when the value of K increases. On the one side some have love-focused topics, while the others have more curse words. In fact, when looking at the average *afinn* sentiment according to Nielsen (2011) weighted by the *beta*-probabilities, shown in figure 4.2, one can observe how the former kind of topics are on average positive, while the other are on average negative. One can also notice how there is always, except for the case of just two topics, (at least) one more neutral topic, which incorporates both love-related words and more general ones, such as 'night' and 'rumor'.

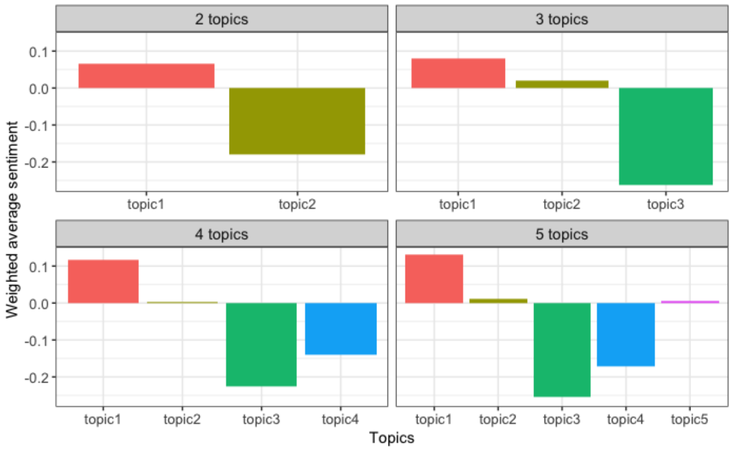


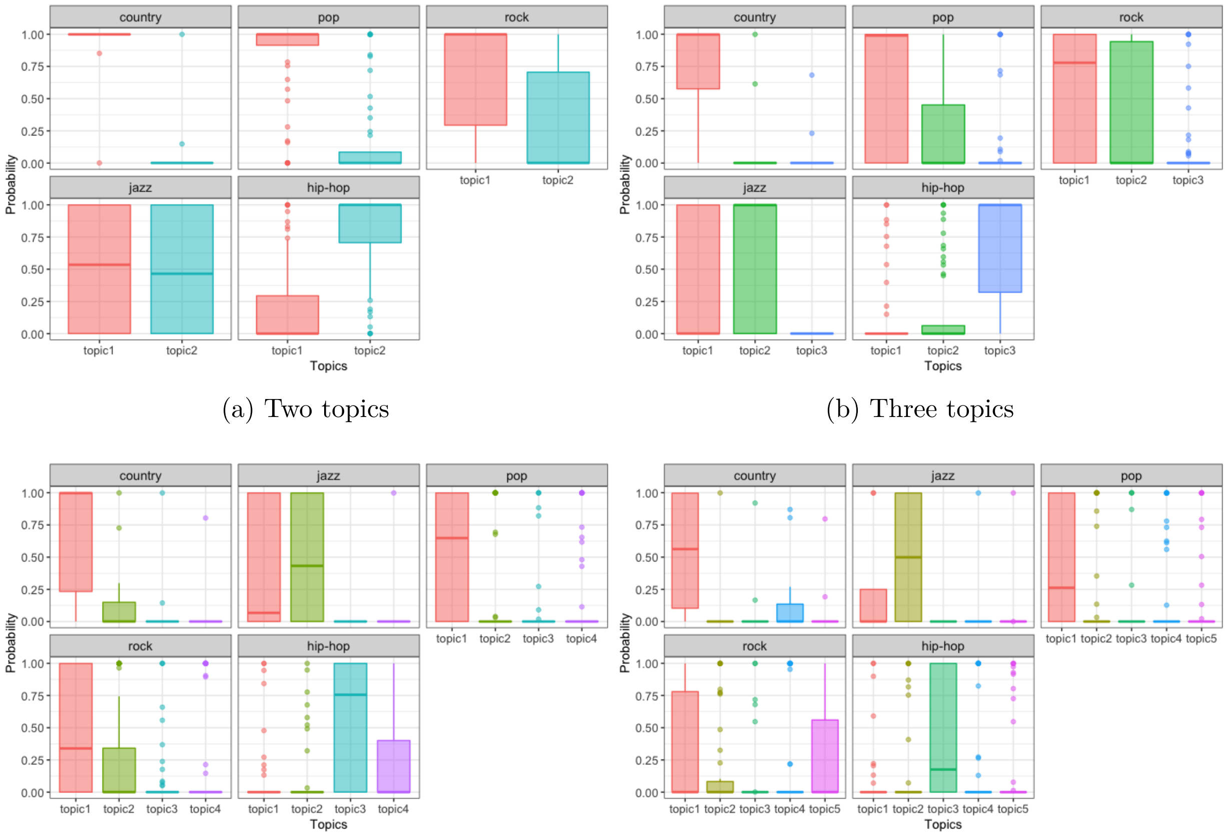
Figure 4.2.: Weighted average *afinn* sentiment of the topics for different values of K

4.1.2. Per-document-per-topic probability

Secondly, we look at the *gamma*-probability distribution for genres, which is shown in figure 4.3.

Considering each song as a document, one can observe from figure 4.3 that the genres Jazz and Rock are more split between topics. This is especially evident in the case of 2 and 3 topics. However, it is to be noted that, in the case of 4 topics, one topic (*topic4*) has in general low probabilities for all songs. This leads us to think that this value of K is not appropriate for this dataset. The same phenomenon can be seen in the case of 5 topics.

One needs to only carefully analyze these figures, since, as shown in figure 3.1, some artists have more songs in the dataset and could drive the distribution of their specific genre. The *gamma*-probability distribution for the specific artists is displayed in figure **??** in the appendix. One could derive from this figure that maybe 4 or 5 topics are appropriate for this dataset, since at least half of the artists show high probabilities for mostly one topic.



(a) Two topics (b) Three topics

(c) Four topics (d) Five topics

Figure 4.3.: Per-document-per-topic probability distribution for music genres

4.2. Findings

Interpretation of the *gamma*-probability with respect of music genres and artists may lead to different conclusions. In the former case, one could favor a low number of topics, while in the latter a larger one.

However, as we increase the number of topics, the words present in the additional ones are quite similar, as it was seen in figure .

Therefore, for the sake of interpretation, we believe that the most appropriate number of topics for the dataset is 3, which are on average split among the genres as follows:

- *topic1*: Country, Pop, Rock, Jazz
- *topic2*: Rock, Jazz, partly Pop
- *topic3*: Hip-Hop

The fact that multiple genres (Pop, Rock and Jazz) have documents with high estimated proportions of words from different topics (*topic1* and *topic2*) might derive from the similarity of these topics, which had also been witnessed in figure 4.2, where the words with the highest *beta*-probability per topic were displayed, and by presence of multiple artists with a different amount of songs.

5. Conclusion

Text mining is a field of research which became especially relevant because of the increasing volume of text data from newspapers and social media (Silge & Robinson 2017). In particular, topic modelling aims at finding the underlying topics in a collection of text documents, in order to subdivide them to better understand their message (Silge & Robinson 2017). The applied algorithm is called Latent Dirichlet Allocation (LDA), which was developed by Blei et al. (2003). It is an unsupervised learning algorithm that "treats each document as a mixture of topics, and each topic as a mixture of words" (Silge & Robinson 2017).

The purpose of the research was to uncover the latent thematic groups in a corpus of song lyrics from different artists and genres and to determine if songs for artists and genres related to a single topic or not. The assumption was that songs of a single artist can talk about different topics and that topics would be quite different from each other, since the dataset includes songs from different artists and genres.

The data contained 200 songs from different artists and genres and their relative lyrics. Before applying the algorithm, the corpus has been pre-processed by performing the steps explained in section 2.4. Furthermore, initial exploratory data analysis showed that Hip-Hop is the most negative genre and the one with the highest amount of words.

The used metrics for selecting the number of topics did not convey a unanimous message, therefore it was made use of previous knowledge about the data to choose the values to test. Out of the four tested, the number of topics that better connected with artists and genres, while still being easily interpretable was 3. Only Hip-Hop and Country were explained almost entirely by only one topic, while the other music genres have songs which are highly related to multiple topics.

Being LDA an unsupervised learning algorithm, the project required a high level of interpretation work, therefore the results are not final. Further research could involve the evaluation of different estimation methods, such as Gibbs sampling proposed by Griffiths & Steyvers (2004) and the application of keyword and topic re-ranking as explained by Song et al. (2009).

Appendix

A. Dirichlet Distribution

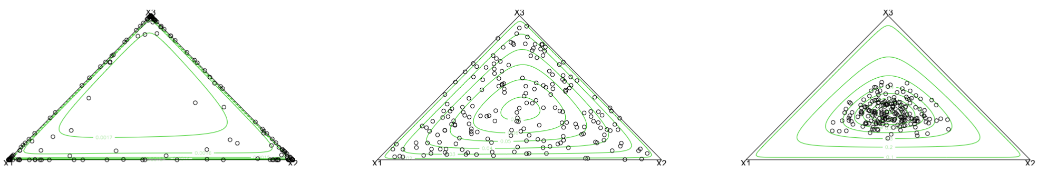
The Dirichlet distribution is a "multivariate generalization of the Beta distribution" (Lin 2016). According to Ng et al. (2011), the Dirichlet is the preferred option when modelling multivariate data consisting of proportions, since all values are on (0,1) and sum to one. In fact, it is often used when dealing with text data (Frigyik et al. 2010).

The Dirichlet distribution family are probability distributions of K-dimensional multinomial distributions' space (Sklar 2014). Written as $\mathbf{X} \sim Dir(\alpha)$, its parameter α is a vector of dimension K. When all α_k, $k \in \{1, ..., K\}$ are the same, then the distribution is called symmetric Dirichlet distribution (Lin 2016). The probability distribution function (pdf) is then defined as follows (Frigyik et al. 2010):

$$f(x_k; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} q_i^{\alpha_i - 1} \tag{.1}$$

and is defined on the K-1 dimensional probability simplex (Lin 2016).

As figure A.1 shows, when α values are low, that is lower than 1, the points are concentrated around the edges of the probability index, while when α has values higher over 1 the points are more towards the center (Frigyik et al. 2010). When all values of α are equal to 1, it is the "uniform distribution over the simplex" (Frigyik et al. 2010).



(a) $\alpha = (0.1, 0.1, 0.1)$ (b) $\alpha = (1, 1, 1)$ (c) $\alpha = (10, 10, 10)$

Figure A.1.: 3-Simplex for different values of α

B. Sentiment Analysis

"Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu 2012). This is a useful start to analyzing text corpora and has become particularly common since the steep increase in the use of social media.

Using `R` for text analysis, one can make use of the dataset `sentiments` from the package `tidytext` which contains the sentiments for English words from four different lexicons:

- *nrc*, a "list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive)" (Mohammad, Saif 2019; Mohammad & Turney 2013)
- *bing*, a "list of English positive and negative opinion words or sentiment words" (Liu, Bing and Hu, Minqing 2019; Hu & Liu 2004)
- *loughran*, list of words "specific to financial reporting" (Silge, Julia 2017; Loughran & McDonald 2016)
- *afinn*, a list of English words "rated for valence with an integer between minus five (negative) and plus five (positive)" (Nielsen 2011)

The choice of the lexicon to use in the analysis depends on the purpose of the text analysis to be performed (MonkeyLearn 2017).

C. Tables

Symbol	Meaning
D	corpus
d	document
w	word
z	topic
j	index over documents
i	index over words
k	index over topics
M	number of documents in corpus
N_j	number of words in document j
V	number of words in corpus
K	number of topics (specified by user)

Table C.1.: List of symbols

Table C.2.: List of songs in the dataset

Genre	Artist	Song	Year
country	dolly parton	comes and goes	2010
country	dolly parton	from here to the moon and back	2014
country	dolly parton	book of life	2010
country	dolly parton	better day	2011
country	dolly parton	banks of the ohio	2014
country	dolly parton	down from dover	2010
country	dolly parton	but you loved me then	2010
country	dolly parton	before you make up your mind	2010
country	george jones	beacon in the night	2016
country	george jones	give my love to rose	2016
country	george jones	day after forever	2012
hip-hop	drake	fireworks	2010
hip-hop	drake	madonna	2015
hip-hop	drake	light up	2010
hip-hop	drake	side pieces	2014

Table C.2.: List of songs in the dataset

Genre	Artist	Song	Year
hip-hop	drake	the language	2013
hip-hop	drake	how about now	2014
hip-hop	drake	six god	2014
hip-hop	drake	shot for me	2011
hip-hop	drake	baby come with me	2011
hip-hop	drake	show me a good time	2010
hip-hop	drake	good ones go	2011
hip-hop	drake	up all night	2010
hip-hop	drake	own it	2013
hip-hop	drake	right hand	2015
hip-hop	drake	take care	2011
hip-hop	drake	hype	2016
hip-hop	drake	10 bands	2015
hip-hop	drake	enough said	2012
hip-hop	drake	notice me	2011
hip-hop	drake	connect	2013
hip-hop	drake	trust issues	2012
hip-hop	drake	unforgettable	2010
hip-hop	drake	karaoke	2010
hip-hop	drake	back to back	2015
hip-hop	drake	hotline bling	2015
hip-hop	drake	my side	2015
hip-hop	drake	go out tonight	2015
hip-hop	drake	cameras	2012
hip-hop	drake	6 man	2015
hip-hop	drake	july	2010
hip-hop	drake	bar mitzvah in 1999	2014
hip-hop	drake	do it all	2010
hip-hop	drake	find your love	2010
hip-hop	drake	doing it wrong	2011
hip-hop	drake	6 god	2014
hip-hop	drake	still here	2016

Table C.2.: List of songs in the dataset

Genre	Artist	Song	Year
hip-hop	drake	grammys	2016
hip-hop	drake	come thru	2013
hip-hop	drake	9	2016
hip-hop	drake	star67	2015
hip-hop	drake	legend	2015
hip-hop	drake	too much	2013
hip-hop	drake	the resistance	2010
hip-hop	drake	wednesday night interlude	2015
hip-hop	drake	girls love beyonce	2013
hip-hop	drake	nothing was the same	2013
hip-hop	drake	independent queen	2011
hip-hop	eminem	give me the ball	2011
hip-hop	eminem	cocaine	2010
hip-hop	eminem	despicable	2010
hip-hop	eminem	above the law	2011
hip-hop	eminem	echo	2010
hip-hop	eminem	get money	2010
jazz	billie holiday	just one of those things	2012
jazz	ella fitzgerald	i wish i were in love again	2013
jazz	ella fitzgerald	almost like being in love	2010
jazz	ella fitzgerald	dedicated to you	2010
jazz	ella fitzgerald	hear me talking to ya	2013
jazz	ella fitzgerald	cow cow boogie	2013
jazz	ella fitzgerald	gone with the wind	2013
jazz	ella fitzgerald	if i were a bell	2013
pop	adele	when we were young	2015
pop	adele	remedy	2015
pop	adele	hello	2015
pop	adele	love in the dark	2015
pop	adele	turning tables	2011
pop	adele	rumour has it	2011
pop	adele	set fire to the rain	2011

Table C.2.: List of songs in the dataset

Genre	Artist	Song	Year
pop	adele	million years ago	2015
pop	adele	all i ask	2015
pop	adele	i miss you	2015
pop	avril lavigne	darlin	2011
pop	avril lavigne	alice	2011
pop	avril lavigne	fly	2013
pop	avril lavigne	adia	2013
pop	avril lavigne	hello kitty	2013
pop	avril lavigne	not enough	2011
pop	avril lavigne	wish you were here	2011
pop	avril lavigne	smile	2011
pop	avril lavigne	making my way down town	2013
pop	avril lavigne	how you remind me	2012
pop	avril lavigne	let me go	2013
pop	avril lavigne	17	2013
pop	avril lavigne	hush hush	2013
pop	avril lavigne	rock n roll	2013
pop	backstreet boys	take care	2013
pop	backstreet boys	soldier	2013
pop	backstreet boys	one phone call	2013
pop	backstreet boys	try	2013
pop	depeche mode	new dress	2014
pop	depeche mode	sacred	2014
pop	depeche mode	the things you said	2014
pop	depeche mode	should be higher	2013
pop	depeche mode	but not tonight	2014
pop	depeche mode	clean	2014
pop	depeche mode	the child inside	2013
pop	depeche mode	waiting for the night	2010
pop	depeche mode	blasphemous rumours	2014
pop	depeche mode	sweetest perfection	2014
pop	depeche mode	welcome to my world	2013

Table C.2.: List of songs in the dataset

Genre	Artist	Song	Year
pop	ed sheeran	shirtsleeves	2014
pop	ed sheeran	let it out	2011
pop	ed sheeran	be like you	2011
pop	ed sheeran	give me love	2011
pop	ed sheeran	runaway	2014
pop	ed sheeran	bloodstream	2014
pop	ed sheeran	wayfaring stranger	2012
pop	ed sheeran	kiss me	2011
pop	ed sheeran	firefly	2010
pop	ed sheeran	heaven	2012
pop	ed sheeran	homeless	2010
pop	ed sheeran	parting glass	2013
pop	ed sheeran	the city	2011
pop	ed sheeran	she	2011
pop	ed sheeran	photograph	2014
pop	ed sheeran	one	2014
pop	ed sheeran	gold rush	2011
pop	ed sheeran	thinking out loud	2014
pop	ed sheeran	grade 8	2011
pop	ed sheeran	fire alarms	2011
pop	ed sheeran	miss you	2011
rock	aerosmith	get it up	2014
rock	aerosmith	sight for sore eyes	2014
rock	aerosmith	chip away the stone	2014
rock	aerosmith	love in an elevator	2012
rock	aerosmith	my fist your face	2014
rock	aerosmith	legendary child	2012
rock	aerosmith	amazing	2012
rock	aerosmith	the other side	2014
rock	aerosmith	head first	2014
rock	aerosmith	hole in my soul	2014
rock	aerosmith	same old song and dance	2014

Table C.2.: List of songs in the dataset

Genre	Artist	Song	Year
rock	aerosmith	rag doll	2012
rock	aerosmith	deuces are wild	2012
rock	aerosmith	sunny side of love	2012
rock	aerosmith	beautiful	2012
rock	aerosmith	big ten inch record	2014
rock	aerosmith	i wanna know why	2014
rock	aerosmith	permanent vacation	2014
rock	aerosmith	draw the line	2014
rock	aerosmith	something	2012
rock	aerosmith	blind man	2012
rock	aerosmith	another last goodbye	2012
rock	aerosmith	dream on	2012
rock	bon jovi	army of one	2013
rock	bon jovi	breakout	2010
rock	bon jovi	come back	2010
rock	bon jovi	every road leads home to you	2013
rock	bon jovi	living in sin	2010
rock	bon jovi	99 in the shade	2010
rock	bon jovi	blood on blood	2010
rock	coldplay	x marks the spot	2015
rock	coldplay	hurts like heaven	2011
rock	coldplay	hymn for the weekend	2015
rock	coldplay	fun	2015
rock	coldplay	adventure of a lifetime	2015
rock	coldplay	us against the world	2011
rock	coldplay	every teardrop is a waterfall	2011
rock	coldplay	paradise	2011
rock	coldplay	everglow	2015
rock	coldplay	magic	2014
rock	coldplay	army of one	2015
rock	coldplay	charlie brown	2011
rock	evanescence	made of stone	2011

Table C.2.: List of songs in the dataset

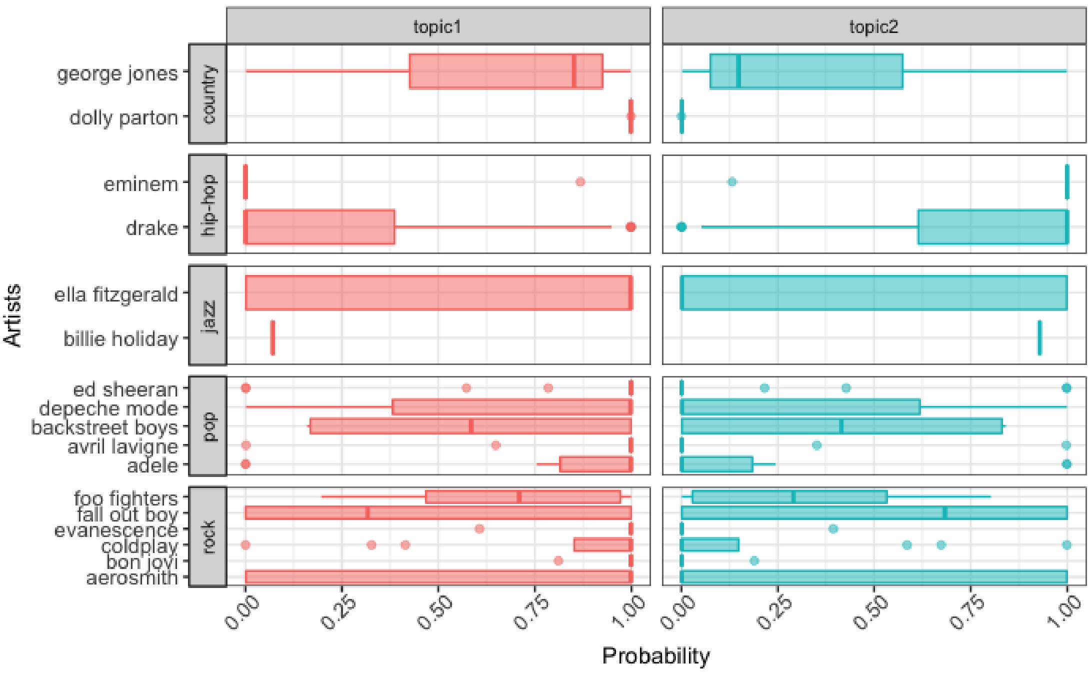
Genre	Artist	Song	Year
rock	evanescence	sick	2011
rock	evanescence	oceans	2011
rock	evanescence	lost in paradise	2011
rock	evanescence	bleed	2010
rock	evanescence	the change	2011
rock	evanescence	end of the dream	2011
rock	evanescence	my heart is broken	2011
rock	fall out boy	jet pack blues	2014
rock	fall out boy	where did the party go	2013
rock	fall out boy	alone together	2013
rock	fall out boy	only the bulls	2014
rock	fall out boy	caffeine cold	2013
rock	fall out boy	the mighty fall	2013
rock	fall out boy	eternal summer	2013
rock	fall out boy	young volcanoes	2013
rock	fall out boy	immortals	2014
rock	fall out boy	the phoenix	2013
rock	fall out boy	rat a tat	2013
rock	fall out boy	centuries	2014
rock	foo fighters	i am a river	2014
rock	foo fighters	white limo	2011
rock	foo fighters	a matter of time	2011
rock	foo fighters	rope	2011
rock	foo fighters	arlandria	2011
rock	foo fighters	something from nothing	2014

Table C.2.: List of songs in the dataset

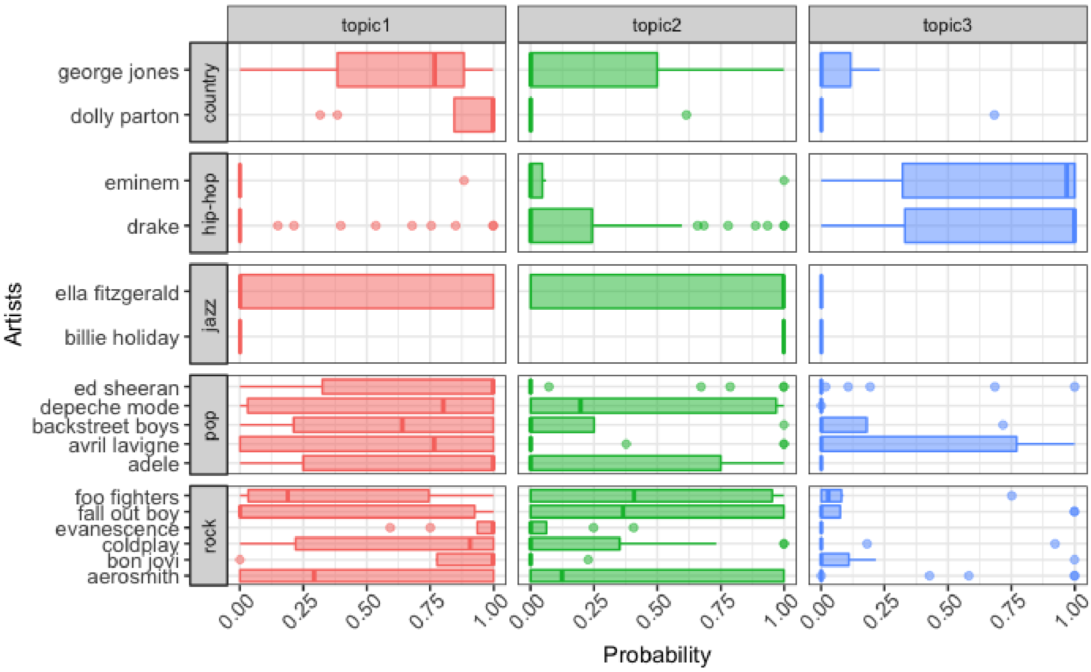
Words
chorus
comma
ding
dong
du
echo
gon
gonna
gotta
hey
hook
m-my
ooh
tat
ti
uh
verse
wanna
ya
yeah
yi

Table C.3.: List of extra stopwords

D. Figures

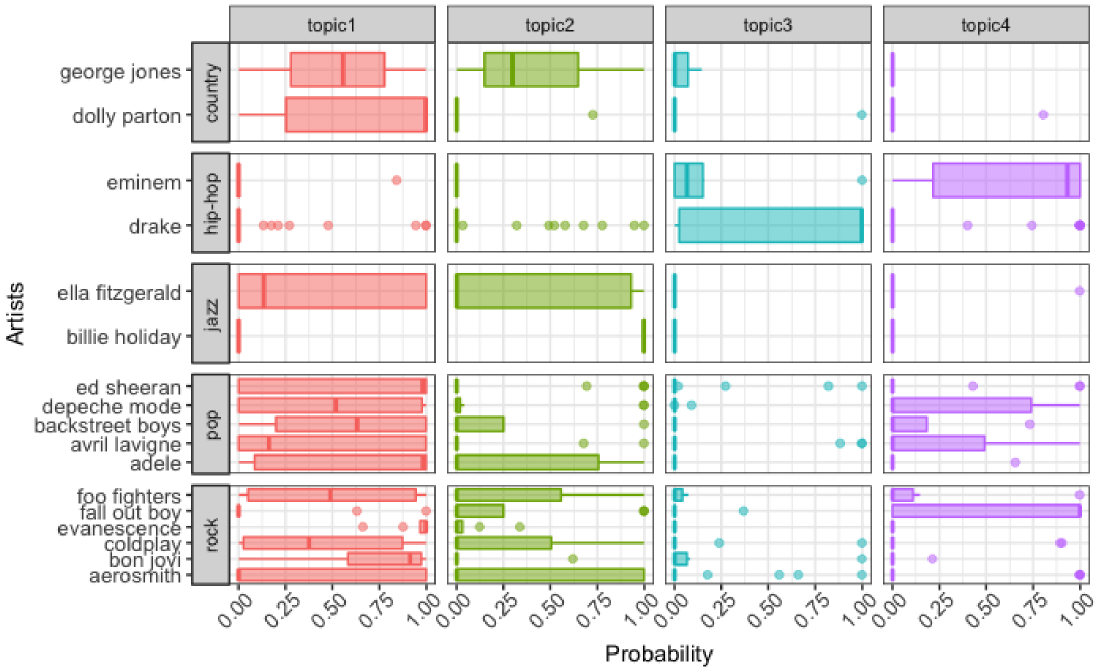


(a) Two topics

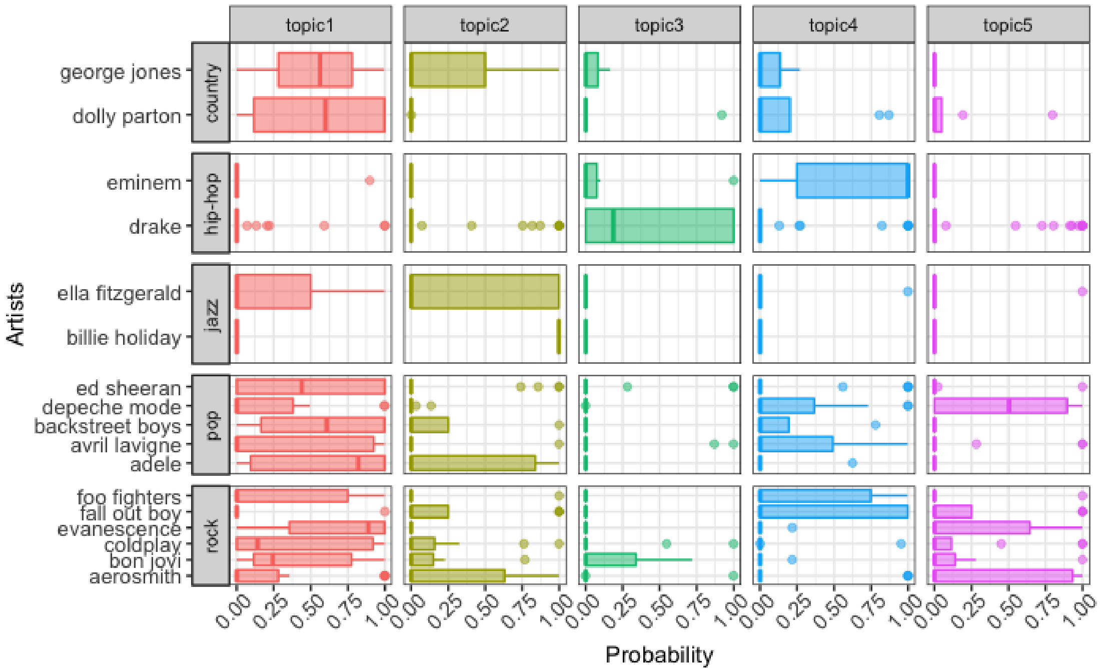


(b) Three topics

Figure D.1.: Per-document-per-topic probability distribution for artists (1)



(c) Four topics



(d) Five topics

Figure D.1.: Per-document-per-topic probability distribution for artists (2)

References

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 391–402).

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In *Text mining* (pp. 101–124). Chapman and Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, *72*(7-9), 1775–1781.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61–84.

Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*(0006), 1–27.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Hornik, K., & Grün, B. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 168–177).

Lin, J. (2016). *On the dirichlet distribution.*

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1–167.

Liu, Bing and Hu, Minqing. (2019). *Opinion mining, sentiment analysis, and opinion spam detection.* `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`. (Online; last accessed on 02. Juli 2019)

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. , *29*(3), 436–465.

Mohammad, Saif. (2019). *Nrc word-emotion association lexicon.* `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`. (Online; last accessed on 02. Juli 2019)

MonkeyLearn. (2017). *Sentiment analysis: Nearly everything you need to know.* `https://monkeylearn.com/sentiment-analysis/`. (Online; last accessed on 06. August 2019)

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, *13*(6), 47–60.

Ng, K. W., Tian, G.-L., & Tang, M.-L. (2011). *Dirichlet and related distributions: Theory, methods and applications* (Vol. 888). John Wiley & Sons.

Nielsen, F. Å. (2011, mar). *Afinn.* Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby: Informatics and Mathematical Modelling, Technical University of Denmark. Retrieved from `http://localhost/pubdb/p.php?6010`

Ponweiser, M. (2012). Latent dirichlet allocation in r.

Silge, J., & Robinson, D. (2017). *Text mining with r: A tidy approach.* " O'Reilly Media, Inc.".

Silge, Julia. (2017). *tidytext 0.1.3.* `https://juliasilge.com/blog/tidytext-0-1-3/`. (Online; last accessed on 02. Juli 2019)

Sklar, M. (2014). Fast mle computation for the dirichlet multinomial. *arXiv preprint arXiv:1405.0099*.

Song, Y., Pan, S., Liu, S., Zhou, M. X., & Qian, W. (2009). Topic and keyword re-ranking for lda-based topic modeling. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 1757–1760).

Tan, A.-H., et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases* (Vol. 8, pp. 65–70).

Tan, P.-N. (2018). *Introduction to data mining.* Pearson Education India.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in r. *Communication Methods and Measures*, *11*(4), 245–265.