Application of LDA topic model to song lyrics

Silvia Ventoruzzo
Freie Universität Berlin

1. Introduction and Motivation

- Topic models are unsupervised learning methods to extract topics from a corpus of documents
- Latent Dirichlet Allocation (LDA) is "generative probabilistic model for collections of discrete data such as text corpora" [1]
- LDA treats each document as a mixture of topics and each topic as a mixture of words
- Bag-of-words assumption: the order of the words in the documents is not important
- This research uses LDA to extract the underlying topics in a corpora of song lyrics

2. Latent Dirichlet Allocation [1]

Model

- 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: $w_i \in \{1, ..., V\}$

Generative process

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: $w_{j,i} \sim Mult(\beta_{z_{j,i}})$

Estimation

Optimize log-likelihood of the marginal distribution of the documents d_j with respect to the coefficients α and β:

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

Inference

Since the function $p(d_j|\alpha,\beta)$ is intractable, variational Bayesian inference is used with the variational parameters γ and ϕ giving the following variational distribution:

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

Variational expectation-maximization (VEM):

- **E-step**: Find $\{\gamma_j^*, \phi_j^*\} = \arg\min_{(\gamma,\phi)} D_{KL}(q(\theta,\mathbf{z}|\gamma,\phi)||p(\theta,\mathbf{z}|d_j,\alpha,\beta))$
- **M-step**: Maximize the resulting lower bound on the log-likelihood with respect to the latent parameters α and β

Repeated until convergence.

3. Creation of corpus

1. Tokenization: split text into the different words
2. Tokens cleaning [6]:
   - Lowercasing: making all text lower case
   - Stemming: reducing the words to their base form
   - Removal of stopwords: filtering out words that are too common and non-informative
   - Removal of other elements, that is numbers, punctuation, symbols and separators
3. Transformation into a document-term matrix [6]:
   - Rows: documents
   - Columns: words
   - Cells: frequencies of the words in the documents

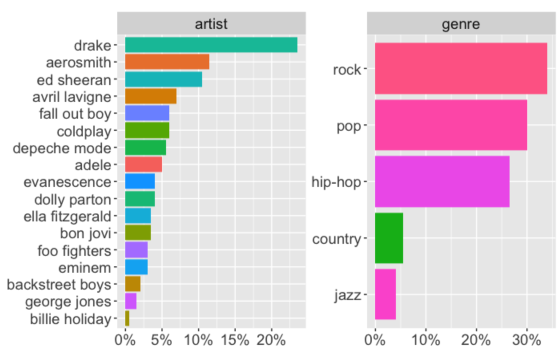
4. Number of topics

Methods to derive the appropriate number of topics:

- Metrics to be minimized:
  - Blei et al. (2003): Perplexity of hold-out set, i.e. geometric mean per-word likelihood
  - Cao et al. (2009): Metric based on the average cosine distance among semantic clusters
  - Arun et al. (2010): Metric based on the information divergence between pairs of topics
- Metric to be maximized:
  - Deveaud et al. (2014): Metric based on the divergence among stochastic matrices
- Chang et al. (2009): Human judgment
  - Word intrusion
  - Topic intrusion

5. Dataset

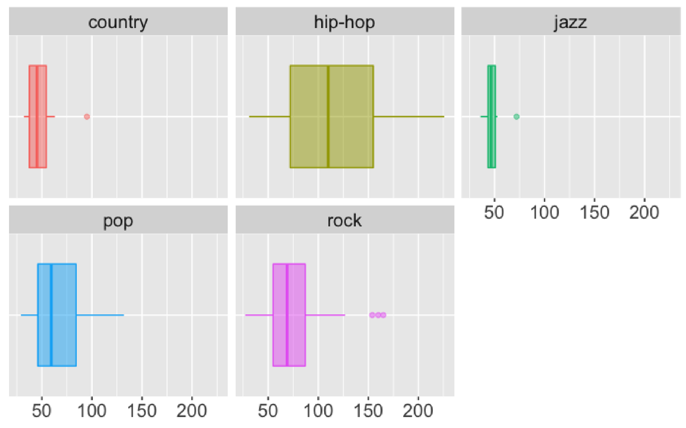
Distribution



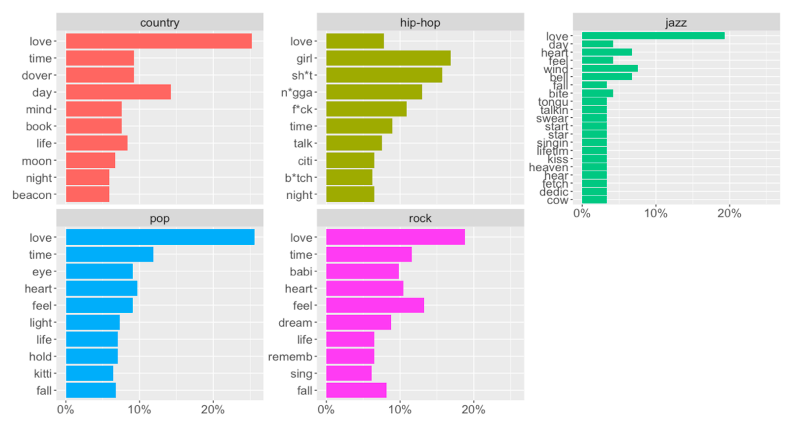
Distribution

- 200 songs
- 17 artists
- 5 genres

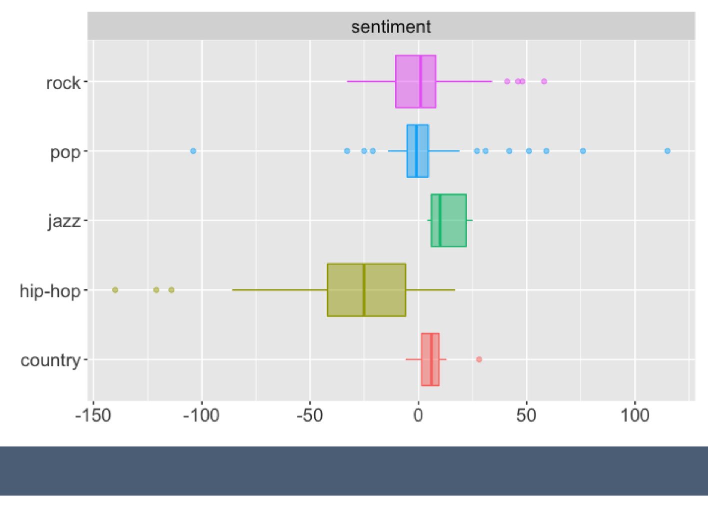
Number of words



Most common words

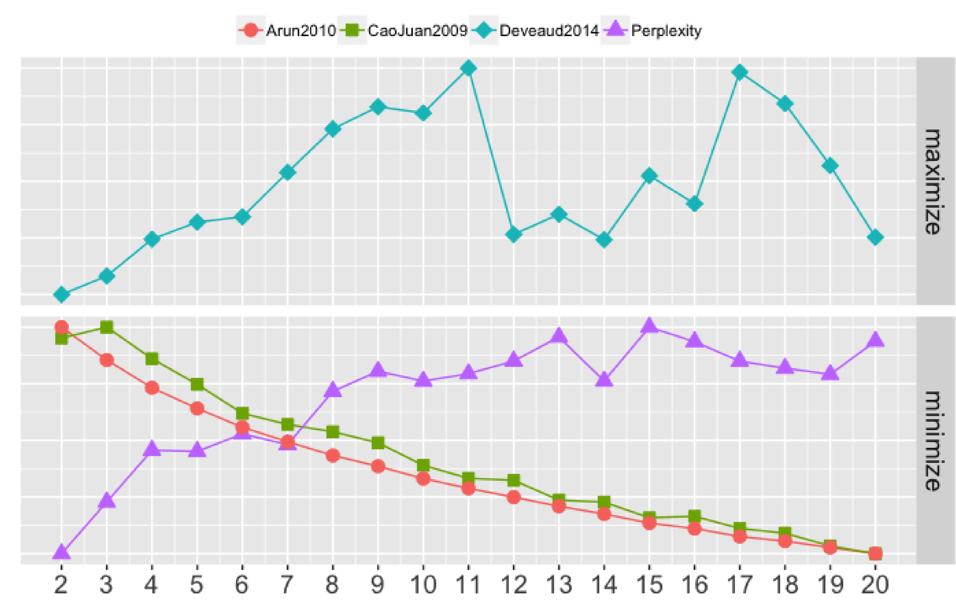


Sentiment distribution



6. Implementation

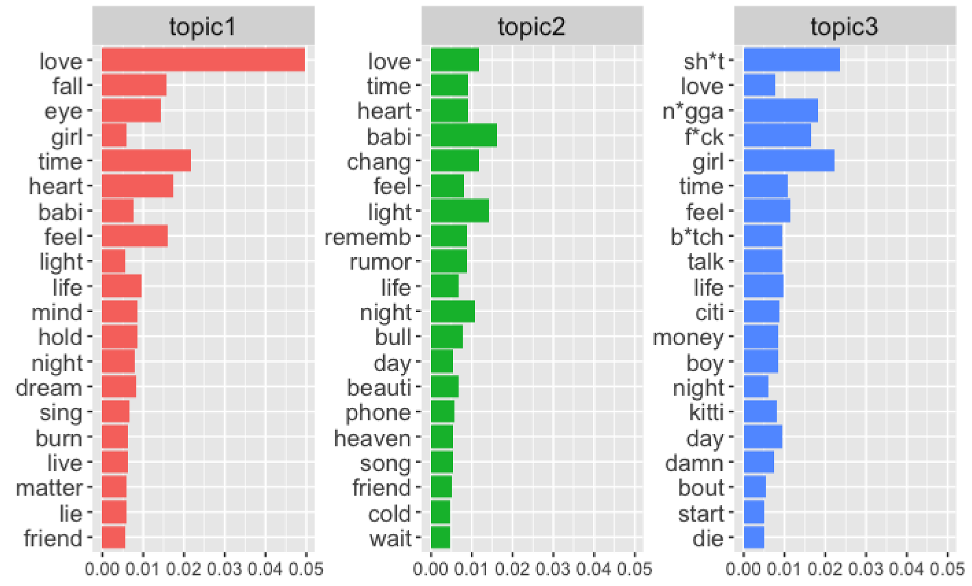
Number of topics



Number of topics

- Values scaled
- No unique results from metrics
- Values too high for interpretation
- Test of values of k between 2 and 5
- Use of human judgment for selection of best k
- $k = 3$ was chosen

Per-topic-per-word probability



Per-topic-per-word probability

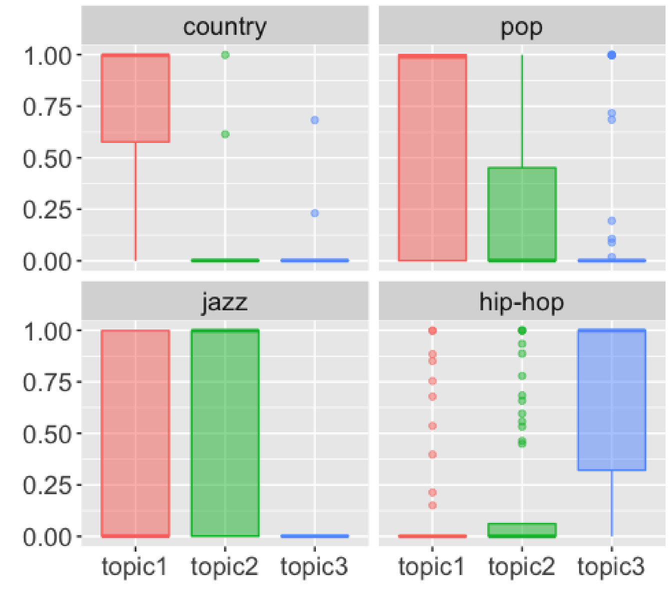
- Some common words across topics
- *topic1* and *topic2* similar, the former more love-related, the latter more general
- *topic3* is more hateful

Topics' weighted average sentiment

Topics' weighted average sentiment

- Sentiment values weighted with respective per-topic-per-word probability
- *topic1* most positive
- *topic2* most neutral, but with positive tendency
- *topic3* most negative

Per-document-per-topic probability



Per-document-per-topic probability

- Jazz and Rock split between *topic1* and *topic2*
- Pop partially split between *topic1* and *topic2*, but tending more towards *topic1*
- Country is associated with *topic1*
- Hip-Hop is almost fully explained by *topic3*

7. Results, conclusions and further research

- 3 topics after balancing topic intrusion and interpretability
- Quite clear distinction between Topic 3 and the others
- According to LDA with VEM estimation, the dataset can be split into 3 topics:
  1. Romantic love: mostly Pop and Country songs, part of Rock ones
  2. Everyday life: primarily Jazz songs, part of Rock ones
  3. Hustling: mainly Hip-Hop songs
- **Further research:**
  - Other estimation methods
  - Keyword re-ranking
  - Topic re-ranking

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3 (Jan), 993–1022
[2] 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.
[3] 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)
[4] Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique, 17 (1), 61–84.
[5] 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).
[6] Ponweiser, M. (2012). Latent dirichlet allocation in R.

For further information

Freie Universität Berlin

Silvia Ventoruzzo
silvia.ventoruzzo@gmail.com
Immatriculation number: 592252 (HU)

Freie Universität Berlin
Institute of Statistics and Econometrics
Seminar: Advanced Statistical Modeling