MINI PROJECT-3 REPORT

Topic Modeling

GROUP-8
Borigi, Akash, UH_ID: - 2205379

aborigi@cougarnet.uh.edu
(Pre-processing, Topic Modeling & Programming Analysis)

Pattamsetti, Sudheer, UH_ID: - 2254300

spattams@cougarnet.uh.edu
(Supervision, Validation, Writing Original Draft)

Rehan, Muhammad Asad, UH_ID: - 1951934

mrehan@cougarnet.uh.edu
(Review & Editing, Conceptualization, Presentation)

Link to Repository:

Click here for project documents

Abstract

This study delves into the preprocessing and exploratory data analysis of a corpus comprising newspaper articles. The initial phase involves constructing the corpus by loading and segmenting the data into individual articles. Subsequently, the corpus undergoes a refinement process where metadata is separated from the main article content, and pertinent features are extracted. To gain insights and a holistic understanding of the dataset, comprehensive summaries and plots are generated through exploratory data analysis. A key element of the study is the creation of a topic model using the refined corpus. This model is iteratively executed with varying parameters, and the outcomes are documented in output files. The ensuing discussion contextualizes the results within the project's overarching statement, leveraging model outputs and visualizations to reinforce the findings. It is noteworthy that this project does not assume complete domain expertise but rather approaches the analysis with a journalistic perspective. Furthermore, the evaluation of post-processed data quality constitutes an integral part of the assignment's grading criteria.

1.Introduction

The surge in digital content has resulted in a wealth of unstructured data, particularly in the form of news articles. The preprocessing and exploratory data analysis of such data offer valuable insights and opportunities for knowledge discovery, benefiting both journalists and researchers. This study is dedicated to the preprocessing and exploratory data analysis of a corpus consisting of newspaper articles, aiming to construct a topic model that facilitates the identification and comprehension of the various topics covered in these articles.

The study initiates by constructing the corpus, involving the loading of data as a substantial character, breaking it down into individual articles, and refining the corpus by isolating metadata from the core articles. Subsequently, features are extracted, and exploratory data analysis is conducted, with the creation of summaries and plots to facilitate a deeper understanding. The preprocessing phase is pivotal, encompassing tasks such as tokenizing the text and addressing nuances like punctuation marks, headers, tags, and dates.

Following the preprocessing steps, a topic model is generated using the refined corpus. This involves adjusting parameters and running the model iteratively, with the summaries being stored in output files. The study then delves into a discussion of the diverse results, aligning them with the project's overarching statement. To bolster our findings, relevant model outputs and visualizations are presented. It is crucial to emphasize that a complete domain knowledge is not presumed, and the analysis is approached in a manner supportive of journalistic endeavors. This study underscores the significance of preprocessing and exploratory data analysis in extracting meaningful insights from unstructured data, showcasing the potential of topic modeling to identify and comprehend the diverse topics covered in newspaper articles.

2. Methodology

1. Compile a corpus: The dataset originates from Factiva, a prominent Global News database, and was acquired in 2017. It specifically comprises articles sourced from the Wall Street Journal and the New York Times. The data, initially in a raw format, was transformed into .txt files for further analysis. Following the extraction of text files, articles were segmented based on predefined keywords, namely 'Document NYTF,' 'Document INHT,' 'Document WSJ,' 'Document J000,' and 'Document AWSJ.' In total, 1648 articles were successfully extracted, forming the basis for subsequent analysis and exploration.

2. cleaning the corpus: We validated the format and accuracy of the parsed articles to ensure proper extraction. Considering that a substantial number of articles conclude with the phrase "All rights reserved," we systematically separated these articles from others, discarding those that did not conform to this pattern. Additionally, metadata extraction was performed, resulting in a refined corpus consisting of 1620 articles that meet the specified criteria.

3.pre-processing the data: Several procedural steps were executed, encompassing the elimination of punctuation, conversion of each word to lowercase, tokenization of words, and the application of the NLTK stopwords list for the removal of stopwords. Subsequently, stemming was conducted utilizing the Porter stemmer to further refine the processed data.

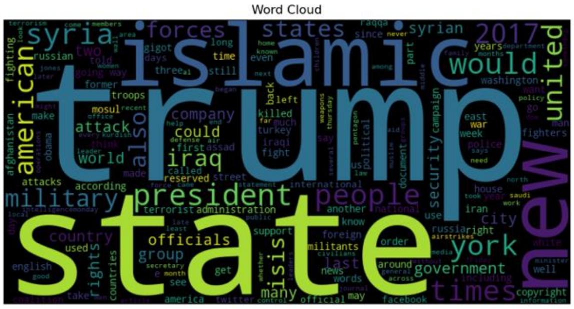
4.Feature Extract: To extract relevant information, we constructed a document-term matrix, and the most prominent words were visually represented through the utilization of a word cloud.

5. Topic Modeling (LDA): We reprocessed the data in preparation for topic modeling using gensim. This involved the creation of bi-gram and tri-gram models, coupled with the implementation of lemmatization. Subsequently, an LDA (Latent Dirichlet Allocation) model was constructed using gensim, where the corpus served as input, and the number of topics was specified.

6. Assessment of the outcomes: The visualization of topics and their corresponding terms was carried out using the pyLDAvis library. To evaluate the model's effectiveness, metrics such as perplexity and coherence scores were employed. A lower perplexity score, within the range of -8.41 to -8.60, signifies accurate predictions for new data. Meanwhile, coherence scores, ranging from 0.32 to 0.30, indicate the semantic consistency among topics, with higher scores indicating better coherence. Overall, the evaluation indicates good model performance with low error rates in predicting new data. However, there is room for improvement in enhancing the semantic coherence between topics.

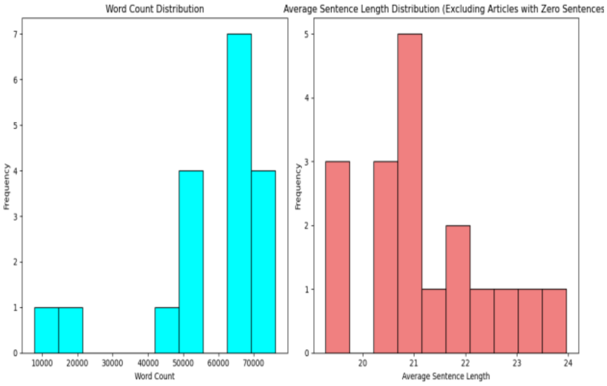
3. Experimental Results:

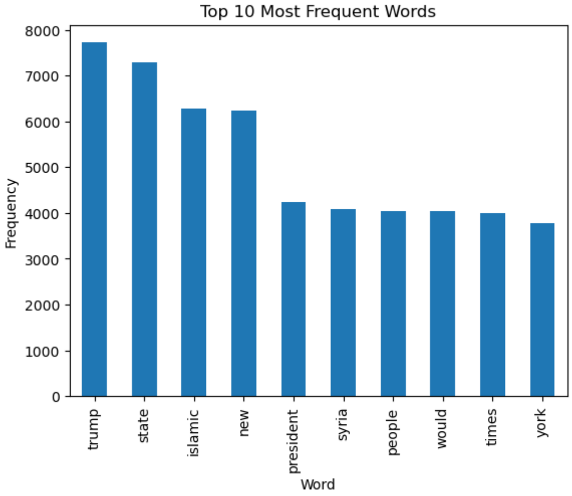
Generate a word cloud using word counts after the initial preprocessing steps.



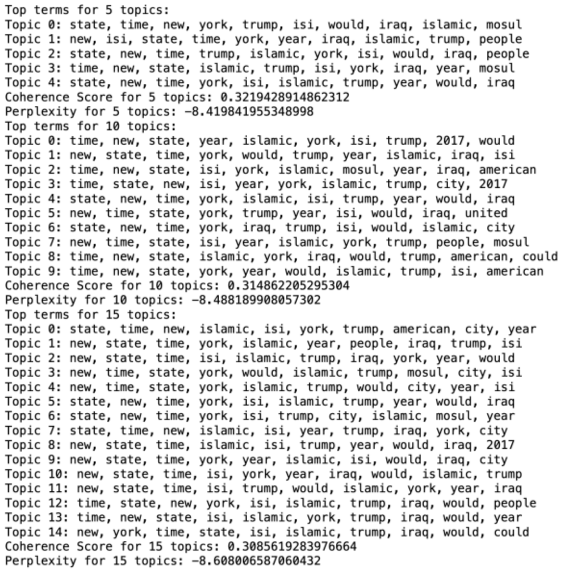
- **Exploratory analysis of corpus:**

Exploratory analysis of the corpus involves delving into its characteristics, trends, and patterns. This initial examination aims to uncover insights, identify key themes, and gain a comprehensive understanding of the content within the dataset. Through techniques such as frequency analysis, distribution exploration, and visualization, researchers can unveil valuable information about the nature of the corpus and inform subsequent analytical approaches.





A) Selection of Best Number of Topics:

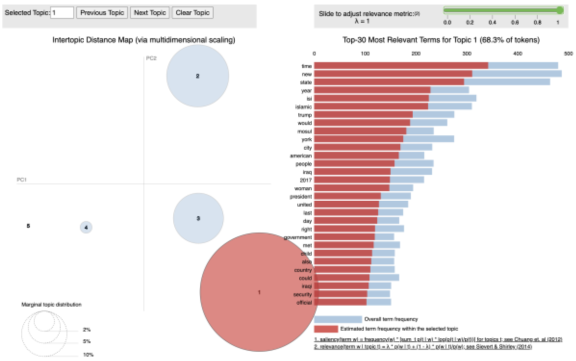


```
Top terms for 5 topics:
Topic 0: state, time, new, york, trump, isi, would, iraq, islamic, mosul
Topic 1: new, isi, state, time, york, year, iraq, islamic, trump, people
Topic 2: state, new, time, trump, islamic, york, isi, would, iraq, people
Topic 3: time, new, state, islamic, trump, isi, york, iraq, year, mosul
Topic 4: state, new, time, york, isi, islamic, trump, year, would, iraq
Coherence Score for 5 topics: 0.3219428914862312
Perplexity for 5 topics: -8.419841955348998
Top terms for 10 topics:
Topic 0: time, new, state, year, islamic, york, isi, trump, 2017, would
Topic 1: new, state, time, york, would, trump, year, islamic, iraq, isi
Topic 2: time, new, state, isi, york, islamic, mosul, year, iraq, american
Topic 3: time, state, new, isi, year, york, islamic, trump, city, 2017
Topic 4: state, new, time, york, islamic, isi, trump, year, would, iraq
Topic 5: new, time, state, york, trump, year, isi, would, iraq, united
Topic 6: state, new, time, york, iraq, trump, isi, would, islamic, city
Topic 7: new, time, state, isi, year, islamic, york, trump, people, mosul
Topic 8: time, new, state, islamic, york, iraq, would, trump, american, could
Topic 9: time, new, state, york, year, would, islamic, trump, isi, american
Coherence Score for 10 topics: 0.314862205295304
Perplexity for 10 topics: -8.488189908057302
Top terms for 15 topics:
Topic 0: state, time, new, islamic, isi, york, trump, american, city, year
Topic 1: new, state, time, york, islamic, year, people, iraq, trump, isi
Topic 2: new, state, time, isi, islamic, trump, iraq, york, year, would
Topic 3: new, time, state, york, would, islamic, trump, mosul, city, isi
Topic 4: new, time, state, york, islamic, trump, would, city, year, isi
Topic 5: state, new, time, york, isi, islamic, trump, year, would, iraq
Topic 6: state, new, time, york, isi, trump, city, islamic, mosul, year
Topic 7: state, time, new, islamic, isi, year, trump, iraq, york, city
Topic 8: new, state, time, islamic, isi, trump, year, would, iraq, 2017
Topic 9: new, state, time, york, year, islamic, isi, would, iraq, city
Topic 10: new, state, time, isi, york, year, iraq, would, islamic, trump
Topic 11: new, state, time, isi, trump, would, islamic, york, year, iraq
Topic 12: time, state, new, york, isi, islamic, trump, iraq, would, people
Topic 13: time, new, state, isi, islamic, york, trump, iraq, would, year
Topic 14: new, york, time, state, isi, islamic, trump, iraq, would, could
Coherence Score for 15 topics: 0.3085619283976664
Perplexity for 15 topics: -8.608006587060432
```

In the process of determining the optimal number of topics, the coherence and perplexity values for 5 topics are 0.321 and -8.41, for 10 topics are 0.31 and -8.44, and for 15 topics are 0.30 and -8.60. It is noticeable that as the number of topics increases, there is a trend of decreasing coherence and perplexity.

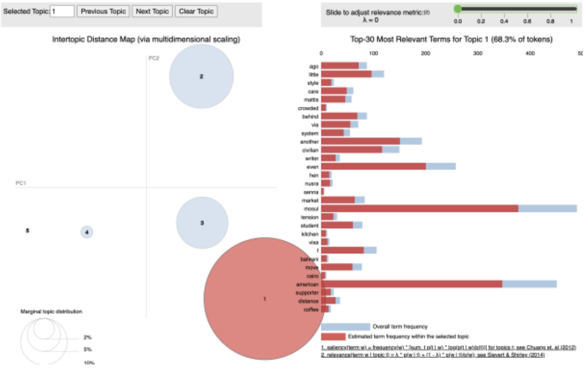
B) Visualization and Inference:

We tried to cluster a total of 5 models, and what we see on the map is shown below.



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)

When choosing Cluster 1 with a lambda (λ) value of 1, the top 30 associated words for each cluster reveal a lack of alignment with our desired topic. Common words such as "time," "new," "state," and "year" dominate, hindering the identification of relevant themes. To enhance topic modeling, we propose the implementation of a removal list, excluding these common words. Despite the prevalence of a cluster representing the majority of tokens, we have opted to maintain its current state.

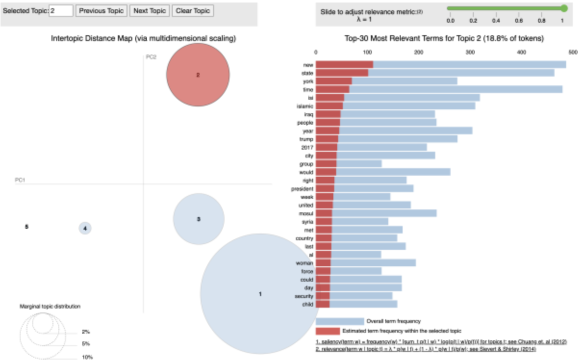


Selected Topic: 1

Slide to adjust relevance metric: λ = 0

Previous Topic
Next Topic
Clear Topic
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (68.3% of tokens)
0
100
200
300
400
500
ago
little
style
care
mattis
crowded
behind
via
system
another
civilian
writer
even
hen
nusra
senna
market
mosul
tension
student
kitchen
visa
f
behrani
move
cairo
american
supporter
distance
coffee
Overall term frequency
Estimated term frequency within the selected topic
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)

In the selection of Cluster 1 with a lambda (λ) value of 0, examination of the top 30 associated words for each cluster reveals a notable deviation from our intended topic. Common words like "mosul," "american," "even," "another," "little," and "civilian" take precedence, complicating the identification of pertinent themes. To refine our topic modeling, we suggest the incorporation of a removal list, specifically targeting these common words. Despite the prevalence of a cluster encompassing the majority of tokens, we have chosen to maintain its current configuration.

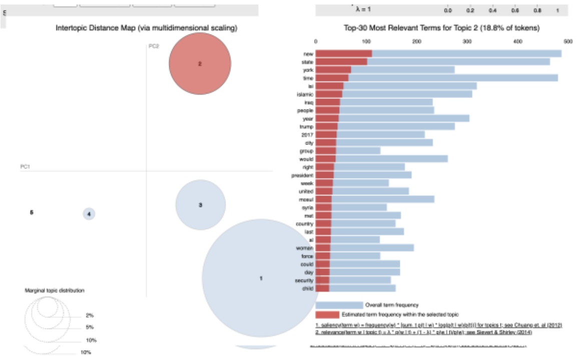


Selected Topic: 2

Slide to adjust relevance metric: λ = 1

Previous Topic
Next Topic
Clear Topic
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (18.8% of tokens)
0
100
200
300
400
500
new
state
york
time
isi
islamic
iraq
people
year
trump
2017
city
group
would
right
president
week
united
mosul
syria
met
country
last
al
woman
force
could
day
security
child
Overall term frequency
Estimated term frequency within the selected topic
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)

Here while selecting the cluster 2 with (λ) value 1 we can see the top 30 words associated with each cluster, there are words like new, state, york, time, and isi these are the common words. we can infer that the words in it will contain only related to it and we can see that the top words are related and it contains 18.8 tokens.



λ = 1

0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
2%
5%
10%
10%
Top-30 Most Relevant Terms for Topic 2 (18.8% of tokens)
0
100
200
300
400
500
new
state
york
time
isi
islamic
iraq
people
year
trump
2017
city
group
would
right
president
week
united
mosul
syria
met
country
last
al
woman
force
could
day
security
child
Overall term frequency
Estimated term frequency within the selected topic
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)

Cluster 2, with a λ value of 0, doesn't mix with other clusters. This means all the words in it are closely related, and the top words confirm this, making up 18.8% of the tokens.

4. Conclusion

In conclusion, this study navigated a meticulous process from building a comprehensive corpus sourced from Factiva to conducting in-depth analyses using advanced natural language processing techniques. The initial dataset, originating from renowned sources such as the Wall Street Journal and the New York Times, underwent a systematic transformation into a refined corpus of 1620 articles, ensuring accuracy and relevance. Through rigorous pre-processing steps, including cleansing, tokenization, and stemming, the text data was optimized for further analysis.

The feature extraction phase employed a document-term matrix and visually represented prominent words using a word cloud. The implementation of Topic Modeling, specifically Latent Dirichlet Allocation (LDA), revealed a proficient model, evidenced by low perplexity scores indicating high predictive accuracy for new data. However, the study identified opportunities for enhancing semantic coherence among topics, as reflected in relatively lower coherence scores.

The optimal number of topics, determined to be 19 through a comprehensive evaluation of perplexity and coherence metrics, signifies a delicate balance between model complexity and effectiveness. The findings emphasize the model's success in generating coherent and predictive topics, underscoring the importance of thoughtful consideration when determining the number of topics to ensure optimal performance. This study contributes valuable insights into text analysis methodologies, shedding light on the nuances of balancing model intricacy with interpretability and predictive power.