Using **Topic Modelling** to determine the relationships between **#Hashtags** and **Tweets.**

Olashile Adebimpe

Quick Overview.

Hashtags

Series of Tweets

Correlation using NLP Topic Modelling Algorithm

Overview

1. Introduction and Related Works
2. Data Sampling & Preprocessing
3. Topic Modelling Approach
4. Set of Experiment
5. Results and Future Works

Overview

1 **Introduction and Related Works**

2 Data Sampling & Preprocessing

3 Topic Modelling Approach

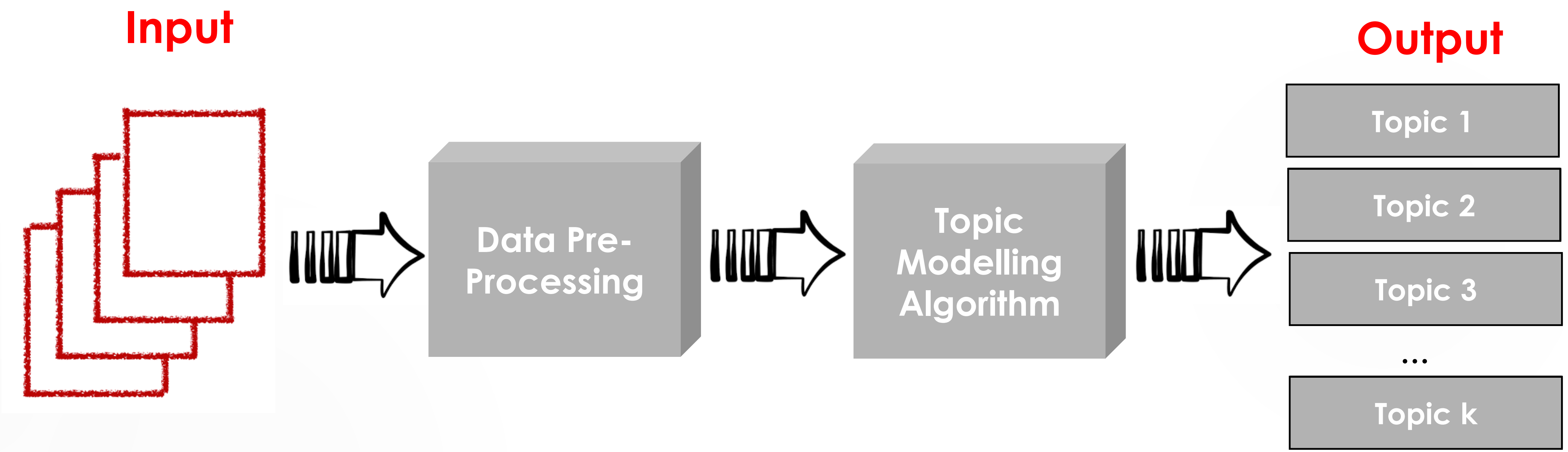
4 Set of Experiment

5 Results and Future Works

What is Topic Modelling?

An unsupervised process used for identifying topics present in text objects and deriving hidden patterns exhibited by a large cluster of text.

What is Topic Modelling?



Input

A corpus of unstructured text document

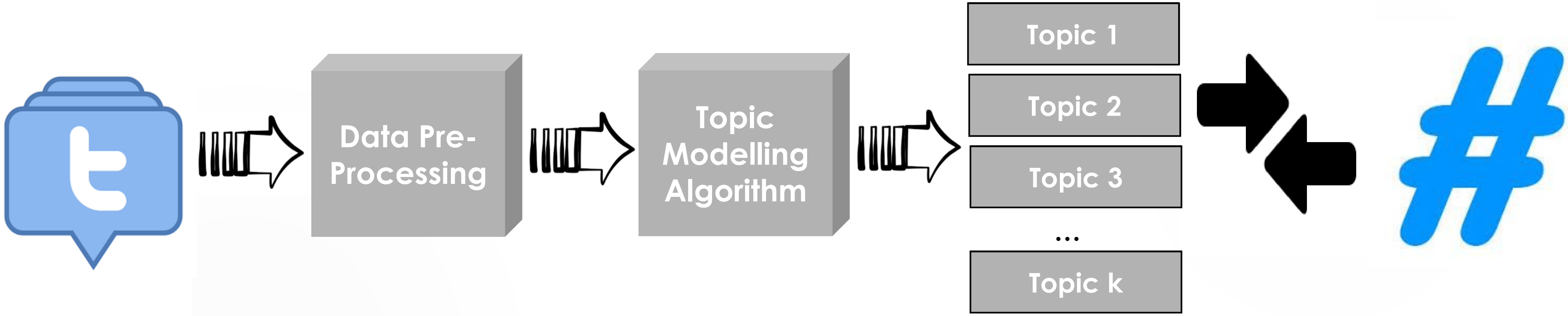
Output

A set of topics represented by top rank terms for the topic

Related Works.

Latent Semantics Analysis (LSA)	Latent Dirichlet Allocation (LDA)
Dimensionality reduction method which uses term matrix with the help of singular value decomposition for topic Modelling[2].	Assumes that documents are represented as a mixture of latent topics, where each topic are characterized by a distribution over words. [1]
Dirichlet-Multinomial Regression (DMR)	**Twitter-LDA model**
DMR an extension of LDA that allows conditioning on arbitrary document feature by including long-linear prior on document-topic distribution. [4]	Discover topics by allowing comparison to traditional news media. [3]

How it applies here?



Extracting topic from tweets and determining the relationship between the modeled topics and related hashtag.

Overview

1 Introduction and Related Works

2 **Data Sampling & Preprocessing**

3 Topic Modelling Approach

4 Set of Experiment

5 Results and Future Works

Data Sampling.

Tweepy

A python library for accessing Twitter for the collection of tweets. [5]

Data Collected

We sampled approximately 330k unique tweets related to 40 hashtags

Tweepy

An easy-to-use Python library for accessing the Twitter API.

Data Sampling.

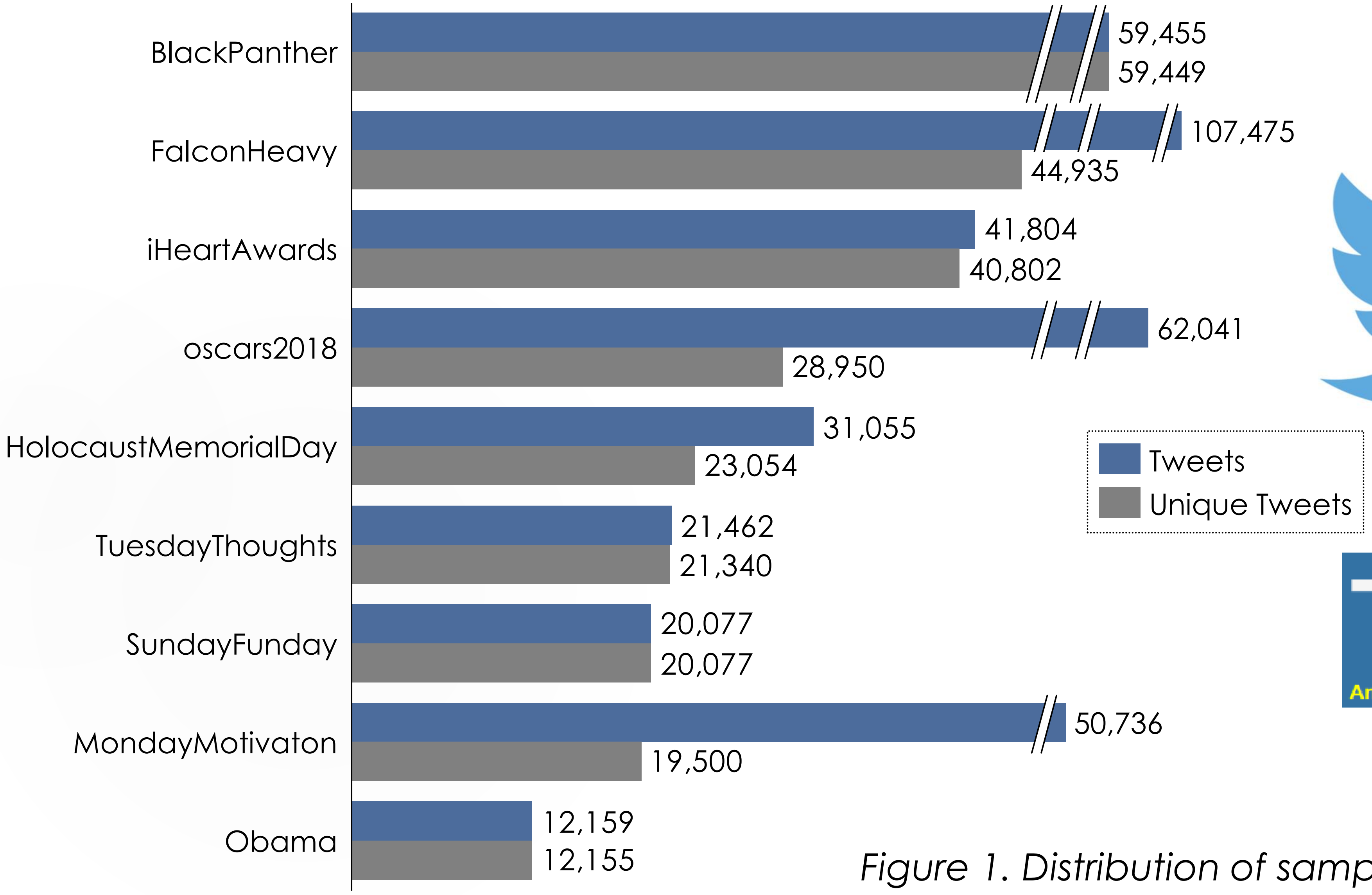
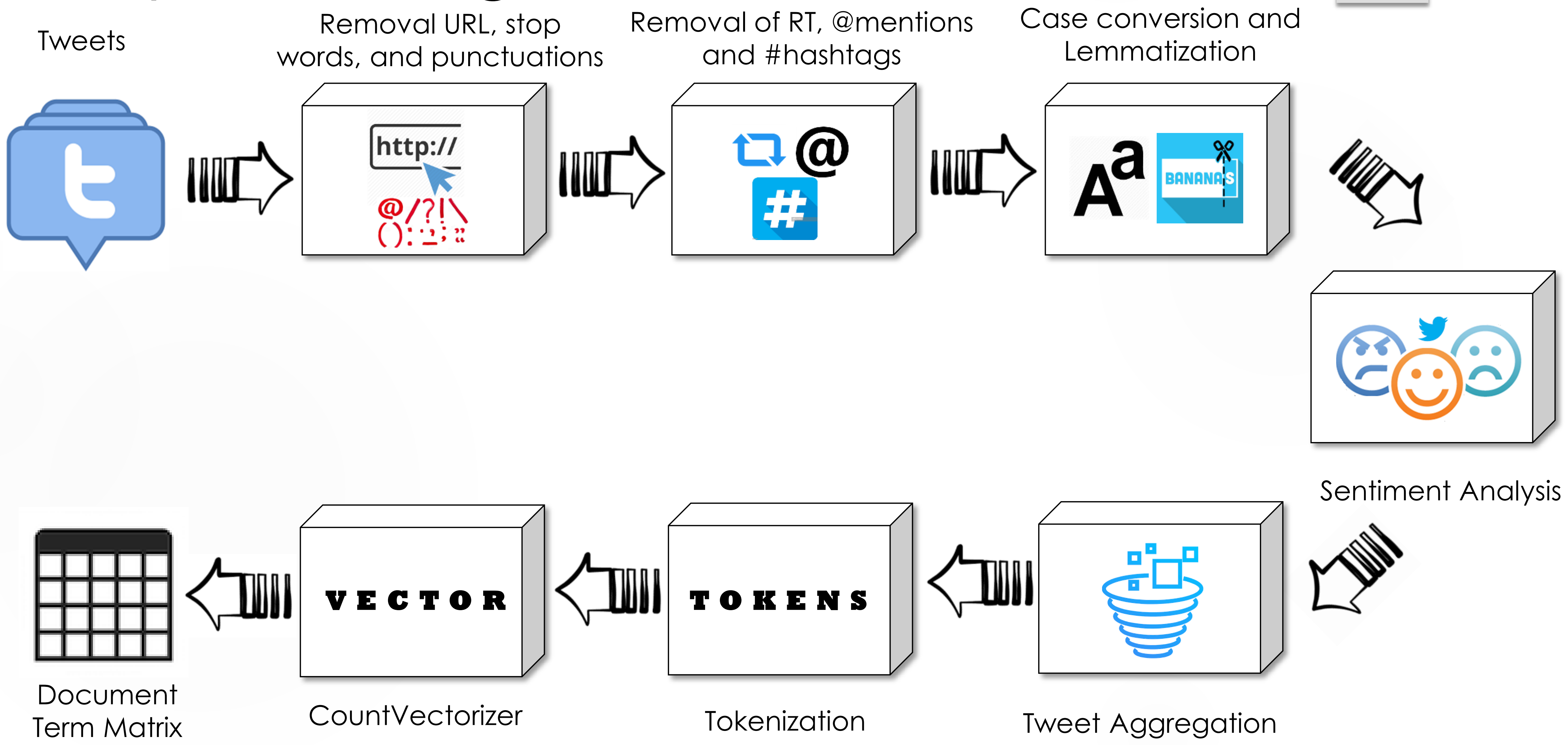


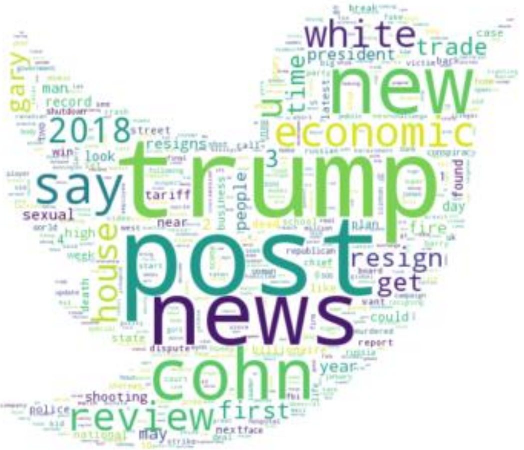
Figure 1. Distribution of sampled Tweets

Preprocessing



Preprocessing

#HolocaustMemorialDay



#BreakingNews

#MondayMotivaton

#FalconHeavy

#TuesdayThoughts

#iHeartAwards

#BlackPanther

#FridayReads

Overview

1 Introduction and Related Works

2 Data Sampling & Preprocessing

3 **Topic Modelling Approach**

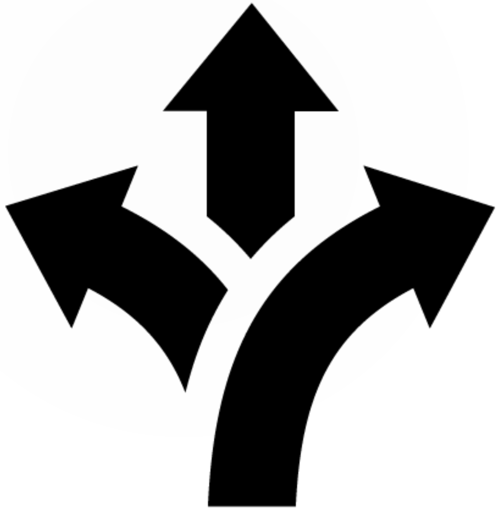
4 Set of Experiment

5 Results and Future Works

Topic Modelling Approach

- 2 popular Topic Modelling Techniques.

1. Probabilistic Approach:
   - E.g. Latent Dirichlet Allocation(LDA)[1]
2. Matrix Factorisation Approach:
   - E.g Non-negative Matrix Factorisation (NMF)[6]



But these approaches are **Unsupervised**

Unsupervised Approach

- They implicitly use document level co-occurrence information to group semantically related words into a single topic. [7]

Our Topic Modelling Approach

We employed a simple and effective method of adding **priors** to the model to guide the direction of our model.

GuidedLDA Topic modeling

- Created by Allen Riddell and Tim Hopper
- Build on Python LDA Algorithm using collapsed Gibbs sampling.
- Uses prior to improve topic-word distribution and document –topic distribution[8]

GuidedLDA - Seed Extraction

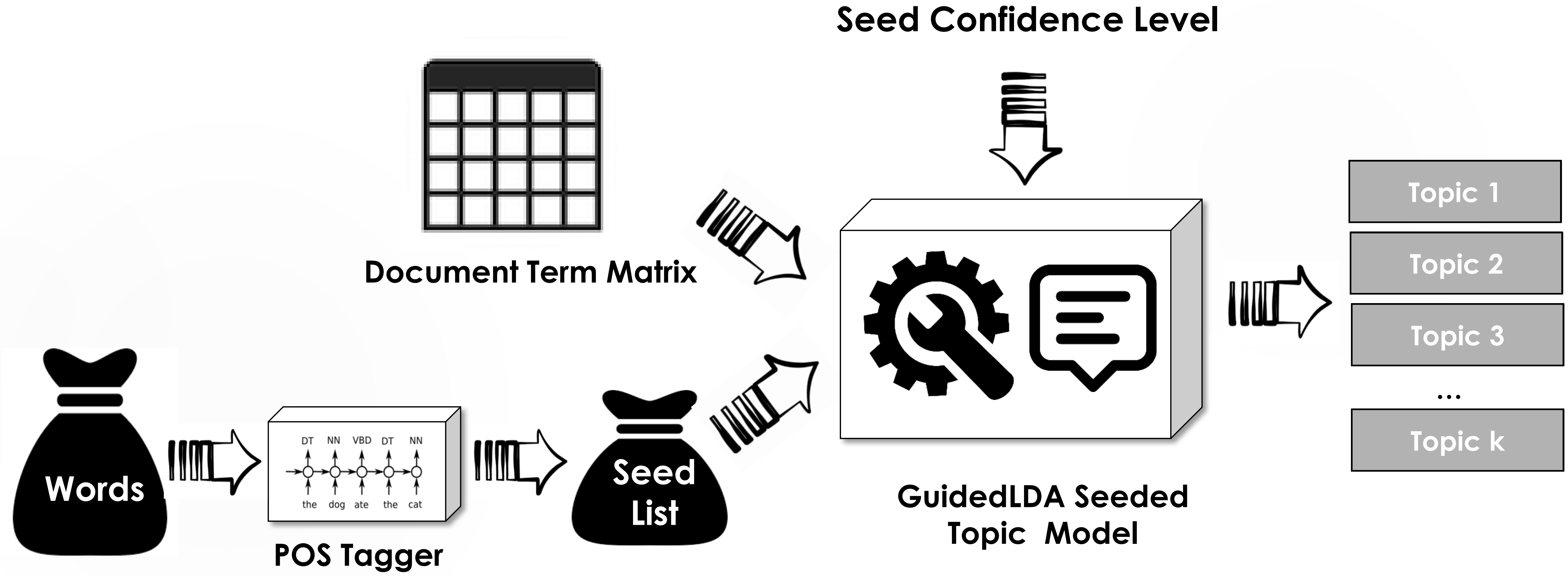
How do we generate the seeds considering the nature of our dataset?

GuidedLDA - Seed Extraction

- Few Assumptions

  Most tokens in topics or title of any document always belong to the **Noun** part of speech.

GuidedLDA Topic Modeling Approach



Overview

1 Introduction and Related Works

2 Data Sampling & Preprocessing

3 Topic Modelling Approach

4 **Set of Experiment**

5 Results and Future Works

Experiment- Parameter Selection

The key parameter selection decision for topic modelling involves choosing the number of topics k.

K = 10

Experiment- Dataset

- Selected hashtag from 4 different category

Hashtags			
Individuals	**Events**	**Days**	**Random**
Obama	SuperBowl	TuesdayThoughts	PressforProgress
Trump	iHeartAwards	MondayMotivation	BreakingNews
	FalconHeavy	ThursdayThoughts	
	HolocaustMemorialDay	FridayReads	

Experiment- Evaluation Techniques

- **Topic Coherence**: The extent to which the top terms representing our modeled topic are semantically related, relative to some "*background corpus*".

- *Measuring the similarity using **NLTK wordnet.synsets** between the modeled topics and the hashtags*

Overview

1 Introduction and Related Works

2 Data Sampling & Preprocessing

3 Topic Modelling Approach

4 Set of Experiment

5 **Results and Future Works**

Results- Topics

#Trump

LDA

GLDA

#HolocaustMemorialDay

LDA

GLDA

#ThursdayThoughts

LDA

GLDA

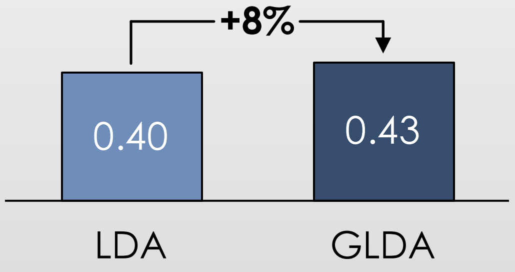
#PressforProgress

LDA

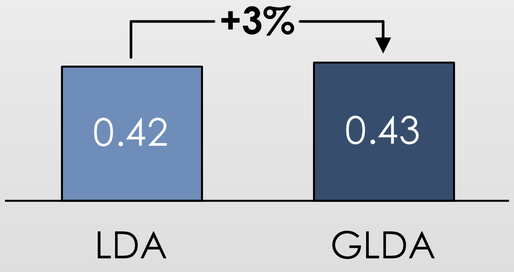
GLDA

Evaluation Metrics - Topic Coherence

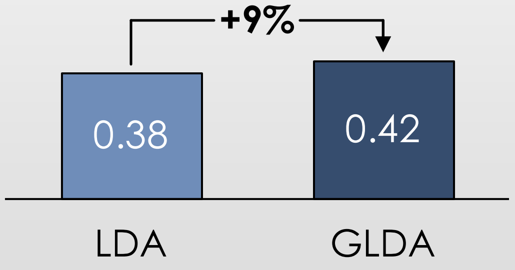
#Trump



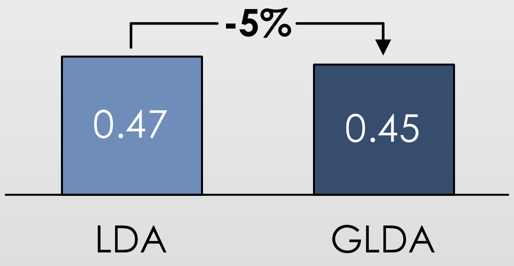
#HolocaustMemorialDay



#ThursdayThoughts

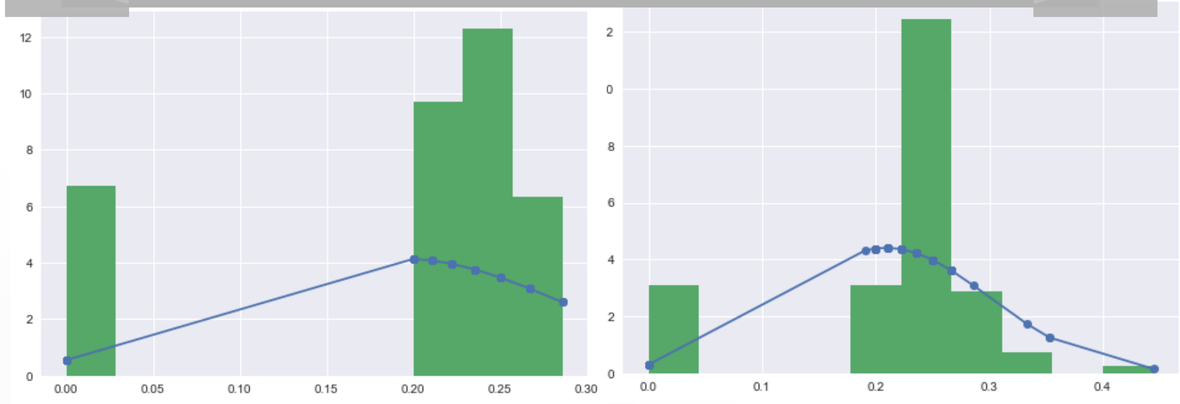


#PressforProgress

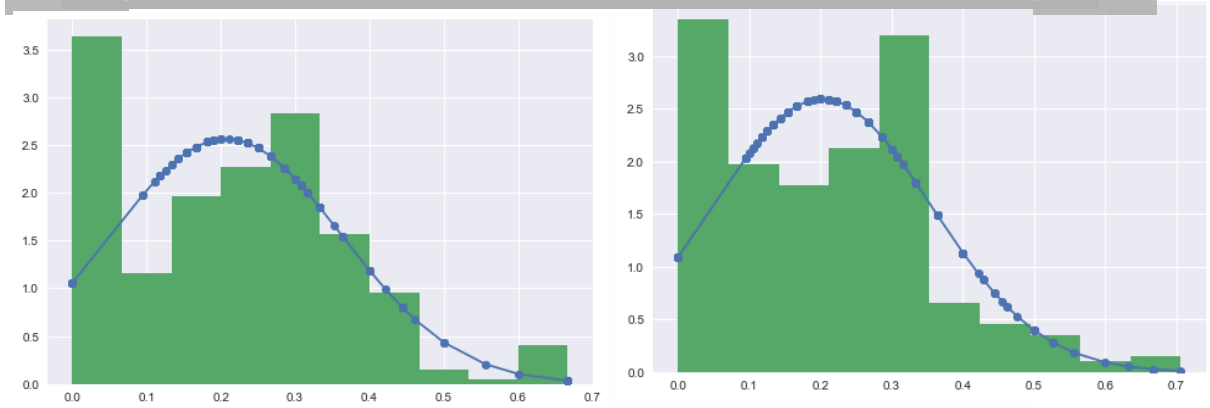


Evaluation Metrics – Similarity Measure

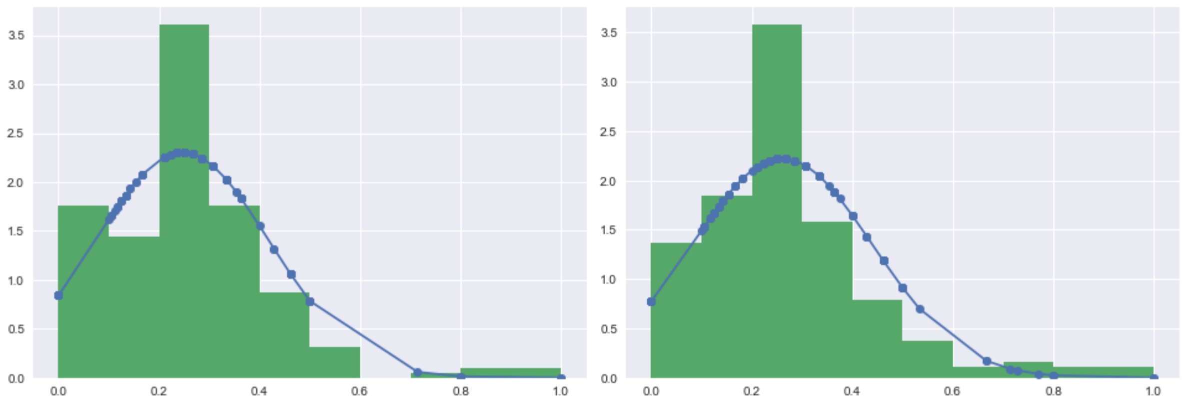
#Trump



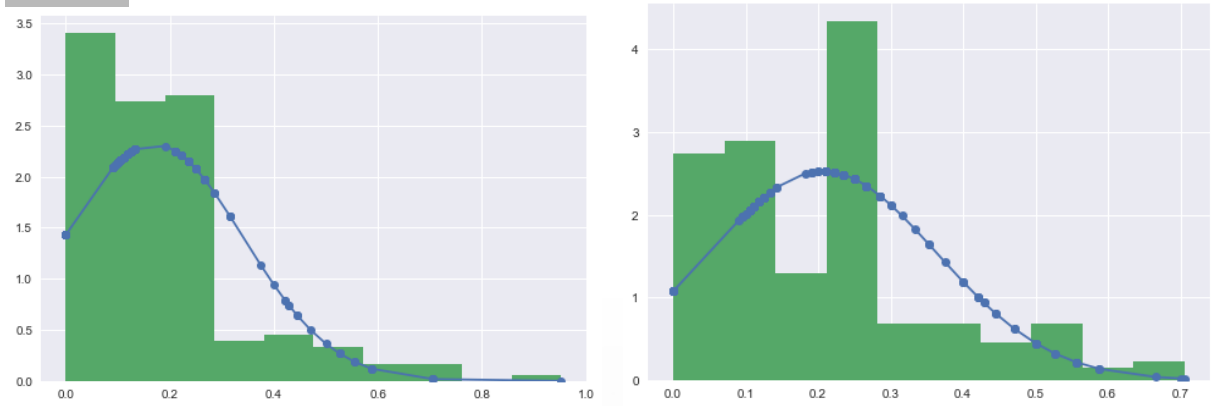
#HolocaustMemorialDay



#ThursdayThoughts

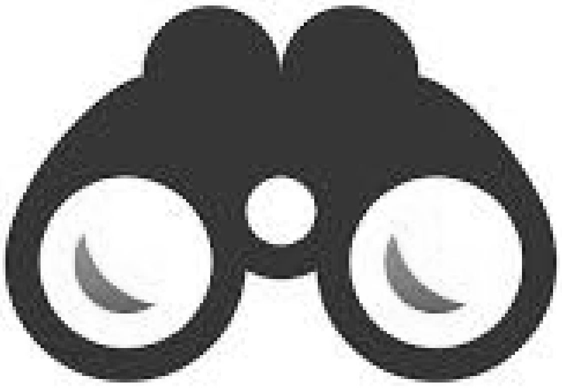


#PressforProgress



Future Works

- Improve on the parameter selection
- Improve the way our seed list was created.
- Try the proposed model on a more structured topic modeling problem
- Using standard information retrieval evaluation techniques
- Also, define a more optimal algorithm for evaluation using WordNet.




THANK YOU

Questions

Reference

[1] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, “Latent Dirichlet Allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[2] G. Maskeri, S. Sarkar, and K. Heafield, “Mining business topics in source code using latent dirichlet allocation,” Proc. 1st Conf. India Softw. Eng. Conf. - ISEC ’08, p. 113, 2008.

[3] W. X. Zhao *et al.*, “Comparing Twitter and Traditional Media using Topic Models,” *Proc. 33rd Eur. Conf. Adv. Inf. Retr.*, pp. 338–349, 2011.

[4] M. Song and M. C. Kim, “RT2M : Real-time twitter trend mining system,” in *Proceedings - 2013 International Conference on Social Intelligence and Technology, SOCIETY 2013*, 2013, pp. 64–71.

[5] J. Roesslein, “Tweepy.” [Online]. Available: http://www.tweepy.org/. [Accessed: 15-Mar-2018].

[6] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[7] J. Chang, S. Gerrish, C. Wang, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” Adv. Neural Inf. Process. Syst. 22, pp. 288--296, 2009.

[8] J. Jagarlamudi and H. Daum, “Incorporating Lexical Priors into Topic Models,” Umiacsumdedu, pp. 204–213, 2009.

[9] “2,300,000+ free and premium vector icons. SVG, PNG, AI, CSH and PNG format.” [Online]. Available: https://www.iconfinder.com/. [Accessed: 08-Apr-2018].