Creating Semantic Spaces Using Large Document Clustering

Team: Tristan Weger, Yahya Emara, & Ryan Rubadue
Advisor: Professor Ali Minai

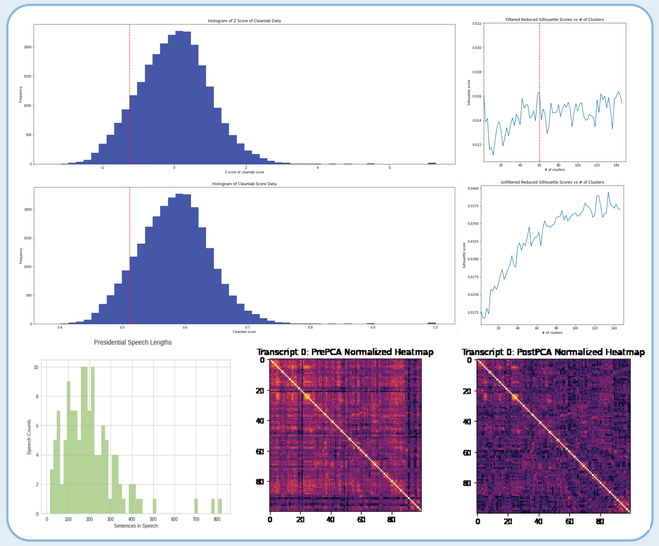
University of CINCINNATI
COLLEGE OF ENGINEERING AND APPLIED SCIENCE

TASK

1. Analyze long text documents, where text for each doc is not confined to a single subject area
2. Create a representation of natural language that captures meaning (semantic space) using long documents
3. Compare performance of algorithm on filtered (using cleanlab) and unfiltered raw data

ANALYSIS

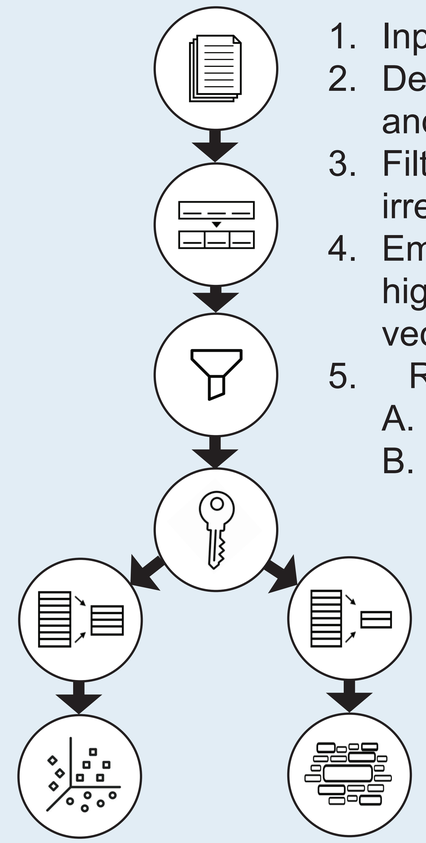
- Obtain various metrics
- Compare
  - Segment vs sentence
  - FItered vs unfiltered
  - Reduced vs not reduced
  - Optimized vs original
- Analyze
  - Cleanlab score
  - Silhouette score
  - Z score
  - Coherence score
  - Cosine Similarity



TECHNOLOGY

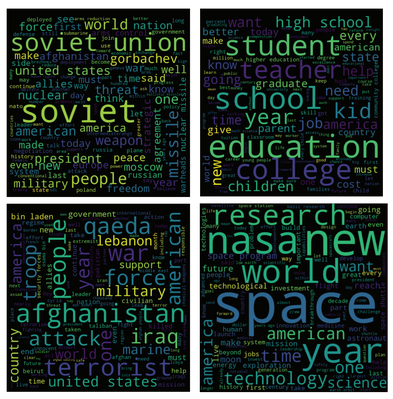
- Python
- NLTK
- SBERT
- UMAP
- Cleanlab
- Numpy
- Pandas
- Sklearn
- Torch
- MatplotLib

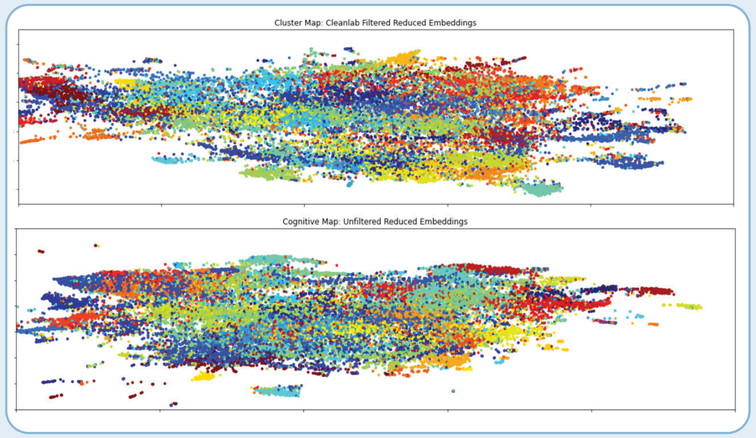
ALGORITHM OVERVIEW

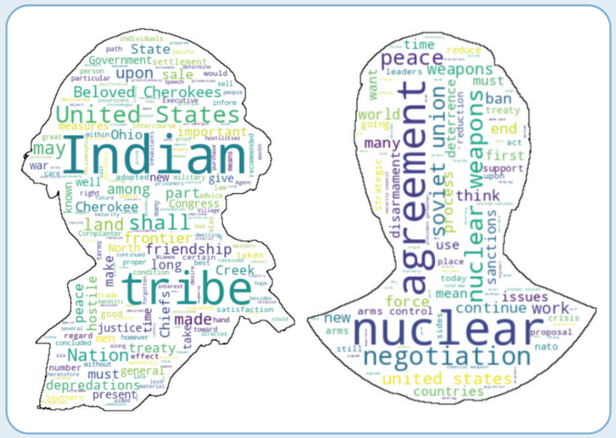


1. Input long text
2. Deconstruct text to segment and sentence tokens
3. Filter tokens to remove irrelevent data
4. Embed filtered tokens into high dimension semantic vectors (786 Dimensions)
5. Reduce dimensions:
   A. 2 dimensions
   B. 95% variance is retained (~200 dimensions)
6. Visualize outputs:
   A. Wordcloud
   B. Cognitive map
7. Compare results using provided metrics

RESULTS







CONCLUSION

- Filtering Documents at Sentence Level is EFFECTIVE
- Observed X percent of sentences in presidential speeches is 'relevent'
- *Splitting documents at sentence level is BETTER than at paragraph
- Reducing embedding dimensions resulted in more coherent outputs
- Creation of a functional semantic space was achieved