Topic Modeling

Vengal Rao Pachava
2156575
vpachava@cougarnet.uh.edu

Samarasimha Reddy
2252338
spunnam@cougarnet.uh.edu

Corwin Bennett
2139400
cbennet7@cougarnet.uh.edu

Abstract—**This study is dedicated to the preprocessing and exploratory analysis of a corpus containing newspaper articles. The initial phase involves constructing the corpus by loading and segmenting the data into individual articles. Subsequently, the corpus undergoes cleanup procedures, separating metadata from the actual article content, and features are extracted. Summaries and visual plots are generated to facilitate exploration. A topic model is established using the refined corpus, running multiple times with varied parameters, and the outcomes are documented in output files. The findings are then discussed in the context of the project's overarching statement, utilizing pertinent model outputs and visualizations to support the conclusions. It's essential to emphasize that the project does not require exhaustive domain knowledge but rather encourages an approach akin to supporting a journalist. The evaluation of post-processed data quality constitutes a crucial aspect of the grading for this assignment.**

Key words: Topic Modeling, Corpus data, Word clouds, pyLDAvis, NLTK tools.

I. INTRODUCTION

The exponential growth of digital content has led to an inundation of unstructured data across diverse formats, notably in the realm of news articles. Undertaking preprocessing and exploratory data analysis on such voluminous data holds immense promise for journalists and researchers seeking valuable insights and opportunities for knowledge discovery. This study focuses on the meticulous preprocessing and exploratory analysis of a corpus composed of newspaper articles, with the primary objective of constructing a topic model to identify and understand the various topics covered.

The study initiates by establishing the corpus, involving the loading of data as a substantial body of text, segmenting it into individual articles, and refining the corpus by distinguishing metadata from the actual article content. Feature extraction follows suit, with the creation of summaries and plots to facilitate a comprehensive exploration of the dataset. Preprocessing emerges as a pivotal phase, encompassing tasks such as tokenizing the text and addressing intricacies such as punctuation, headers, tags, and dates. Following this, a topic model is meticulously developed using the refined corpus, involving adjustments to parameters and multiple runs to capture meaningful summaries in output files.

The diverse results generated are scrutinized within the broader framework of the project's objectives, supported by relevant model outputs and visualizations. It is essential to note that a comprehensive understanding of the domain is not presupposed, and the analysis is designed to align with the support provided to a journalist. This study underscores the critical importance of preprocessing and exploratory data analysis in extracting meaningful insights from unstructured data. Moreover, it showcases the potential of topic modeling as a powerful tool for discerning and comprehending the various topics covered in newspaper articles, contributing to a deeper understanding of the media landscape.

II. METHODOLOGY

- **Create a corpus:** The dataset is sourced from Factiva, a comprehensive Global News database, and pertains to the year 2017. It comprises articles from distinguished sources such as the Wall Street Journal and the New York Times. The initial data, after being downloaded, was processed to obtain .txt files. The subsequent step involved segmenting the articles based on specified keywords like 'Document NYTF', 'Document INHT', 'Document WSJ', 'Document J000', and 'Document AWSJ'. The result is a collection of 1648 articles, each categorized accordingly.

- **Refine the corpus:** The next step involved scrutinizing the parsed articles to ensure proper formatting and accurate extraction. We conducted checks to verify the integrity of the data extraction process. Articles typically concluded with the phrase" All rights reserved." To streamline the corpus, we segregated articles that concluded with this phrase and excluded them from further analysis. Additionally, metadata was extracted during this cleanup process.

- **Preprocessing the data**: In the pre-processing phase, several sequential operations were executed. These included the elimination of punctuation, uniform conversion of words to lowercase, and the segmentation of text into individual tokens. Leveraging the NLTK stop words list, we systematically removed common stop words. Subsequently, stemming was applied using the Porter stemmer. These measures collectively contribute to the refinement and standardization of the textual data for subsequent analysis.

- **Feature extraction:** The process of information extraction involved the creation of a document-term matrix. This matrix serves as a structured representation of the relevant information. Additionally, a visual representation of the most prominent words was generated using a word cloud, providing a concise and visually intuitive summary of the key terms within the corpus.

- **Topic modeling (LDA):** To prepare the data for topic modeling using the genism library, a secondary round of pre-processing was undertaken. This involved the creation of bi-gram and tri-gram models, coupled with lemmatization. Subsequently, an LDA (Latent Dirichlet Allocation) model was constructed using genism. The corpus served as input for the model, and the number of topics was specified to facilitate the identification and exploration of key themes within the dataset.

- **Results Evaluation**: The pyLDAvis library was employed for visualizing topics and associated terms. Evaluation metrics such as perplexity and coherence scores were utilized to assess the model's performance. A lower perplexity score, ranging from -8.38 to -8.37, indicates effective prediction of new data. On the other hand, coherence scores, ranging from 0.39 to 0.46, reflect the semantic coherence between topics, with higher scores denoting improved coherence. The evaluation suggests that the model generally performs well, offering low error rates in predicting new data. However, there is room for enhancing semantic coherence between topics.

  Number of Topics:

  • The models with 5 topics consistently show the best coherence scores across different numbers of passes.

  • As the number of topics increases, the coherence scores generally decrease.

  Number of Passes:

  • Increasing the number of passes generally results in slightly lower perplexity scores but doesn't always lead to a significant improvement in coherence.

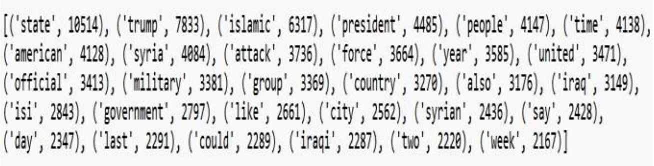
  Considering the coherence scores and the trade-off with perplexity, the model with 5 topics and 20 passes appears to be performing relatively well. It's essential to strike a balance between model complexity and performance, and the optimal choice may depend on the specific goals of your analysis and the interpretability of the topics.

III. EXPERIMENTAL RESULTS

After obtaining the preprocessed corpus data, we can now start analyzing the preprocessed data. Performed the frequency distribution for individual words with stop words removal using NLTK library, this process to find the best features and Wordcloud used to plot those features. Now we are going to perform Topic model techniques like LDA to analysis our data, based on different numbers of topic selections and passes. Below are the visualizations for our results.

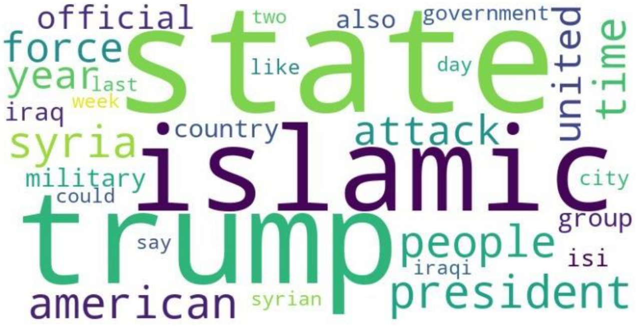
3.1 Common features

Top 30 Features extracted using frequency distribution:

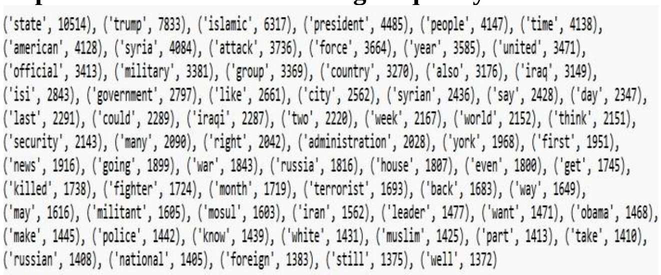


```
[('state', 10514), ('trump', 7833), ('islamic', 6317), ('president', 4485), ('people', 4147), ('time', 4138),
('american', 4128), ('syria', 4084), ('attack', 3736), ('force', 3664), ('year', 3585), ('united', 3471),
('official', 3413), ('military', 3381), ('group', 3369), ('country', 3270), ('also', 3176), ('iraq', 3149),
('isi', 2843), ('government', 2797), ('like', 2661), ('city', 2562), ('syrian', 2436), ('say', 2428),
('day', 2347), ('last', 2291), ('could', 2289), ('iraqi', 2287), ('two', 2220), ('week', 2167)]
```

Word cloud generated using the extracted 30 features:



Top 70 Features extracted using frequency distribution:



```
('state', 10514), ('trump', 7833), ('islamic', 6317), ('president', 4485), ('people', 4147), ('time', 4138),
('american', 4128), ('syria', 4084), ('attack', 3736), ('force', 3664), ('year', 3585), ('united', 3471),
('official', 3413), ('military', 3381), ('group', 3369), ('country', 3270), ('also', 3176), ('iraq', 3149),
('isi', 2843), ('government', 2797), ('like', 2661), ('city', 2562), ('syrian', 2436), ('say', 2428), ('day', 2347),
('last', 2291), ('could', 2289), ('iraqi', 2287), ('two', 2220), ('week', 2167), ('world', 2152), ('think', 2151),
('security', 2143), ('many', 2090), ('right', 2042), ('administration', 2028), ('york', 1968), ('first', 1951),
('news', 1916), ('going', 1899), ('war', 1843), ('russia', 1816), ('house', 1807), ('even', 1800), ('get', 1745),
('killed', 1738), ('fighter', 1724), ('month', 1719), ('terrorist', 1693), ('back', 1683), ('way', 1649),
('may', 1616), ('militant', 1605), ('mosul', 1603), ('iran', 1562), ('leader', 1477), ('want', 1471), ('obama', 1468),
('make', 1445), ('police', 1442), ('know', 1439), ('white', 1431), ('muslim', 1425), ('part', 1413), ('take', 1410),
('russian', 1408), ('national', 1405), ('foreign', 1383), ('still', 1375), ('well', 1372)
```

Word cloud generated using the extracted 70 features:

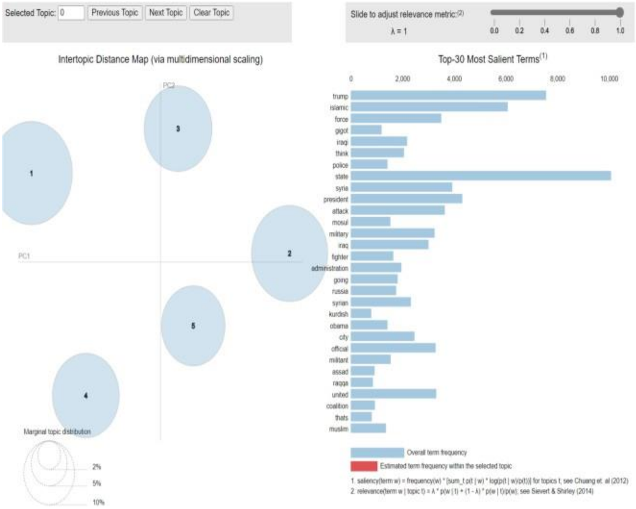
russia
obama
say
fighter
syrian
islamic
time
people
terrorist
iraq
muslim
part
state
last
isi
killed
city
get
make
militant
like
united
first
world
right
news
going
many
take
two
month
house
could
year
also
american
official
leader
even
president
military
security
trump
syria
war
want
way
well
week
still
may
police
york
iraqi
country
back
iran
group
russian
national
foreign
administration
mosul
white
think
know
force
attack
government

The above results show the most frequently occurring features for top30 and top70 in the given data with removing stop words. it Clearly shows that removing stop words has significantly reduced ambiguity. A quick look at the topmost 30 common frequency words likes indicates that by seeing some words like "trump", "State", "President", "Islamic", "United", gives an assumption like this article might be speaking about politics in United State that related to Trump and Islamic terrorist related topic. A quick look at the topmost 70 common frequency words indicates that by seeing some words like "people", "attack", "year", "military", "terrorist", givens as assumption like there might be a terrorist attack on people in New York and Military will be coming to save all this people. The inclusion of stop words significantly influenced the generation of the data.

3.2 LDA (Latent Dirichlet Allocation):

Here we are selecting topics 5, 10 and passes 10, 20, and performing combination of each topic with all passes it will give 4 different number of models and results. We can see how topics are distributed in each visualization. In each visualization we choose one topic, it gives the total token percentage of each topic, with having highest number of tokens for topic 1 in each LDA and we have done this for top 30 terms in the corpus. The lowest is topic 5. The top 15 most important words and their respective weights have been taken for analysis. These weights tell us how important each word is in its corresponding topic.

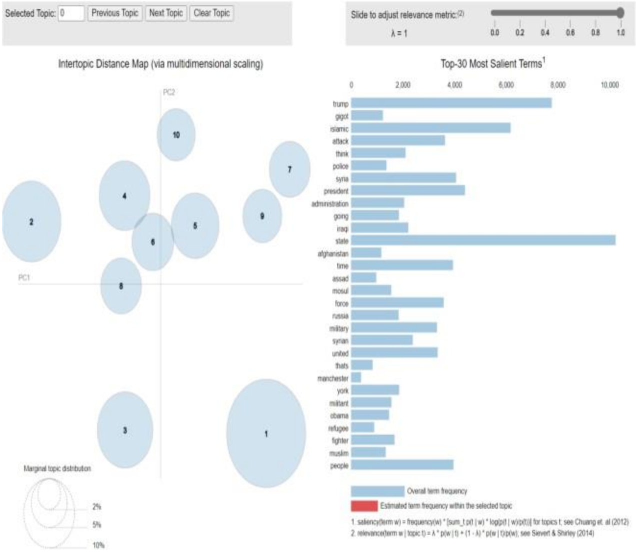
Topic 5 with Passes 10:



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 0 is related to year which Trump elected as President by people by seeing word "Trump", "President", "Year", "People". Topic 1 is relating to something like Muslim Islamic and about terrorist in that year by seeing words "terrorist", "muslim", "year", "islamic". Topic 2 is related to group of Islamic people attacked by police words like "Islamic", "police", "group". Topic 3 is related to Trump speaking about Obama in administration words like "Trump", "Obama", "administration". Topic 4 is related to Islamic force that came from Iran to New York words like "Islamic", "Iran", "force". Overall, the topic model provides insight into the variety of topics that could potentially be found in a large corpus of text and the relative importance of certain words within each topic.

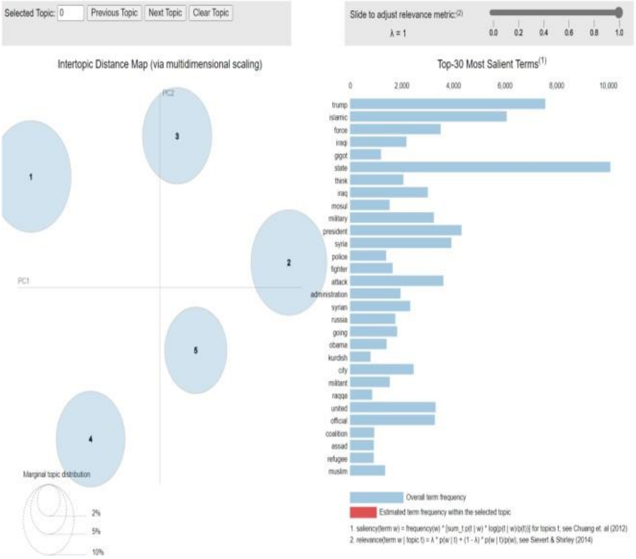
Topic 10 with Passes 10:



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 0 is related to topic that Trump mention about Afghanistan people by seeing words like "Trump", "Afghanistan", "state". Topic 1 relates to united state attack by muslim terrorist by seeing words like "United", "State", "Muslim". Topic 2 is related to attack happened between police and muslim people by seeing words like "police", "muslim", "attack". Topic 3 is related to having a conversation between Russia and United states by Trump by seeing words like "Trump", "Russia", "administration". Topic 4 is related to the attack that happened in Manchester state be seeing words like "state", "Manchester", "attack". Topic 5 is related to where refugee trump as president at time by seeing words like "Trump", "President", "refugee". Topic 6 is related to Trumps things people going to gigot by seeing words like "gigot", "trump", "people. Topic 7 relates to a company that help trump government by seeing words like "trump", "company", "government". Topic 8 is related to women are get attack by people with gun by seeing words like "gun", "women", "people". Topic 9 is related to islamic people got attacked by military by seeing words like "military", "islamic", "Iran". Overall, the topic model provides insight into the variety of topics that could potentially be found in a large corpus of text and the relative importance of certain words within each topic.

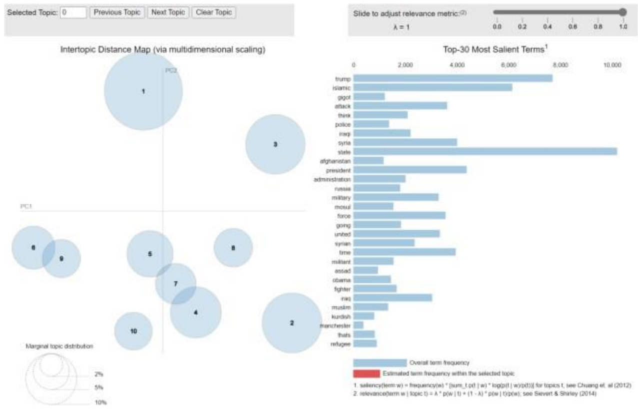
Topic 5 with Passes 20:



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 0 is related to year which Trump elected as President by people in New York state by seeing word "Trump", "President", "Year", "People", "state". Topic 1 is relating to something like Muslim Islamic and about terrorist in that year by seeing words "terrorist", "muslim", "year", "islamic". Topic 2 is related to group of Islamic people attacked by police words like "Islamic", "police", "group". Topic 3 is related to Trump speaking about Obama in administration words like "Trump", "Obama", "administration". Topic 4 is related to Islamic force that came from Iran to New York words like "Islamic", "Iran", "force". Overall, the topic model provides insight into the variety of topics that could potentially be found in a large corpus of text and the relative importance of certain words within each topic.

Topic 10 with Passes 20:

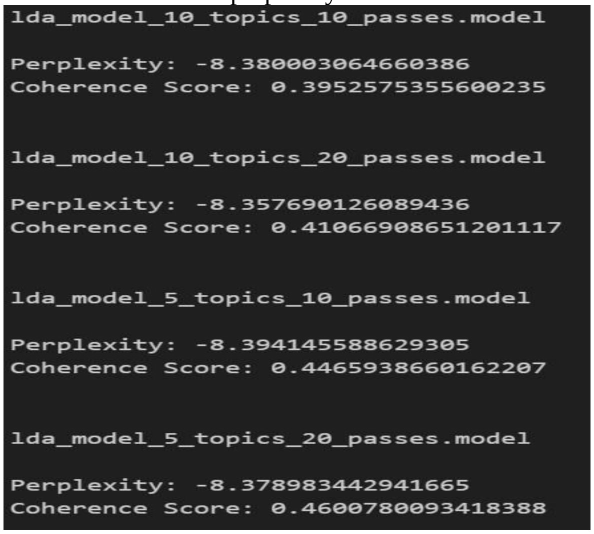


1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t, see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic 0 is related to topic that Trump mention about Afghanistan people by seeing words like “Trump”, “Afghanistan”, “state”, “President”. Topic 1 relates to united state attack by muslim terrorist by seeing words like “United”, “State”, “Muslim”. Topic 2 is related to attack happened between police and muslim people by seeing words like “police”, “muslim”, “attack”. Topic 3 is related to having a conversation between Russia and United states by Trump by seeing words like “Trump”, “Russia”, “administration”. Topic 4 is related to the attack that happened in Manchester state be seeing words like “state”, “Manchester”, “attack”. Topic 5 is related to where refugee trump as president at time by seeing words like “Trump”, “President”, “refugee”. Topic 6 is related to Trumps things people going to gigot by seeing words like “gigot”, “trump”, “people. Topic 7 relates to a company that help trump government by seeing words like “trump”, “company”, “government”. Topic 8 is related to women are get attack by people with gun by seeing words like “gun”, “women”, “people”. Topic 9 is related to islamic people got attacked by military by seeing words like “military”, “islamic”, “Iran”. Overall, the topic model provides insight into the variety of topics that could potentially be found in a large corpus of text and the relative importance of certain words within each topic.

Coherence Score and Perplexity:

Coherence score and perplexity for all the four models.



```
lda_model_10_topics_10_passes.model

Perplexity: -8.380003064660386
Coherence Score: 0.3952575355600235


lda_model_10_topics_20_passes.model

Perplexity: -8.357690126089436
Coherence Score: 0.41066908651201117


lda_model_5_topics_10_passes.model

Perplexity: -8.394145588629305
Coherence Score: 0.4465938660162207


lda_model_5_topics_20_passes.model

Perplexity: -8.378983442941665
Coherence Score: 0.4600780093418388
```

IV. Conclusion

In conclusion, this research underscores the crucial role of preprocessing and exploratory data analysis in the realm of newspaper articles, culminating in topic modeling. Through the construction and refinement of a corpus, along with the extraction of relevant features, we were able to generate summaries and visualizations that greatly aided in our understanding of the data. The application of topic modeling further allowed us to pinpoint significant themes and topics within the corpus, offering valuable insights for journalists to better grasp the data and enhance their reporting. This study highlights the essential nature of data analysis in journalism and stresses the importance of adequately preparing data for analysis. While the process of preprocessing and analyzing data can be demanding and time-consuming, the insights derived from the analysis can empower journalists to create well-informed and impactful stories. For future research endeavors, it may prove beneficial to eliminate common words from the analysis and explore additional analytical techniques, or to apply the same methodology to another dataset for comparative.