# Topic Modelling

Deepak Mahapatra, Mridula Maddukuri and Gopika Jayadev

November 2016

## 1 Introduction

Data exploration and text mining have become very popularly used tool for decision making and other applications in the recent times. One of the interesting and useful part in this is topic modeling. Topic modeling is used for a variety of tasks including but not limited to compact document topic representation, topic identification, rough category classifications etc. Topic models provide a simple way to analyze large volumes of unlabeled text and to identify any underlying semantic structure in a set of documents.

Topic modeling can be classified as semi-supervised or unsupervised learning, but the discussions in this project are limited to the unsupervised versions of topic modeling. Topic modeling can be performed using various techniques like LDA [2], HDP [3], NMF [4], the most common and arguably the most accurate is Latent Dirichilet Allocation (LDA). The project uses two of the above discussed techniques, namely NMF and LDA. The generative model on Non-negative matrix factorization (NMF) is of a more deterministic nature mainly due to the use of a point estimate for the parameters. LDA on the other hand uses a Bayesian estimation for the parameters which results in the reduction of over-fitting when applying to huge data sets. We compare the performance of both these algorithms on large real world document corpuses.

### 1.1 Latent Dirichilet Allocation (LDA)

LDA assumes a probabilistic generative model for each document. Let us assume we know $K$ topic distributions for our dataset ($K$ multinomials containing V elements each), where V is the number of words in our corpus. Let $\beta_i$ represent the multinomial for the $i$-th topic, where the size of $\beta_i$. Given these distributions, the LDA generative process is as follows. For each document:

1. randomly choose a distribution over topics (a multinomial of length $K$)
2. for each word in the document:
   (a) Probabilistically draw one of the $K$ topics from the distribution over topics obtained in (1), say topic $\beta_j$
   (b) Probabilistically draw one of the V words from $\beta_j$

The LDA algorithm [2] infers the parameters of this generative model given data, through variational inference. We use the Sklearn library version of LDA, on the document corpus which we pre-process ourselves.

### 1.2 Non-Negative Matrix Factorization (NMF)

Suppose there are $N$ documents and $M$ words in the vocabulary. NMF assumes that the document versus word matrix (expressed as counts or tf-idf weights) can be factorized in to non-negative factors whose latent dimension $K \ll N, M$. Thus the matrix of documents and words, $\mathbf{Y} \in \mathbb{R}^{N \times M}$ can be factorized as : $\mathbf{Y} = \mathbf{AW}$, where $\mathbf{A} \in \mathbb{R}^{N \times K}$ and $\mathbf{W} \in \mathbb{R}^{K \times M}$. The $K$ rows of $\mathbf{W}$ have the interpretation of topics and the words with highest weights in these are the key words in the topics.

The NMF algorithm infers these factors given data, provided some assumptions are satisfied [4]. We implement the algorithm in [4] ourselves and also compare the results with the NMF library function in Python.

## 2 Data Description and Preprocessing

We have tested all the algorithms on various data-sets like the NIPS full paper data-set, AP news and New York times data-set. For ease of exposition we only include the results from the AP news [1] data-set only. It has 2246 documents. Each document can be represented as a bag of words, and vocabulary is the total set of words in the data. Thereby, the document vs word count matrix can be formed after the following pre-processing steps which we implement (note that for NMF the count matrix is further processed into tf-idf weights).

- Tokenizing the document: Breaking into array of words
- Stop word removal: Removal of words which do not add any meaning to the topics
- Stemming: Merging of words that are equivalent in meaning like singular and plural versions of the same word.

## 3 Methodology and Description

The steps we follow in our experiments are as follows:

- Preprocess Data as in Section 2.
- Create document vs word count matrix. If the method is NMF we further preprocess to for the tf-idf matrix.
- Apply the algorithm to get topics and the following list of words.
- Human supervision to infer each of the topic qualitatively from keywords.
- Qualitative validation of the results.

## 4 Results

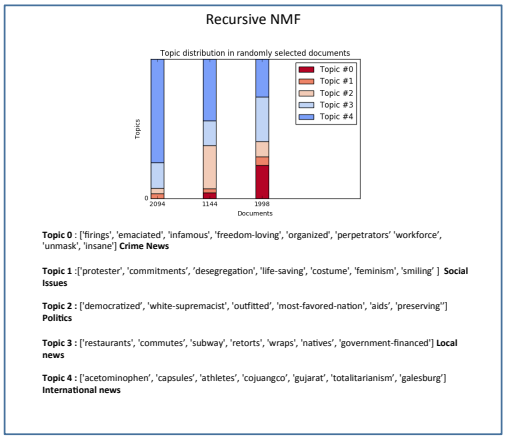
In this section we present our main experimental results. The topics have been qualitatively inferred from the main key word. The topic distributions in some of the chosen documents have also been displayed as a bar plot, for better interpretation.

### 4.1 Results from NMF

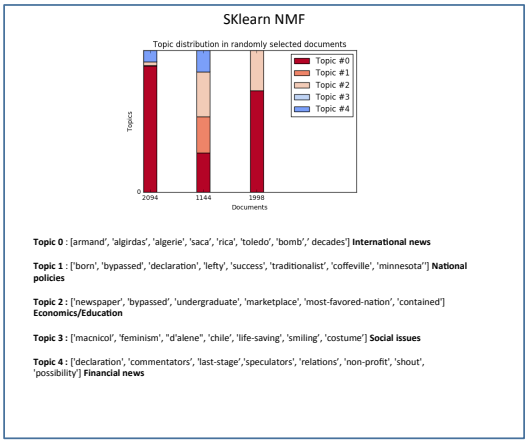
In this section we compare the results obtained from our implementation of the NMF algorithm in [4] with that of the NMF library in Python. We show the results in Figure 1. The topics have been inferred from the keywords (words with highest weights). We see almost comparable performance qualitatively. The distribution of the topics in three randomly chosen documents have been plotted under both the algorithms.We see that our version has slightly more spread in these topic distributions, in the same documents.

### 4.2 Results from LDA

In this section we present the results of LDA on the same data-set. We demonstrate the results through topic distributions in the documents and the key-words in the documents. Figure 2a, while in Figure 2b we plot the the word distributions in different topics. The results obtained are very easily interpretable, mainly owing to the generative model of LDA.

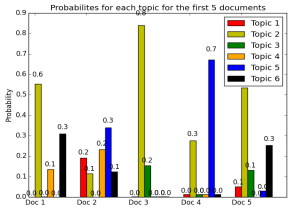



(a) Our Implementation




(b) SKlearn Implementation

Figure 1: Comparison of library implementation of NMF vs fast recursive NMF implemented separately by us. We factorize into 5 topics in both the cases. The topic distribution is plotted for three randomly chosen documents.




Doc 1: Seizure of tax payments made by U.S. businesses operating in Panama

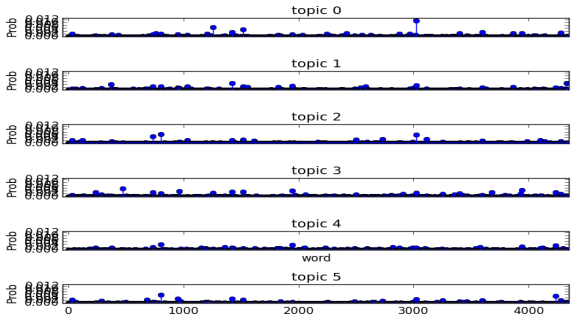
Doc 2: Protected species of alligator were smuggled

Doc 3: Democratic caucuses in Michigan

Doc 4:Death of National Front after party leader Jean-Marie Le Pen

Doc 5: Flood in some states in India

(a) Topic distribution in some documents and keywords in document.




(b) Topic vs. word distributions

Figure 2: Results from LDA applied to AP data-set. The topics have been inferred from the key-words.

# 5 Conclusion

LDA produces more humanly distinguishable topics according to keywords on a larger data set as compared to NMF. The number of topics can be varied as required to get pristine classifications and results can be cross-validated The performance of the algorithm is evaluated by human in the loop technique wherein the user can successfully distinguish all words on each topic easily. A possible future direction is to explore quantitative methods to validate and compare the results of different topic modeling algorithms.

# References


[1] David Blei. AP News Data-Set. http://www.cs.princeton.edu/~blei/lda-c/, 2008.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Jordan L Boyd-Graber and David M Blei. Syntactic topic models. In *Advances in neural information processing systems*, pages 185–192, 2009.

[4] Nicolas Gillis and Stephen A Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):698–714, 2014.