Topic Modeling

Presented by
Ayesha Asif
2017-MS-EE-100

Problem Statement

Extracting topics form a set of documents and finding probability of topics over documents using topic modeling.

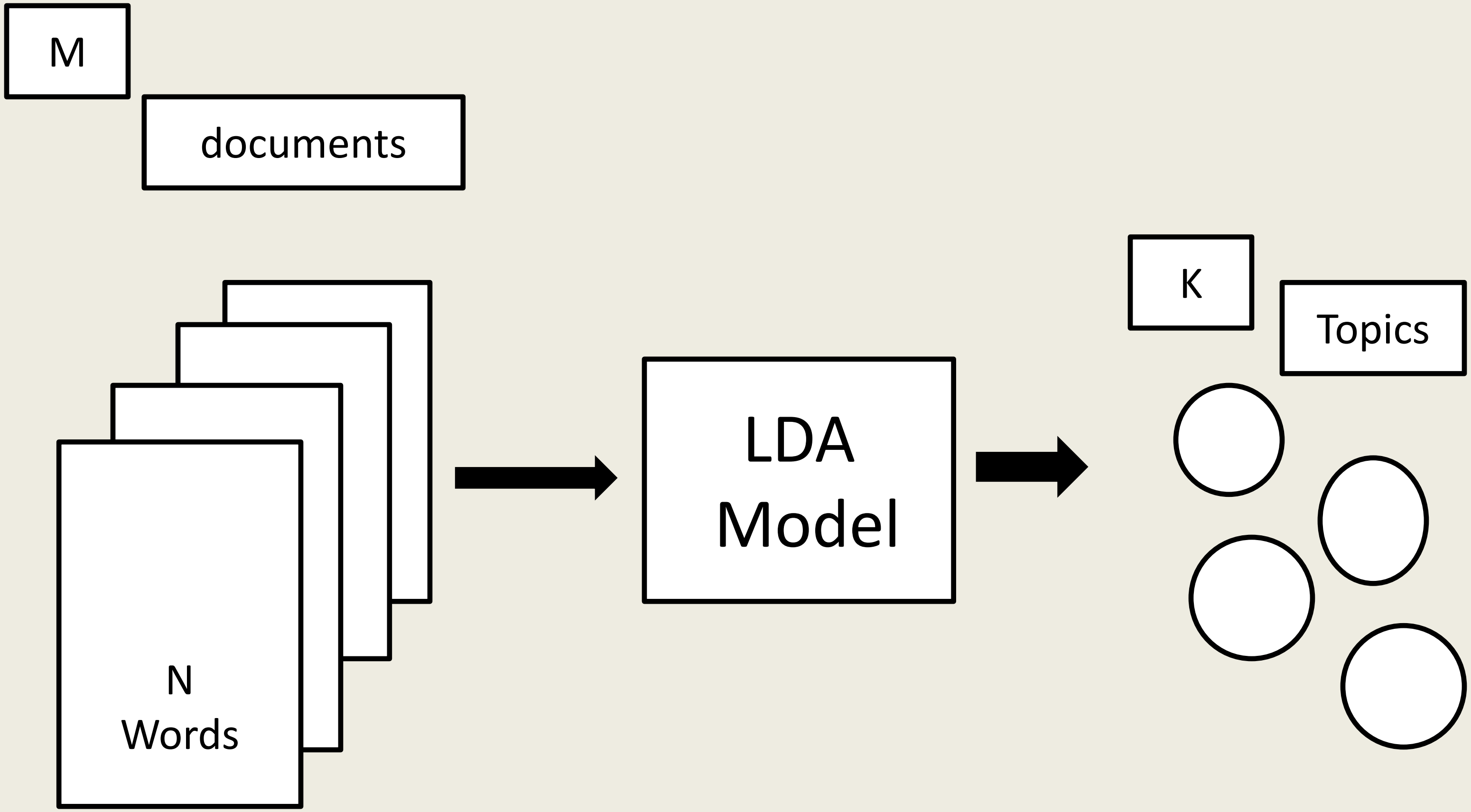
Introduction

Belongs to Unsupervised Learning

A topic model is a statistical model for discovering the abstract “topics” and the hidden thematic structure that occur in a collection of documents.

A topic consists of cluster of words that occur frequently together.

Methodology



M

documents

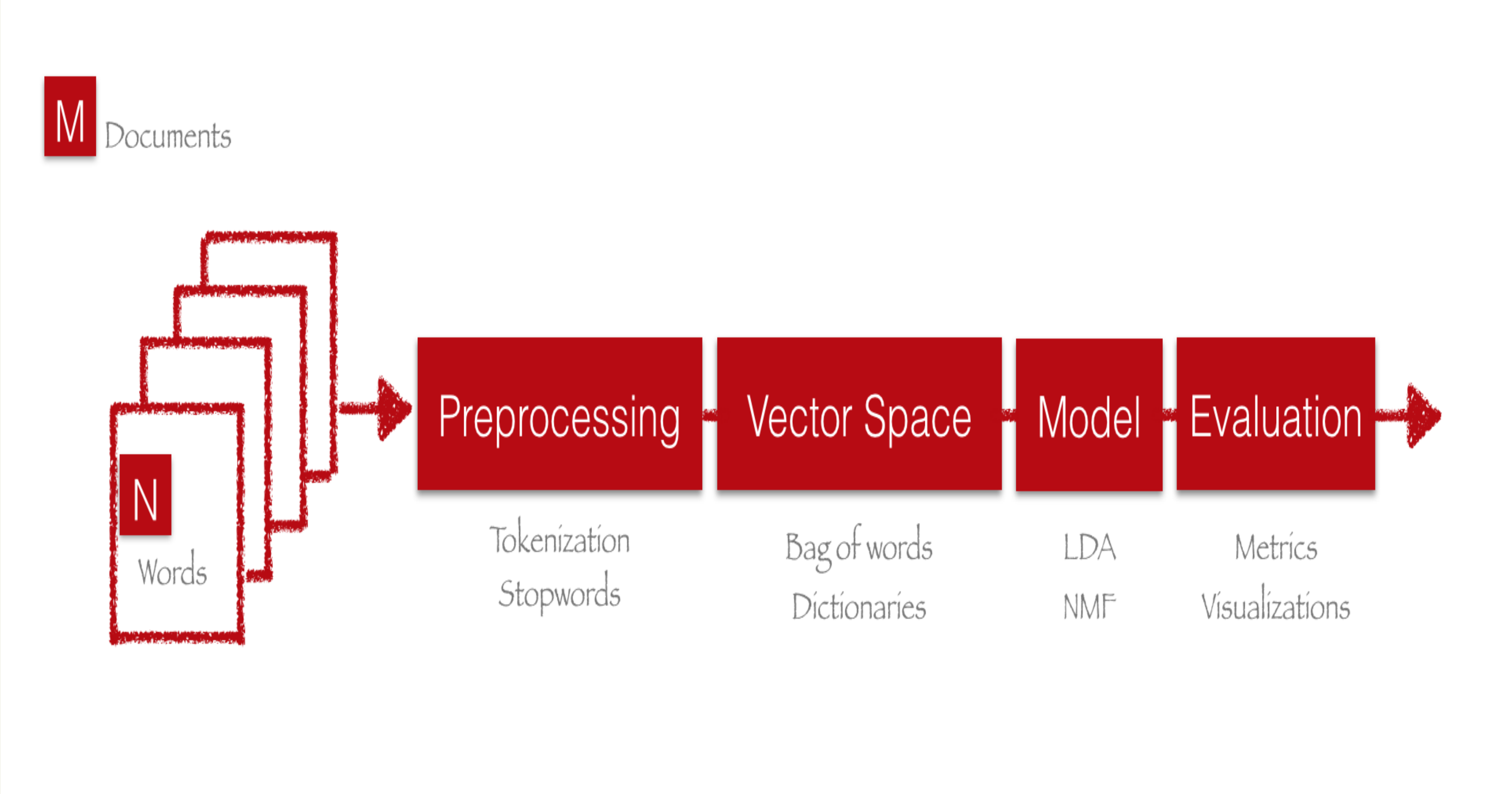
N
Words

LDA
Model

K

Topics

LDA Model


M Documents
N
Words
Preprocessing
Vector Space
Model
Evaluation
Tokenization
Stopwords
Bag of words
Dictionaries
LDA
NMF
Metrics
Visualizations

Preprocessing

- Tokenizing

  separating each word from each document

- Removing Stop Words

  removing words like a, the, of, and …

- Stemming

  removing form of verbs like ing, ed …

Vector Space

- Dictionary

  cricket, technology, investment, ….

- Corpus – Bag of words

  coverts the words to its integer id, and count the number of occurrence of words in each document

Model

- TF model

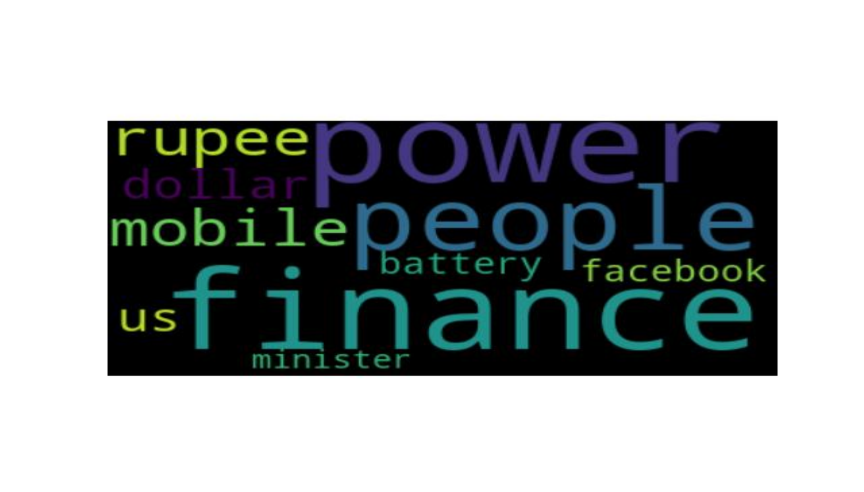
Term frequency model tells us how important a word is to the model and its value increases in proportionality to the number of times a word appears in a document.

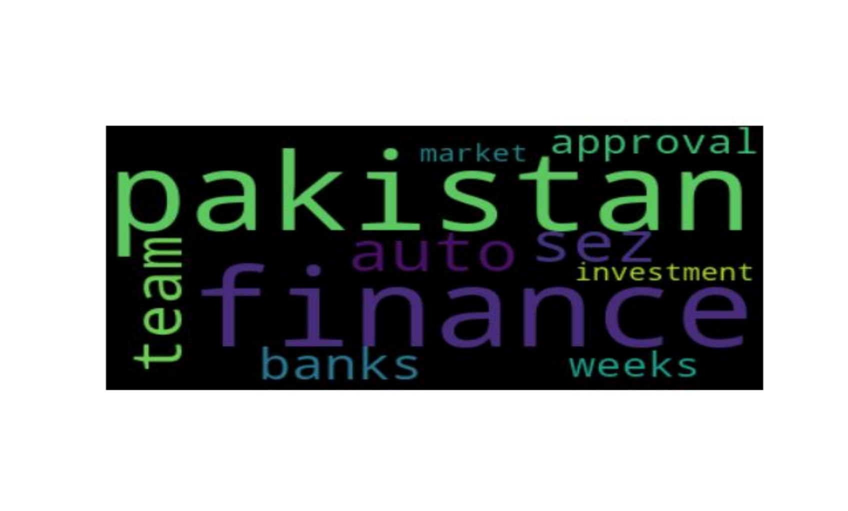
- LDA (Latent-Dirichlet allocation)

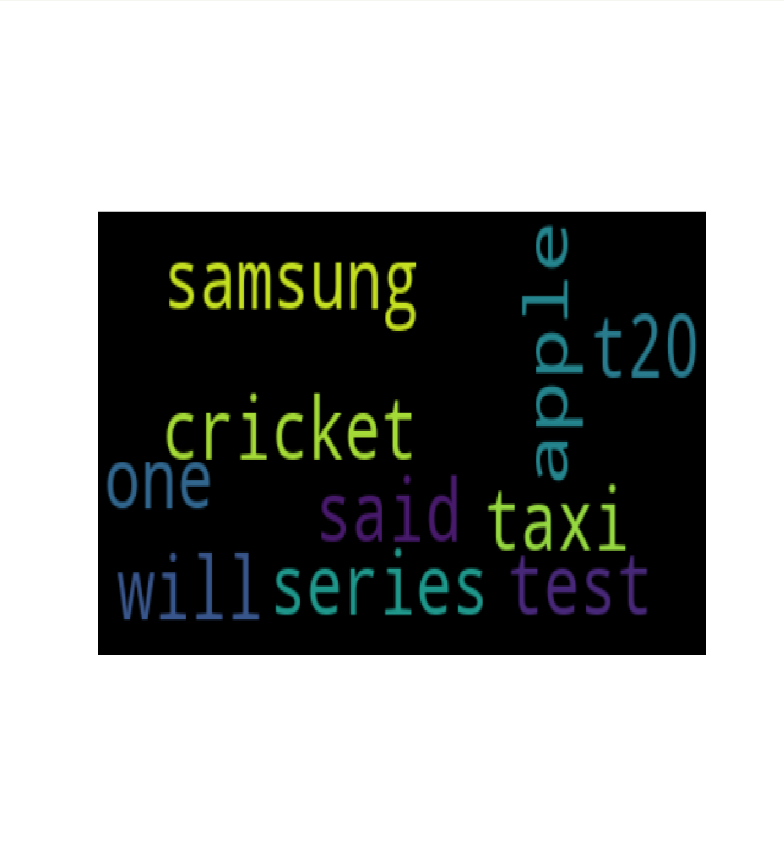
This model generates topics based on word frequency from a set of documents.

Gives us a representation that each document is a mixture of topics.

Topics


rupee
power
dollar
mobile
people
battery
facebook
us
finance
minister


market
approval
pakistan
auto
sez
team
investment
finance
banks
weeks


samsung
apple
t20
cricket
one
said
taxi
will
series
test

Representation of words over topics

Document 1: [(0, 0.90895), (1, 0.04612), (2, 0.04492)]

Document 2:[(0, 0.06031), (1, 0.88831), (2, 0.05133)]

Document 3:[(0, 0.05683), (1, 0.89173), (2,0.051436)]

Document 4:[(0, 0.88827), (1, 0.05674), (2, 0.05497)]

Document 5:[(0, 0.90047), (1, 0.05130), (2, 0.04822)]

Document 6:[(0, 0.89641), (1, 0.05394), (2, 0.04963)]

Document 7:[(0, 0.88054), (1, 0.06082), (2, 0.058622)]

Document 8:[(0, 0.06455), (1, 0.878264), (2, 0.057182)]

Conclusion

The Topic Modeling technique Latent Dirichlet Allocation (LDA) has been applied on a news group dataset. 3 topics are presented form 150 instances covering all the documents. By seeing the topic-document distribution, we can tell that which document contains which topic in highest percentage.