Assessing Quality of Unsupervised Topics in Song Lyrics

Lucas Sterckx, Thomas Demeester,
Johannes Deleu, Laurent Mertens, and Chris Develder

Ghent University - iMinds, Belgium
<firstname>.<lastname>@intec.ugent.be

Abstract. How useful are topic models based on song lyrics for applications in music information retrieval? Unsupervised topic models on text corpora are often difficult to interpret. Based on a large collection of lyrics, we investigate how well automatically generated topics are related to manual topic annotations. We propose to use the kurtosis metric to align unsupervised topics with a reference model of supervised topics. This metric is well-suited for topic assessments, as it turns out to be more strongly correlated with manual topic quality scores than existing measures for semantic coherence. We also show how it can be used for a detailed graphical topic quality assessment.

1 Introduction

This paper presents an analysis of how well topic models can be used to detect lyrical themes for use in Music Information Retrieval (MIR), an interdisciplinary science developing techniques including music recommendation.
Probabilistic topic models are a tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Latent Dirichlet Allocation (LDA) is a Bayesian graphical model for text document collections represented by bags-of-words [1]. In a topic model, each document in the collection of documents is modeled as a multinomial distribution over a chosen number of topics, each topic is a multinomial distribution over all words. We evaluate the quality and usefulness of topic models for new music recommendation applications.
Although lyricism and themes are undeniably contributing to a musical identity, they are often treated as mere secondary features, e.g., for obtaining music or artist similarity, which are dominantly determined by the audio signal. Nevertheless, previous works have analyzed lyrics, mainly aimed at determining the major themes they address. Mahadero et al. [2] performed a small scale evaluation of a probabilistic classifier, classifying lyrics into five manually applied thematic categories. Kleedorfer et al. [3] focused on topic detection in lyrics using an unsupervised statistical model called Non-negative Matrix Factorization (NMF) on 32,323 lyrics. After clustering by NMF, a limited evaluation was performed by a judgment of the most significant terms for each cluster. We expand

on this work by performing a large-scale evaluation of unsupervised topic models using a smaller dataset of labeled lyrics and a supervised topic model.
While state-of-the-art unsupervised topic models lead to reasonable statistical models of documents, they offer no guarantee of producing results that are interpretable by humans and require a thorough evaluation of the output. When considering lyrics, there is no general consensus on the amount and nature of the main themes, as opposed to news-corpora (sports, science,...). A useful topic model for MIR, appends the music with a representation of the thematic composition of the lyrics. For use in applications like music recommendation, playlist generation, ..., the topics should be interpretable. Evaluation methodologies based on statistical [1] or coherence [4] measures are not optimal for this purpose since they do not account for interpretability and relevance to the application. Chuang et al. [5] introduced a framework for the large-scale assessment of topical relevance using supervised topics and alignment between unsupervised and supervised topics.
Our contributions presented in this paper apply and build on aforementioned work, by assessing quality of unsupervised topics for use in MIR, and by introducing a new method for measuring and visualizing the quality of topical alignment, based on the kurtosis of the similarity between unsupervised topics and a reference set of supervised topics.
In Section 2, we present the data and our experimental set-up. The main topic model analysis is presented in Section 3, followed in Section 4 by conclusions.
We used the results presented below to create an online demo that demonstrates the use of high-quality topics for MIR with an application which automatically generates playlists based on preference of lyrical themes. This demo can be found at `http://users.ugent.be/~lusterck` (Login = `demo:ldamir`).

2 Experimental Setup

The main dataset used for this research is the 'Million Song Dataset'(MSD) [6], with metadata for one million songs, and lyrics as bags-of-words for a subset of 237,662 songs from a commercial lyrics catalogue, 'musiXmatch'.
LDA was applied on the set of lyrics, using the java-based package MALLET [7]. Three topic models were inferred from the subset of 181,892 English lyrics for evaluation[1], one with 60 (T60), 120 (T120) and 200 (T200) topics. A manual quality assessment of all of these topics was performed, with scores ranging from 1 (useless) to 3 (highly useful).
As an additional resource, a clean dataset of labels was provided by the website, 'GreenbookofSongs.com®'[2] (GOS), a searchable database of songs categorized by subject. This dataset contains 9,261 manually annotated song lyrics matched with the MSD (a small subsample of the GOS' complete database), with multiple labels from a large class-hierarchy of 24 super-categories with a total of 877 subcategories. Labeled Latent Dirichlet Allocation (L-LDA) is a variation of LDA

[1] Track-id's can be provided upon request
[2] `http://www.greenbookofsongs.com`, the authors would like to thank Lauren Virshup and Jeff Green for providing access to the GOS-database

for labeled corpora by incorporating user supervision in the form of a one-to-one mapping between topics and labels [8]. An L-LDA model with 38 supervised topics was inferred from the much smaller set of GOS-labeled lyrics, based on the GOS super-categories (but with the omission of minor categories like 'Tools', and splitting up of major categories like 'Love'). These are high-quality topics, but because of the limited size of the GOS data set, less representative for the entire scope of themes in the complete MSD lyrics collection.

3 Topic Model Assessment

The suitability of topic models for use in MIR is determined by the amount of relevant and interpretable topics they produce. We first introduce suitable metrics to evaluate to what extent unsupervised topics can be mapped to supervised topics obtained from tagged documents. We then show how these can be used as a better measure for the interpretability of topics than an existing alternative, and provide a visual analysis of topical alignment.

3.1 Measuring Topical Alignment

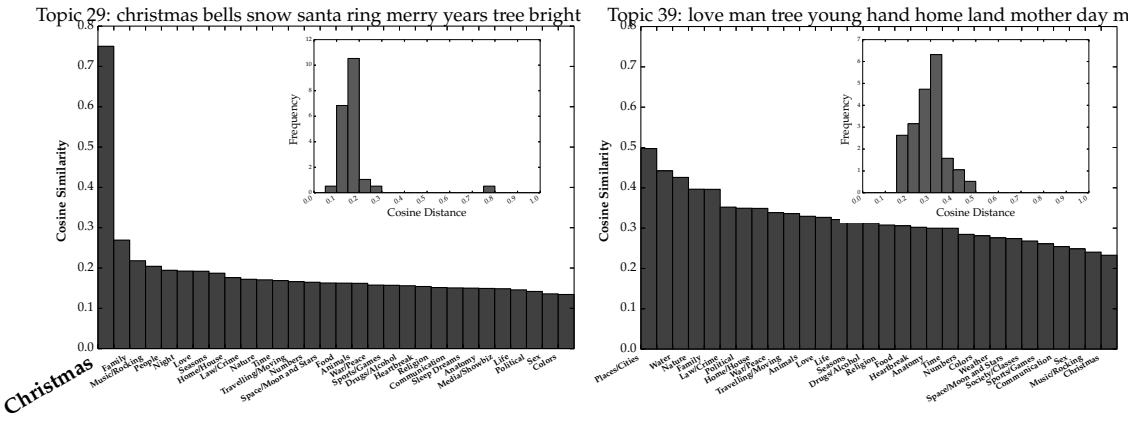
We define high-quality topics as topics for which a human judge finds a clear semantic coherence between the relevant terms in relation to an underlying concept (such as 'Love', or 'Christmas'). Such concepts are made explicit by an L-LDA model based on tagged documents, and we detect high-quality LDA-topics as those that bear a strong resemblance with L-LDA topics. For an unsupervised topic to represent a highly distinctive theme, ideally it should be highly similar to only a single supervised topic. For each of the unsupervised LDA-topics, the cosine similarity between the word-topic probability distribution is calculated with the distribution of each L-LDA topic.
We introduce two metrics to assess the distribution of these similarities per LDA-topic, which measure how strongly the variance of the mean cosine similarity depends on extreme values (in this case, because of similarities that are much higher than the average). The first is the excess kurtosis (γ_2), traditionally used to detect peakedness and heavy tails. The second is the normalized maximum similarity ($z_{\max}$), used in several outlier detection strategies.

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3, \qquad z_{\max} = \frac{X_{\max} - \mu}{\sigma} \tag{1}$$

with μ_4 the fourth moment about the mean μ, σ the standard deviation, and $X_{\max}$ the maximum similarity. Figure 1 shows the similarities with the unsupervised topics for the high-quality LDA-topic 29, with a clearly matched supervised topic (and high values for γ_2 and $z_{\max}$), and for the low-quality LDA-topic 39 (with low γ_2 and $z_{\max}$). The insets show the histograms of the similarities. Various other metrics were evaluated as well, but with a lower ability of detecting the interesting distributions.

3.2 Semantic Coherence

A second evaluation was performed using metrics presented in [4], where the authors show that measures for semantic coherence are highly correlated with



$\gamma_2 = 25,91$
$z_{\max} = 5.75$

$\gamma_2 = 0,15$
$z_{\max} = 2.70$

Fig. 1: Kurtosis measure and Normalized Maximum Similarity for topic evaluation

human quality judgments. These metrics use WordNet, a lexical ontology, to score topics by measuring the average distance between words of a topic using a variety of distance metrics based on the ontology. The best performing metric was reported to be the LESK-metric [9], based on lexical overlap in dictionary definitions. Table 1 shows the Spearman rank correlation between the LESK score for each topic and the manually assigned quality scores. For comparison, the rank correlation between the manual quality scores and γ_2 and $z_{\max}$ (as calculated in Section 3.1) are shown as well, and lead to significantly higher correlation values than with the LESK metric.

Table 1: Spearman correlations with manual quality-scores for the three topic models

Evaluation Metric	**T60**	**T120**	**T200**
Semantic Coherence using Wordnet (LESK)	0,35	0,23	0,31
Kurtosis (γ_2)	0,49	0,49	0,56
Normalized Maximum Similarity ($z_{\max}$)	0,49	0,50	0,53

3.3 Graphical Alignment of Topics

We can visualize the alignment between the supervised and unsupervised topics by calculating the kurtosis on the similarities between both topic sets. These are shown in Fig. 2, a *correspondence chart* similar to the one presented in [5], for the 60 topics LDA-model (T60). Our chart differs from the one presented in [5] in that it uses topics from an L-LDA model for the matching of unsupervised topics instead of a list of words generated by experts, and uses bar-charts to display the automatically calculated kurtosis scores instead of likelihoods of

human-assisted matching. The size of the circles denotes the cosine similarity between the corresponding supervised and unsupervised topics, and the coloring shows which concepts are matched in a one-to-one fashion by the unsupervised and supervised topics using the harmonic mean of both kurtosis' values. Note that the detection of topics is dependent on the labels included in the supervised data. High-quality LDA-topics, not present in the supervised set, are not detected. The chart shows that topics involving *Christmas*, *Fire* and *Water* are all very distinguishable by statistical models and human-assisted labeling, or resolved. Other topics are linked to more labels and contain fused concepts or junk. Another use of this chart is evaluating the reference topics by the experts of GOS. Some concepts devised by experts may be chosen too broadly. For example, the supervised topic of *Music/Rocking* is close in cosine-distance to Topic 6 and to Topic 54, which in turn is close to the supervised theme *Dancing/Party*. This indicates that labeling for *Music/Rocking* should be confined more to music and exclude songs about dancing. Topics like *Love* and *Heartbreak* correlate with many LDA-topics which demonstrate their dominance in lyrical themes.

4 Conclusion

This paper provides insights into the quality of topic models constructed from song lyrics. We showed that the kurtosis is a suitable metric to align unsupervised topics with supervised reference topics, which allows detecting high-quality topics in accordance to manual quality assessments.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3** (2003) 993–1022
2. Logan, B., Kositsky, A., Moreno, P.: Semantic analysis of song lyrics. In: International Conference on Multimedia and Expo, ICME 2004. Volume 2., IEEE (2004) 827–830
3. Kleedorfer, F., Knees, P., Pohle, T.: Oh oh oh whoah! towards automatic topic detection in song lyrics. In: Proceedings of the 9th ISMIR. (2008) 287–292
4. Newman, D., Karimi, S., Cavedon, L.: External evaluation of topic models. In: Australasian Document Computing Symposium (ADCS). (2009) 1–8
5. Chuang, J., Gupta, S., Manning, C.D., Heer, J.: Topic model diagnostics: Assessing domain relevance via topical alignment. In: ICML. (2013)
6. McFee, B., Bertin-Mahieux, T., Ellis, D.P., Lanckriet, G.R.: The million song dataset challenge. In: Proceedings of the 21st international conference companion on World Wide Web, ACM (2012) 909–916
7. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
8. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on EMNLP, ACL (2009) 248–256
9. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on Systems documentation, ACM (1986) 24–26

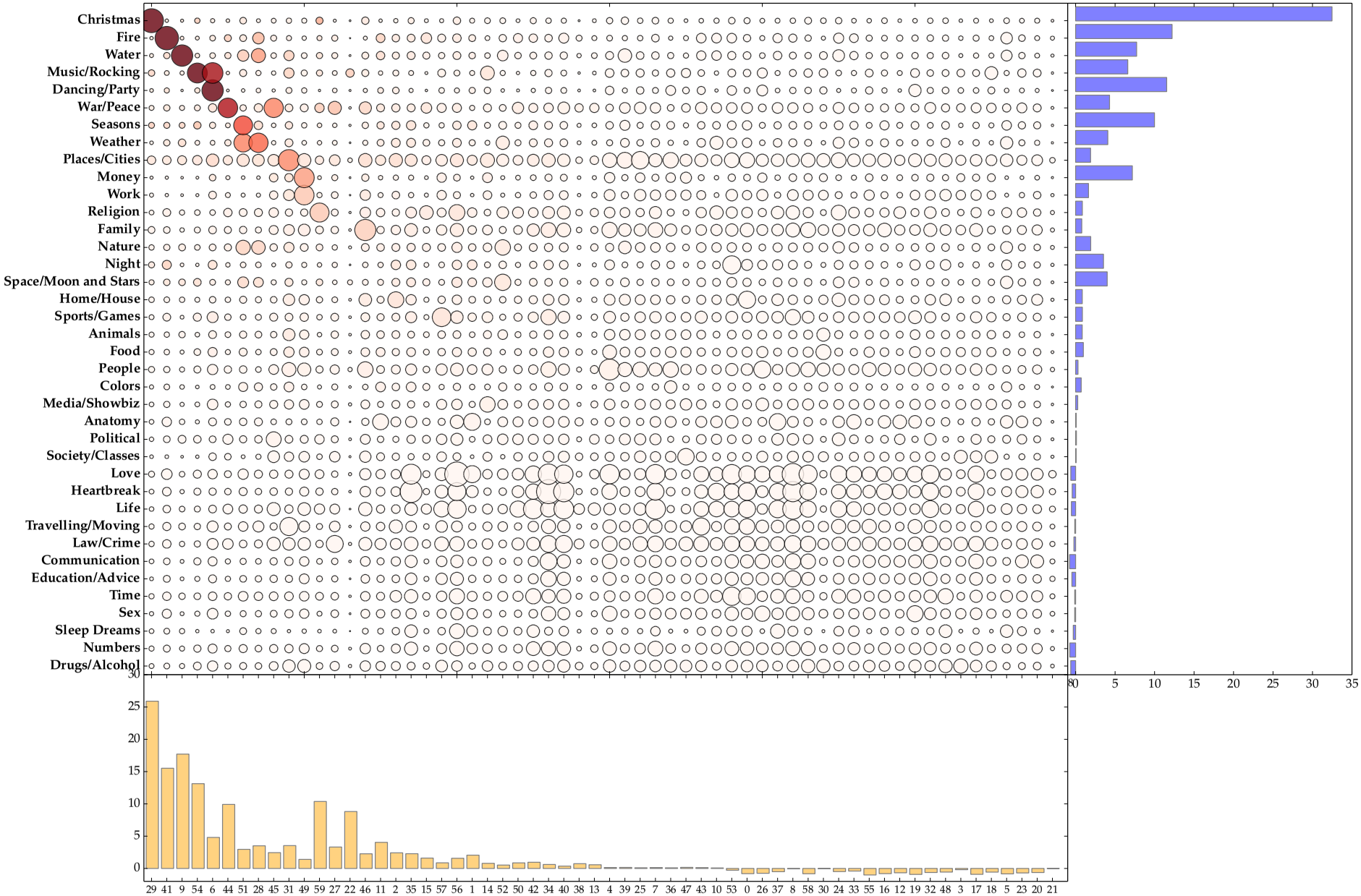


Fig. 2: Correspondence Chart between topics. The size of circles depicts cosine similarity between corresponding supervised and unsupervised topics. Bars on the sides of the graph show the kurtosis scores for their corresponding topics. High scores show that topics are aligned, low scores mean topics are not matched or junk in the case of LDA-topics. Circle coloring means a strong match between LDA and L-LDA topics, thus a useful unsupervised topic.