# A Comparison of Topic Modeling Approaches for a Comprehensive Corpus of Song Lyrics

**Alen Lukic**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
alukic@cs.cmu.edu

**Abstract**

In the interdisciplinary field of music information retrieval (MIR), applying topic modeling techniques to corpora of song lyrics remains an open area of research. In this paper, I discuss previous approaches to this problem and approaches which remain unexplored. I attempt to improve upon previous work by harvesting a larger, more comprehensive corpus of song lyrics. I will apply previously attempted topic modeling algorithms to this corpus, including latent Dirichlet allocation (LDA). In addition, I will also apply the Pachinko allocation (PAM) technique to this corpus. PAM is a directed acylic graph-based topic modeling algorithm which models correlations between topics as well as words and therefore has more expressive power than LDA. Currently, there are no documented research results which utilize the PAM technique for topic modeling in song lyrics. Finally, I will obtain a collection of human-annotated songs in order to apply a supervised topic modeling approach and use the results in order to produce one of the topic quality metrics which will be investigated in this paper.

## 1 Introduction

The field of music information retrieval encompasses the research efforts which address several problems, including genre classification, mood categorization, and topic detection in lyrics. The focus of this paper is the latter of these problems.

Within the study of topic modeling, topics are defined as a set of $k$ item $I = \{i_1, \ldots, i_N\}$ distributions $P_1(I), \ldots, P_k(I)$. In the context of song lyrics, each item is a word in the corpus vocabulary. The corpus discussed in this paper was extracted by a Web crawl over the SongMeanings website [1]. Each topic is a probability distribution over the items where items with higher probabilities of being drawn from the distribution are more representative of that topic. Typically, topics are represented by the $m$ most probable items in the distribution, where $m \ll N$. For example, the top 10 words in one topic determined by running latent Dirichlet allocation (LDA) over the documents in the SongMeanings corpus are {*dream, lovin, baby, night, good, lover, sleep, sweetheart, make, sad*}. One might assign the label "love" to this topic. In the case of this corpus, each document represented a bag-of-words model for the lyrics of a single artist.

In section 2, I will discuss related work in this domain and highlight potential areas for improvement and expansion. Section 3 describes the dataset upon which the results in this paper are based and the methodology used to acquire the data. Section 4 details the results of applying each of the following unsupervised statistical topic modeling techniques to the SongMeanings corpus: LDA, non-negative matrix factorization (NNMF), and Pachinko allocation (PAM) [2] [3] [4]. Section 5

discusses conclusions based on the evaluation results. Finally, section 6 outlines plans for future work in this area.

# 2 Related Work

Research focusing specifically on unsupervised topic modeling for song lyrics is somewhat sparse. In one of the earliest works on extracting topics from lyrical corpora, Logan et. al. applied latent semantic analysis (LSA) to the uspop2002 corpus and generated a set of topics which were used as features for calculating artist similarity [5] [6]. Kleedorfer, et. al. employed NNMF on a lyrics corpus extracted from Verisign's content-download platform to generate and manually label topic clusters [7]. Most recently, Streckx et. al. used LDA on a subset of the Million Song Dataset for the specific purpose of assessing the quality of the generated topics [8] [9]. I seek to expand upon this work in the following ways.

The corpora used in previous work were constructed from already-existing resources. There is no indication that the lyrics which constituted the corpora were a representative sample of song lyrics in general. I mitigate this problem by crawling SongMeanings and extracting all lyrics on the website. The resulting corpus is significantly larger and more likely to be a good sample than any previous corpora.

Topic modeling techniques used in previous works include LSA, LDA, and NNMF. However, there is no documented usage of the PAM technique. PAM is a graph-based algorithm which models correlations between topics in addition to correlations between words. As such, it is more expressive than LDA, which models only word correlations. Using PAM may result in higher quality topic clusters. I will test this hypothesis by applying each of the aforementioned techniques to the SongMeanings corpus and then using some set of standardized metrics to measure the quality of the topics produced by each method. One such metric is the maximum similarity between the unsupervised topics and supervised topics produced via Labeled LDA and manually annotated songs.

# 3 Dataset Methodology

The song lyrics used in this paper were acquired by crawling the entirety of the SongMeanings lyrics directory, processing the raw data, and creating a bag-of-words model for each artists' lyrics.

## 3.1 Crawler

The SongMeanings crawler was written in Python. Lyrics on SongMeanings are organized in alphabetic directories. Each directory contains several pages of artist links, and there are 50 links on each page. Artist links yield a single page containing all of that artist's songs. Subsequently, following song links yields the lyrics for that song. The crawler collected song lyrics assuming this structure held for all artists and songs on the website. Raw HTML was downloaded, from which pertinent information (e.g. artist links, song links, and the lyrics themselves) was extracted and processed. The crawler created a file on disk for each artist, and each artist file contained all of the song lyrics for that artist, each of which was separated by an explicit delineator.

## 3.2 Pre-processor

This program performed several pre-processing functions on the data. A number of the downloaded song lyrics were either empty (e.g. instrumentals), copyright-restricted, or non-English. The pre-processor removed all such lyrics from the corpus. Artists which ended up with no remaining songs

after this step were removed from the corpus altogether. Next, each artist file was converted to a bag-of-words representation. First, all stopwords were removed from the artist file, Then, the remaining words were lemmatized using the WordNet lemmatizer. I chose lemmatization over stemming because popular stemmers tend to actually perform rather poorly and often yield non-real words. Finally, after lemmatization, a word histogram replaced the original contents of the artist file, effectively transforming it into a bag-of-words model.

## 3.3 Corpus Statistics

The SongMeanings corpus statistics may be found below. By comparison, the uspop2002 corpus used in the paper by Logan et. al. contained 15,589 song lyrics, the Verisign corpus used in the paper by Kleedorfer et. al. contained 33,863 song lyrics, and the MSD corpus used in the paper by Streckx et. al. contained 181,892 song lyrics.

| **Statistic** | **Value** |
|---|---|
| Number of song lyrics | 763,491 |
| Number of artists | 118,438 |
| Average songs per artist | 6.446 |
| Vocabulary size | 276,685 |

The following are the top 25 most frequent terms in the corpus.

| **Rank** | | | | | |
|---|---|---|---|---|---|
| **1** | love | time | day | make | heart |
| **6** | night | eye | feel | life | thing |
| **11** | back | find | dream | hand | world |
| **16** | long | light | mind | give | hear |
| **21** | good | face | hold | man | thought |

# 4 Topic Model Evaluation

## 4.1 Latent Dirichlet Allocation

### 4.1.1 Qualitative Evaluation

I used the LDA implementation in the open-source MALLET package in order to produce topics from the SongMeanings corpus, choosing $k = 38$ in order to match the number of supervised topics obtained in the work by Streckx et. al [8] [11]. The following word sets represent select unsupervised topics generated by the algorithm which match supervised topics in the aforementioned work.

Select Topics

| **Matched Label** | **Word Set** |
|---|---|
| Heartbreak | thing begin break face heart strange remind |
| Fire | run head fire dead burn back handy |
| War/Peace | people world war make human life line |
| Religious | god lord heaven angel jesus man hand |
| Dance | dance move shake ready party beat body |
| Life | life world live day time find give |
| Christmas | town city santa comin claus york watch |
| Nature | sea water river wind ocean wave tree |
| Music/Rocking | song sing hear music play sound singing |
| Traveling/Moving | home ride road back train town man |
| Sex | hot sugar sweet candy honey sex ice |
| Night/Dreaming | night tonight light shine star dream morning |

There were also several generated topics which did not seem to fit well into any of the manually annotated topics.

Novel Topics

| **Given Label** | **Word Set** |
|---|---|
| Rap | back shit man rhyme make rap check |
| Anger/Frustration | fuck money fucking shit hate give sick |
| Evil/Death | blood death black soul fire evil dead |

Other topics closely resembled one of the aforementioned topics.

### 4.1.2 Quantitative Evaluation

In order to replicate the evaluation methodology by Streckx, et. al., I divided the data into training and validation sets, and had the validation set labeled by human annotators using the same set of labels employed by Streckx, et. al. in their work. I then employed the Labeled LDA algorithm in the Stanford Topic Modeling Toolbox in order to generate a supervised word distributions against which to compare the unsupervised word distributuons generated by LDA over SongMeanings training set [12]. I then measured the similarity between each supervised and each unsupervised distribution using the Kullback-Leibler distance metric in order to compute the normalized maximum similarity for each unsupervised topic. The following table shows the minimum, average, and maximum values for the normalized maximum similarities and kurtoses values for each unsupervised topic.

| Statistic | Minimum | Mean | Maximum |
|---|---|---|---|
| Normalized maximum similarity | 4.087 | 4.342 | 4.601 |
| Kurtosis | 8.619 | 9.741 | 11.065 |

### 4.1.3 Analysis

The results indicate that the distributional similarities between the unsupervised and supervised topics were generally invariant. The range of similarities is very small, as is the distributional skewness as measured by the kurtosis. The most likely explanation for this is that the validation set was simply too small. There were well over 700,000 songs in the training set, but only approximately 8,000 in the validation set; of those, less than 500 were assigned labels by human annotators. As such, it is possible that such a small validation set was simply not a representative sample of the entire song lyrics corpus, and none of the learned topics aligned well with the human-annotated distributions.

## 4.2 Pachinko Allocation

### 4.2.1 Qualitative Evaluation

Pachinko allocation produces two sets of topics. The first is a superset of topics, which is itself composed of related topics. The second is a subset of topics, which are equivalent to word distributions as determined by algorithms like latent Dirichlet allocation. The subtopics produced by a run of PAM with the supertopic parameter set to 5 and the subtopic parameter set to 20 produced analogous subtopics to that produced by LDA on the same corpus. On the other hand, the 5 supertopics produced were virtually non-informative due to their extremely similar nature. For reference, the supertopics are listed in the below tables. The subtopics are listed in order of decreasing likelihood of belonging to the supertopic.

Supertopic 0

| Subtopic ID | Word Set |
|---|---|
| 0 | feel back eye run hold inside turn fall heart head |
| 14 | life time world live find mind day give free make |
| 3 | make thing good time gonna feel friend wrong bad care |
| 1 | back time day long remember left thought knew home year |
| 11 | blood life dead die eye death lie pain soul fear |
| 12 | love heart cry feel true give make baby hold time |
| 4 | girl drink hair head put boy house car big red |
| 8 | sun wind sky sea water rain river tree cold light |
| 19 | night day light tonight dream wait morning star sleep long |
| 6 | people man war world gun kid fight line power make |

Supertopic 1

| Subtopic ID | Word Set |
|---|---|
| 0 | feel back eye run hold inside turn fall heart head |
| 14 | life time world live find mind day give free make |
| 3 | make thing good time gonna feel friend wrong bad care |
| 1 | back time day long remember left thought knew home year |
| 11 | blood life dead die eye death lie pain soul fear |
| 12 | love heart cry feel true give make baby hold time |
| 4 | girl drink hair head put boy house car big red |
| 8 | sun wind sky sea water rain river tree cold light |
| 6 | people man war world gun kid fight line power make |
| 19 | night day light tonight dream wait morning star sleep long |

Supertopic 2
(isomorphic to supertopic 1)

Supertopic 3
(isomorphic to supertopic 0)

Supertopic 4
(isomorphic to supertopic 0)

#### 4.2.2 Analysis

The homogeneity of the supertopics produced by Pachinko allocation indicate that there is a small subset of latent subtopics in the corpus which are highly related to such an extent that they occupy each of the supertopics. By extension, this implies that the remaining subtopics are generally unrelated to each other. The difference between these two sets of subtopics possibly points to a quality difference between them; that is, topics produced because they are of high quality, and topics which consist mainly of noise in the data.

## 5 Conclusions

The experiments confirm the hypothesis that new topics, previously undetected in state-of-the-art papers such as the work by Streckx, et. al, would surface when topic modeling algorithms were applied to a comprehensive collection of song lyrics. Investigating Pachinko allocation revealed a homogeneic composition of supertopics, indicating some previously undetected relatedness between the subtopics which compose those supertopics. Finally, the lack of topical alignment between the unsupervised and supervised topics is likely a symptom of insufficient data in the human-annotated validation set.

# 6 Future Work

The most important improvement necessary to better analyze and evaluate the efficacy of the topic modeling algorithms mentioned in this paper on the SongMeanings corpus is a significant expansion in the amount of human-annotated data. For this purpose, a crowdsourcing platform like Amazon's Mechanical Turk would prove beneficial [13]. Also, developing a framework for efficiently exploring topic models' parameters is worth consideration, as model parameters significantly impact the quality of the model produced.

# References


[1] SongMeanings. http://songmeanings.net/

[2] Blei, D.; Ng, A.Y; & Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3 (pp. 993 - 1022), 2003.*

[3] Lee, D.D.; & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature 401 (pp. 788-791), 1999.*

[4] Li, W.; & McCallum, A. Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. *In Proceedings of the* $23^{rd}$ *International Conference on Machine Learning, 2006.*

[5] Logan, B.; Kositsky, A.; & Moreno, P. Semantic Analysis of Song Lyrics. *In Proceedings of IEEE International Conference on Multimedia and Expo, 2004.*

[6] The "uspop2002" Pop Music data set. http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html

[7] Kleedorfer, F.; Knees, P.; & Pohle, T. Oh Oh Oh Whoah! Towards Automatic Topic Detection in Song Lyrics. *In Proceedings of ISMIR, 2008.*

[8] Sterckx, L.; Demeester, T.; Deleu, J.; Mertens, L. & Develder, C. Assessing quality of unsupervised topics in song lyrics. *In Proceedings of ECIR, 2014.*

[9] Million Song Dataset. http://labrosa.ee.columbia.edu/millionsong/

[10] Lesk, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *In Proceedings of SIGDOC, 1986.*

[11] MALLET. http://mallet.cs.umass.edu/

[12] Stanford Topic Modeling Toolbox. http://www-nlp.stanford.edu/software/tmt/tmt-0.4/

[13] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome