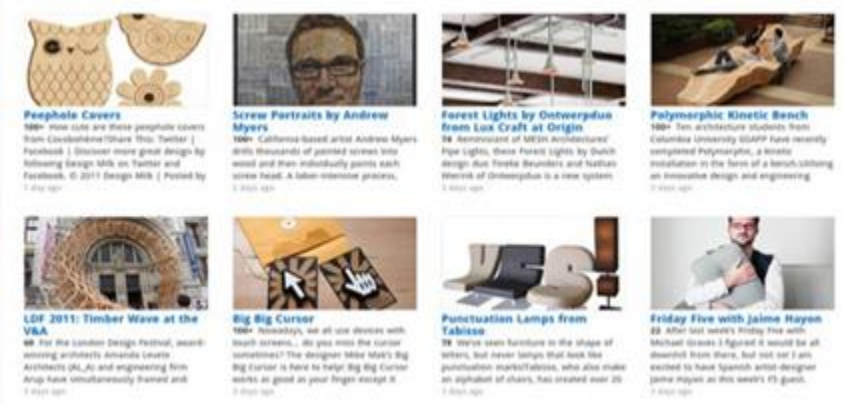
TOPIC MODELING

Siddhesh Tiwari

University of Illinois at Chicago

Motivation



Plethora of documents being published by news agencies

Organizing articles is important and challenging

Solution:
TOPIC MODELING

Data

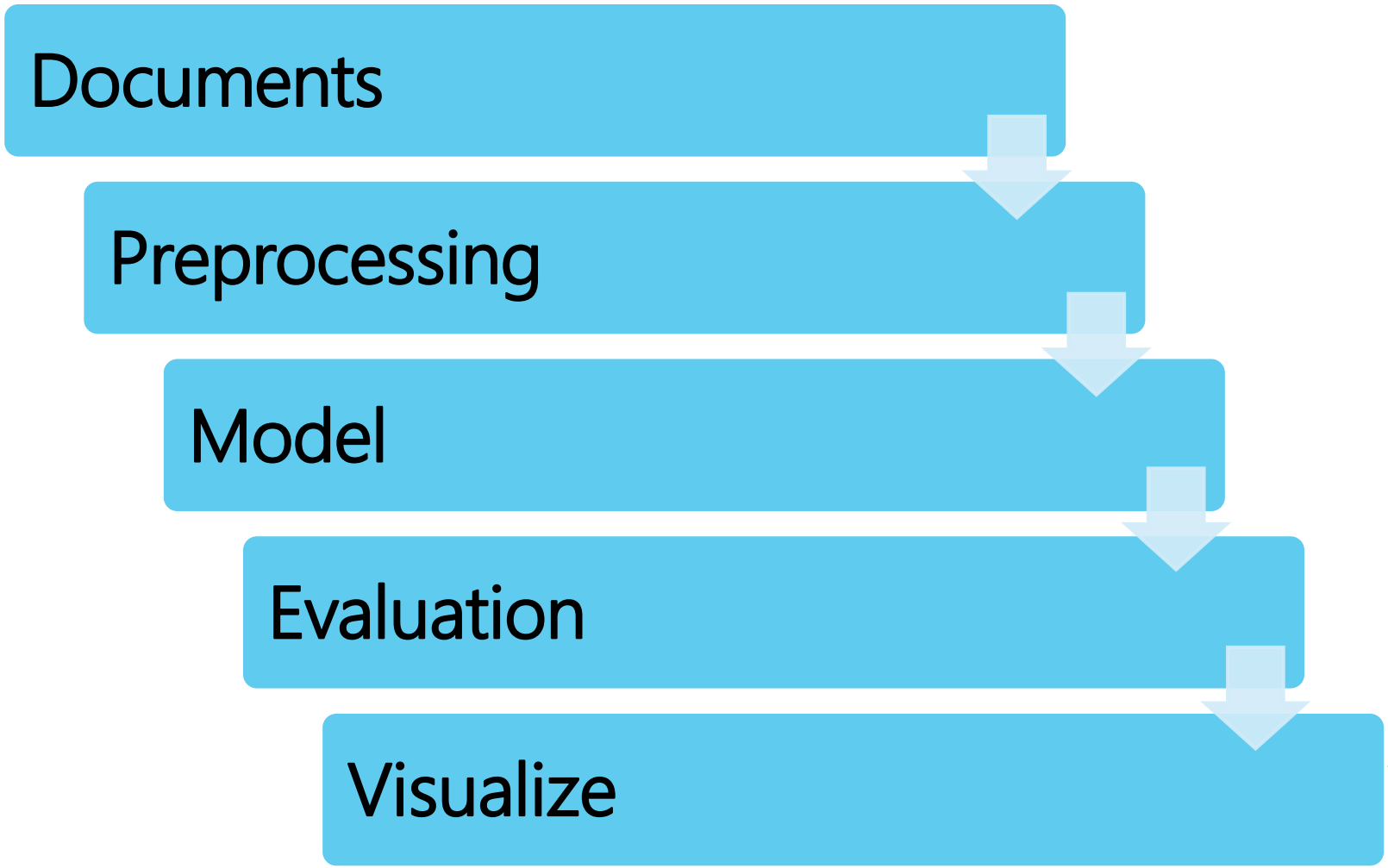
Corpus of 2246 articles from Associated Press

Articles in corpus comes from variety of domain

Tools & Algorithm

- Python
- Gensim
  - Python library for Semantic Analysis and Topic Modelling
  - Contains implementation of LDA, LSA and other unsupervised learning algorithm for language processing
- pyLDAvis
  - Python library to visualize topics generated from LDA
- Latent Dirichlet Allocation

Implementation Flow



Preprocessing

Divided single text file into 2246 text documents

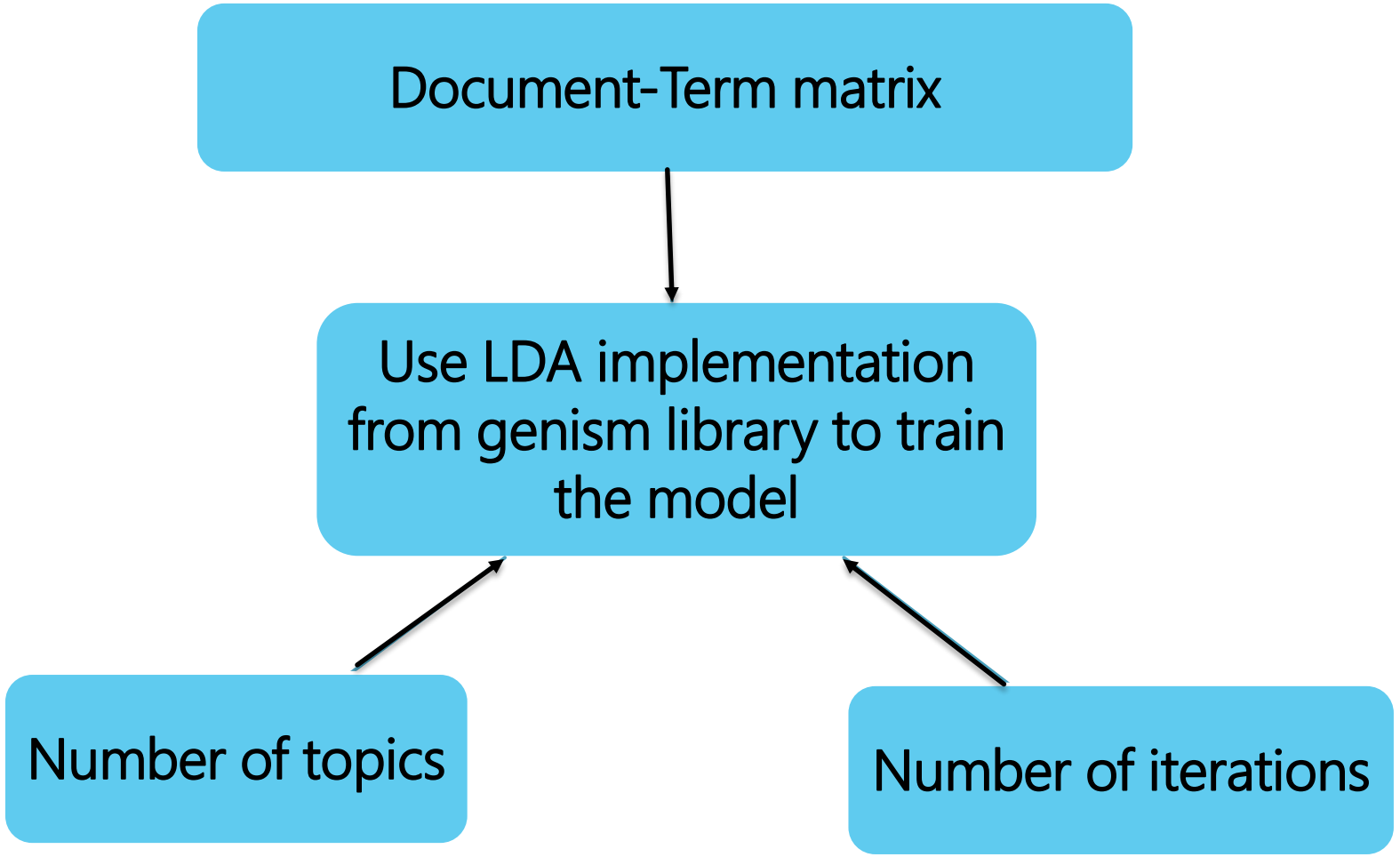
Remove Non-ASCII values

Tokenize

Remove Stopwords

Remove word with frequency less than 2 and length less than 5

Model

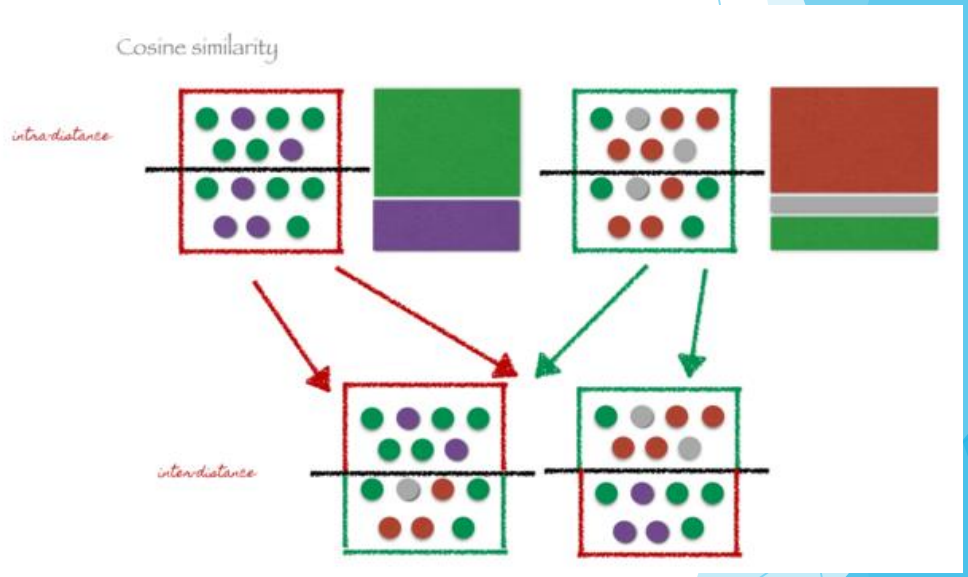

Document-Term matrix
Use LDA implementation from genism library to train the model
Number of topics
Number of iterations

Evaluation

Split each testing documents into two parts

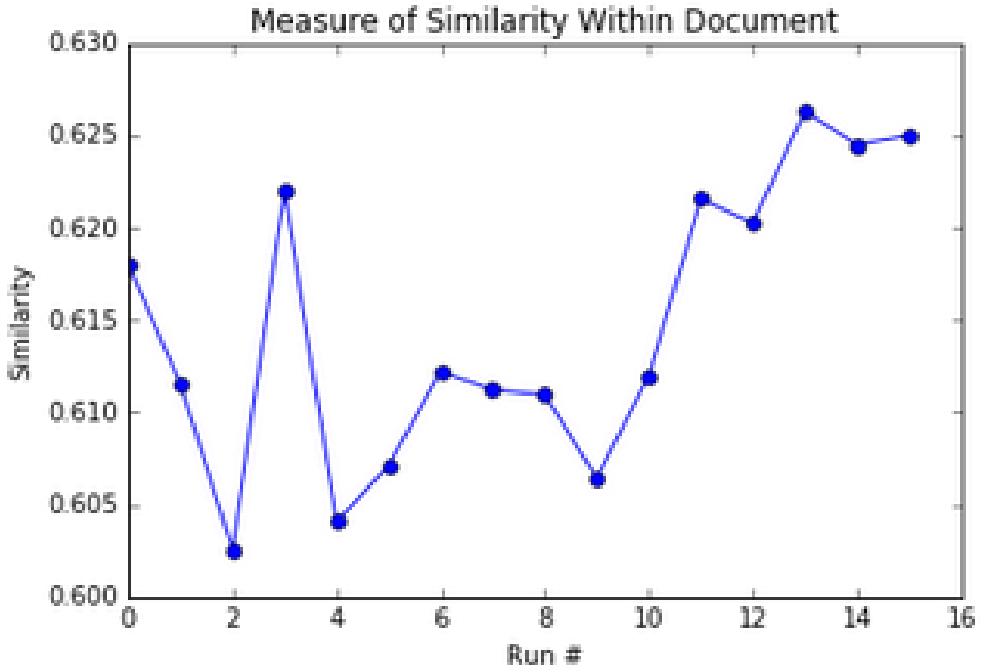
Compute average cosine similarity between corresponding halves with respect to topics. This Metric should be higher.

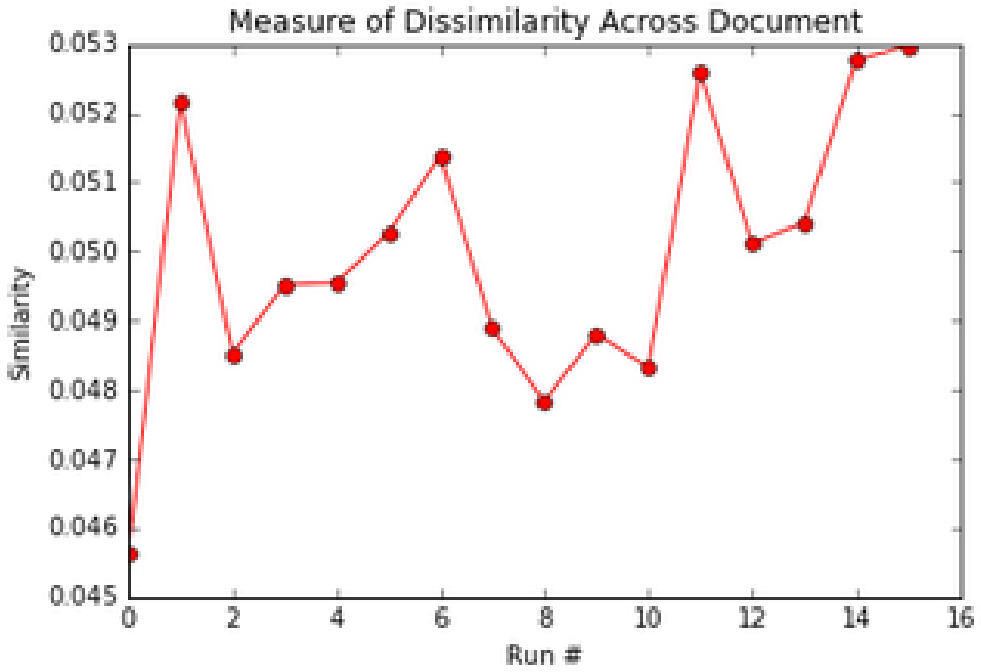
Compute average cosine similarity between halves of different documents. This metric should be small.



http://chdoig.github.io/pygotham-topic-modeling/#/3/15

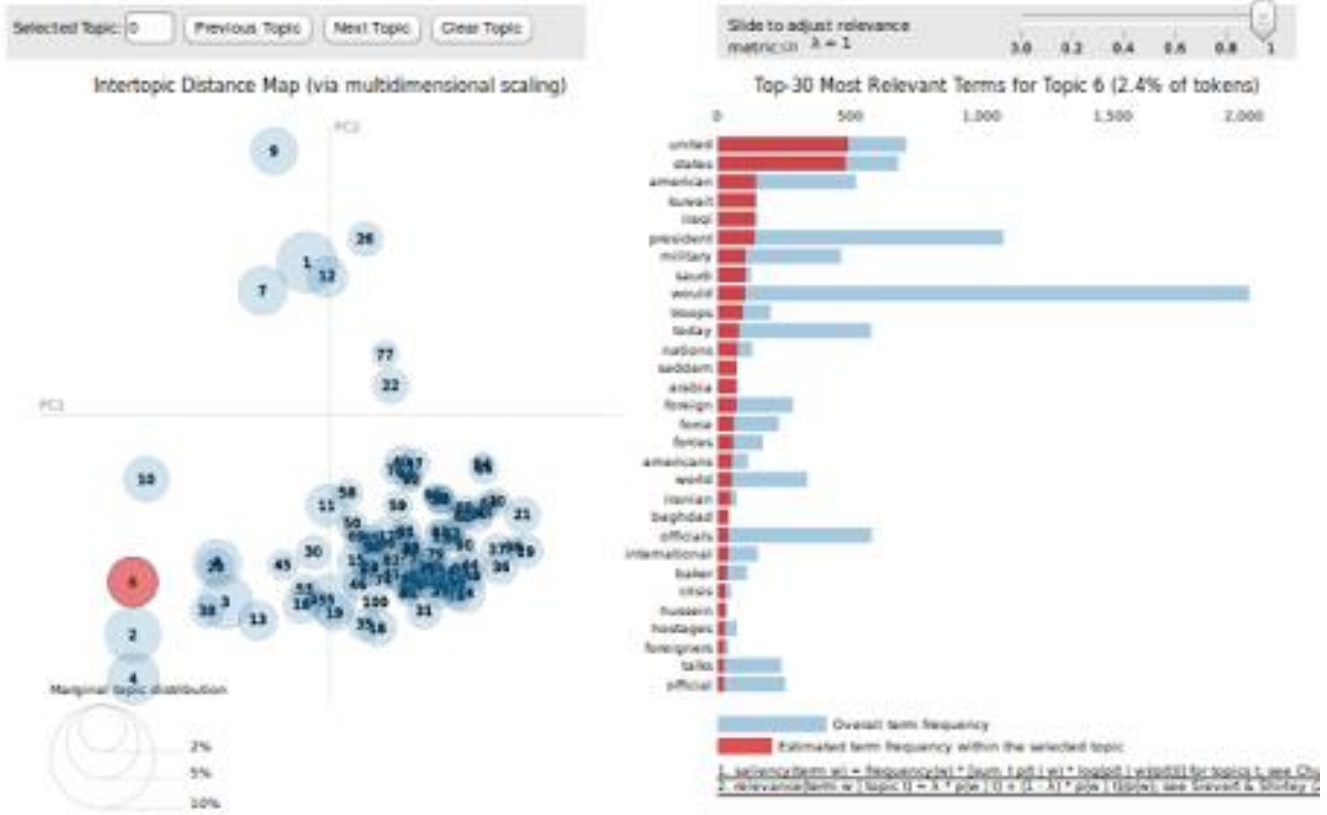
Evaluation


Measure of Similarity Within Document
Similarity
Run #
0.630
0.625
0.620
0.615
0.610
0.605
0.600
0
2
4
6
8
10
12
14
16


Measure of Dissimilarity Across Document
Similarity
Run #
0.053
0.052
0.051
0.050
0.049
0.048
0.047
0.046
0.045
0
2
4
6
8
10
12
14
16

Visualization

Used pyLDAvis library for generating interactive visualization for topics



Applications

- Topic similarity among documents can be used to organize documents in news portals, online library and research paper publishers
- Topic modeling can be used to build recommendation engine for online magazines by using topic similarity between documents and preferences of users.


Thanks for listening!
Any Questions?
No?
SUPER!