PRML Course Project Evaluation

# Topic Modeling using LDA


*Akshat Karani, Hritik Kumar & Arpit Agrawal*

*Indian Institute of Technology Dharwad*

*Computer Science and Engineering*

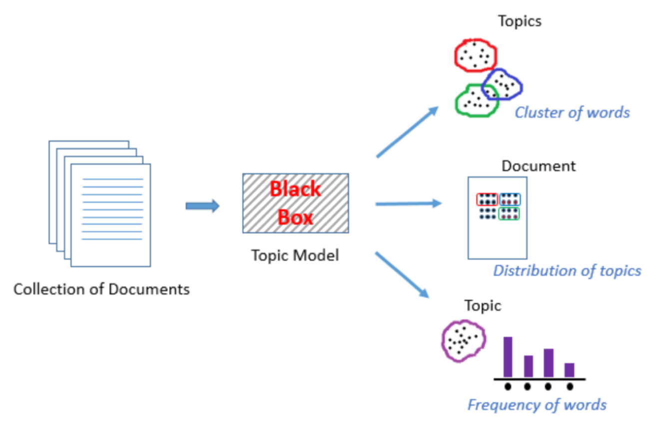
*Email: {170010003,13,40}@iitdh.ac.in*


## Abstract


In the project we have done topic modeling using Latent Dirichlet Allocation to find the topics in the IIT Dharwad broadcast E-mail.
We explored the mathematics behind LDA and it's other applications like using LDA for images. We also used LDA to model topics in the popular 20 newsgroups dataset.


## Topic Modeling

A topic model is an unsupervised technique to discover topics across various text documents. These topics are abstract in nature, i.e., words that are related to each other form a topic. There can be multiple topics in an individual document.
Topic modeling helps in exploring large amounts of text data, finding clusters of words, the similarity between documents, and discovering abstract topics.



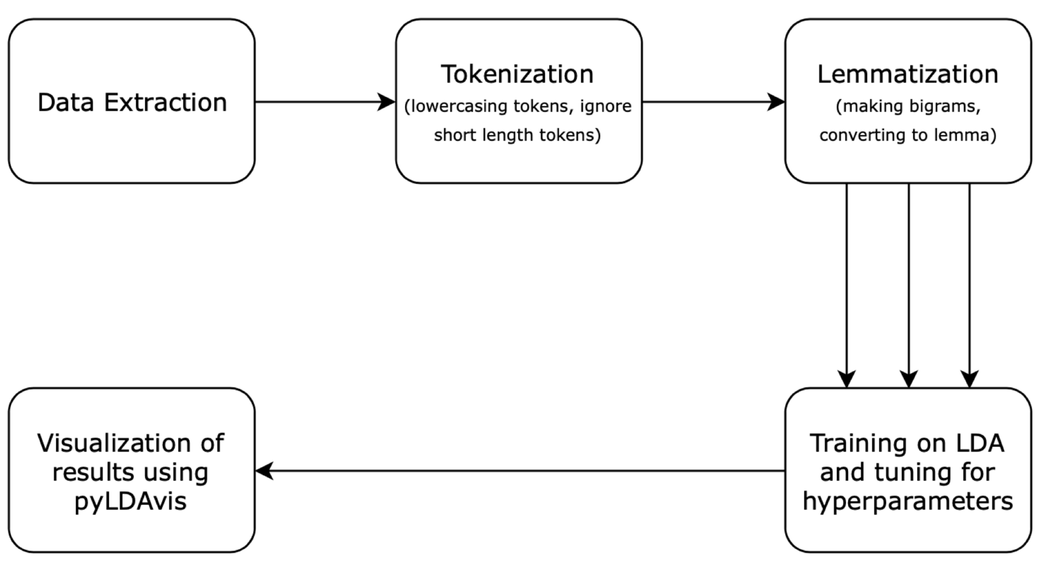

## Introduction to LDA

Latent Dirichlet Allocation (LDA) is an example of a topic model and is used to classify text in a document to topics. In this project, we applied LDA to a set of documents that we collected from the IIT Dharwad broadcast E-mail.
Input given to LDA is a document term matrix and the number of topics, each document is represented as a Bag of Words.
LDA is a generative model and the generative process is as follows:

- Each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words.
- It also assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution.
- Now, a document is assumed to be generated as follows: first, you select a distribution over the topics.
- Then draw a topic from this distribution, then draw a word from the distribution over words corresponding to the topic. This is the first word of the document.
- This is repeated for all the words in the document.

LDA then uses a Variational EM algorithm to estimate the topic distributions and the distribution of words for each topic during training.

## Basic Block Diagram




## Proposed Modeling Scheme

### Training Phase :-

Following process involved in building the LDA model for the dataset:

- Pre-processing the raw text, removing extra characters, punctuations and stop words.
- Tokenization - Convert a document into a list of lowercase tokens, ignoring tokens that are too short or too long.
- Making bigrams of a document.
- Lemmatization - Words in the third person are changed to first person and verbs in past and future tenses are changed into the present.
- Converting text of bag of words:

1. Prior to topic modeling, we convert the tokenized and lemmatized text to a bag of words — which you can think of as a dictionary where the key is the word and value is the number of times that word occurs in the entire corpus.
2. Now for each pre-processed document we use the dictionary object just created to convert that document into a bag of words.

- After this, the data is fed to the LDA model. The output obtained is document-topic distribution and topics-word distribution.

for each word   select topic   select word from topic

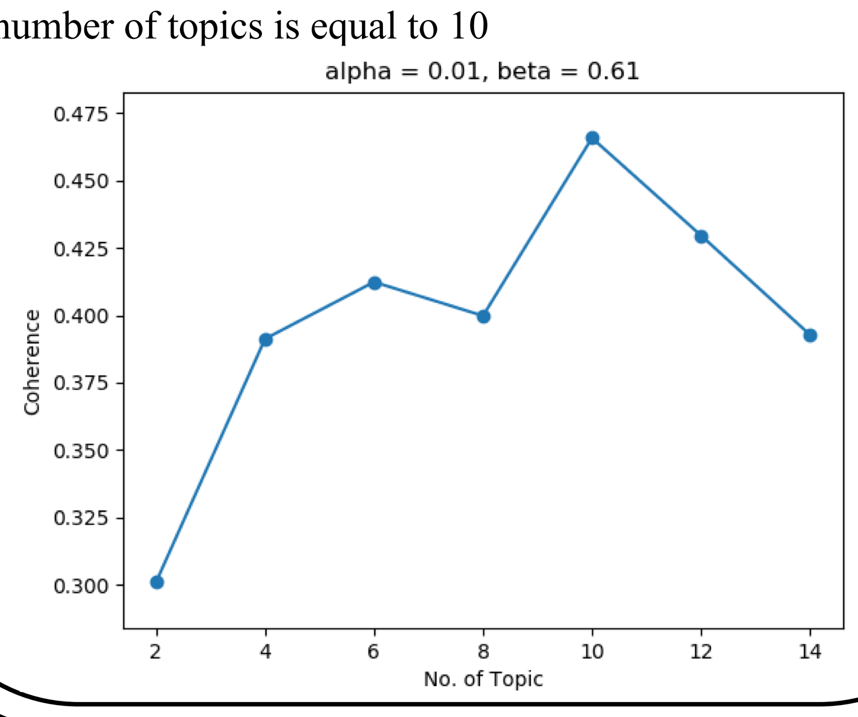
$$p(W, Z, \Theta) = \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn})$$

for each document   generate topic probabilities

LDA tries to find the joint posterior probability of distribution of topics, one for each document, N topics for each document, a distribution of words, one for each topic given the corpus.
Variational EM algorithm can be used for this.

### Tuning of Hyperparameters

We had to tune for finding the optimal number of topics . We modeled with varying number of topics (2, 4, …, 14) and chose the one for which the coherence score was maximum.
Coherence score was observed to be maximum when the number of topics is equal to 10



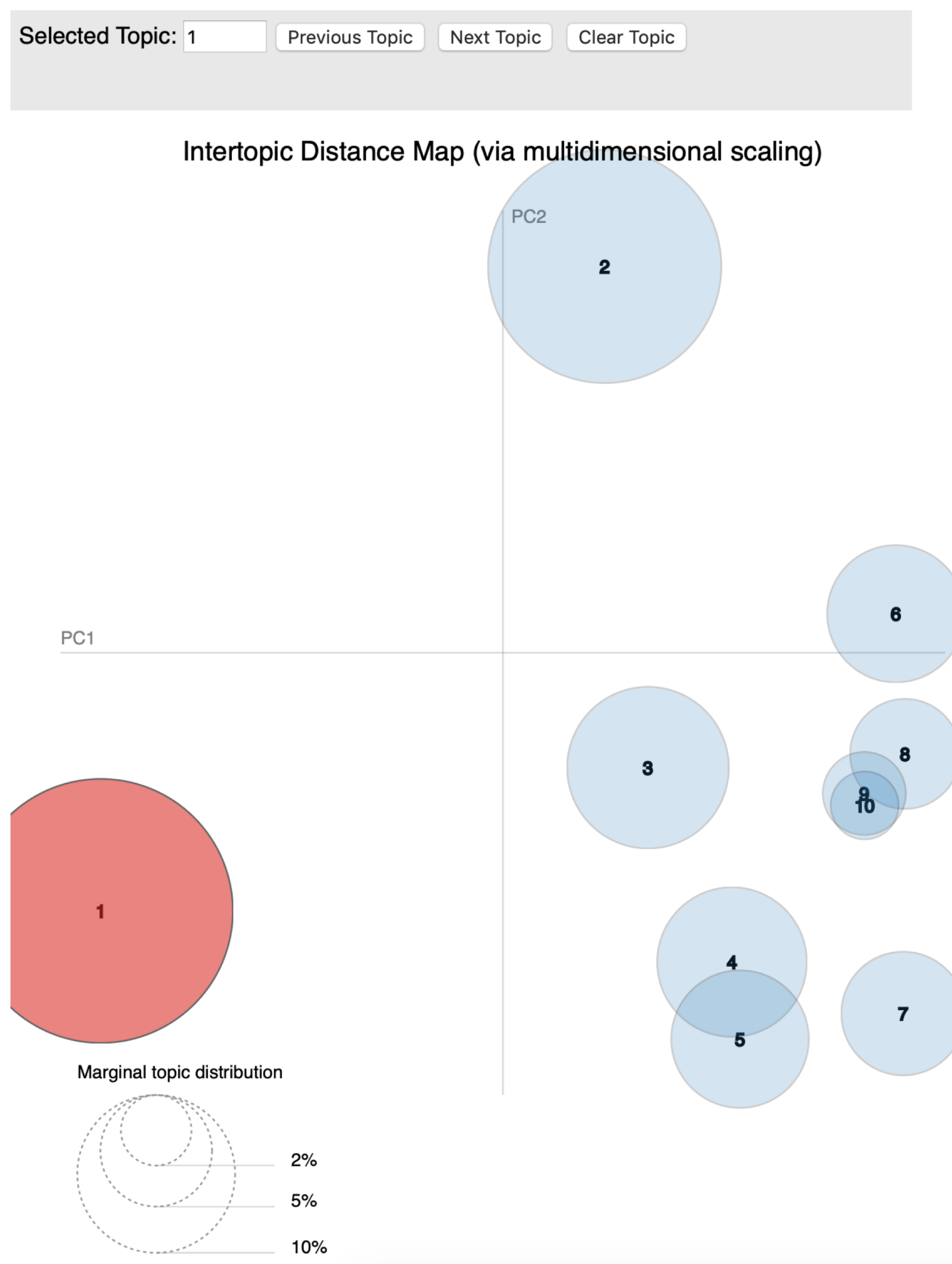

## Results

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| event | intended recipient | team | hostel | course |
| team | strictly prohibited | sport | ashrith adepu | academic |
| club | confidential | match | election | office |
| registration | unauthorized review | player | campus | slot |
| workshop | disclosure dissemination | football | vote | notice |
| competition | copying | table tennis | general | credit course |
| participate | unlawful | tournament | manifesto | general |
| register | print | game | contestant | schedule |
| cultural | notify sender | final | wifi | rule |
| technical | action taken | schedule | cycle | staff |

## Results

| Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|
| talk | music | water | course | analysis |
| research | internship | library | internet | seminar |
| speaker | program | journal | speed | mechanical |
| university | company | omprakash bhendigeri | rajshekar | wind farm |
| design | yourdost | questions panel | wifi | mathematic |
| development | apply | cube | network | simulation |
| science | expert | fearless step | computer science | patil |
| application | skill | rubik | sandeep | amlan |
| communication | saurav dosi | interspeech | moodle | research |
| program | scholarship | corpus | password | turbulence |

## Visualization




## Conclusion

We had to tune for finding the optimal number of topics. We modeled with a varying number of topics (2, 4, …, 14) and chose the one for which the coherence score was maximum.
Coherence score was observed to be maximum when the number of topics is equal to 10

## Acknowledgement


We would like to acknowledge Prof. Prabuchandran K J for mentoring us throughout this project. We would also like to thank Prof. SR Mahadeva Prasanna for providing us with this opportunity.


## References


[1] David M. Blei, Andrew Y. Ng and Michael I. Jordan. Latent Dirichlet Allocation, Journal of Machine Learning Research 3
[2] Carson Sievert and Kenneth E. Shirley. LDAvis: A method for visualizing and interpreting topics
[3] Jason Chuang, Daniel Ramage, Christopher D. Man- ning and Jeffrey Heer. 2012a. *Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis*. CHI.
[4] pyLDAvis(https://pyldavis.readthedocs.io/en/latest/index.html)
[5] genism(https://radimrehurek.com/gensim/index.html)