Labeling News Headline Topics with Unsupervised Learning

Matthew Radovan, Chris Cross, Sasankh Munukutla

CS 221: Artificial Intelligence
Stanford University

Problem

- Unsupervised topic modeling
  - Extracting representative topic words in a text dataset
  - Assigning texts to topics
- Challenging when working on short texts without pretrained semantic knowledge

Motivation

- 70% of Americans are fatigued by the amount of news available
- Effective topic modelling makes news understandable

Dataset

- Dataset of ~8000 article headlines and descriptions from NewsAPI

Pre-processing

Initial Text	UK Supreme Court hears government side in vital Brexit case
Lowercase + No Punctuation (Simple)	uk supreme court hears government side in vital brexit case
Stopwords Removed	uk supreme court hears government side vital brexit case
Only Nouns (PoS Tagged)	UK Supreme Court government side Brexit case
Pruned Word Length (Word Length)	supreme government brexit

Clustering Algorithms

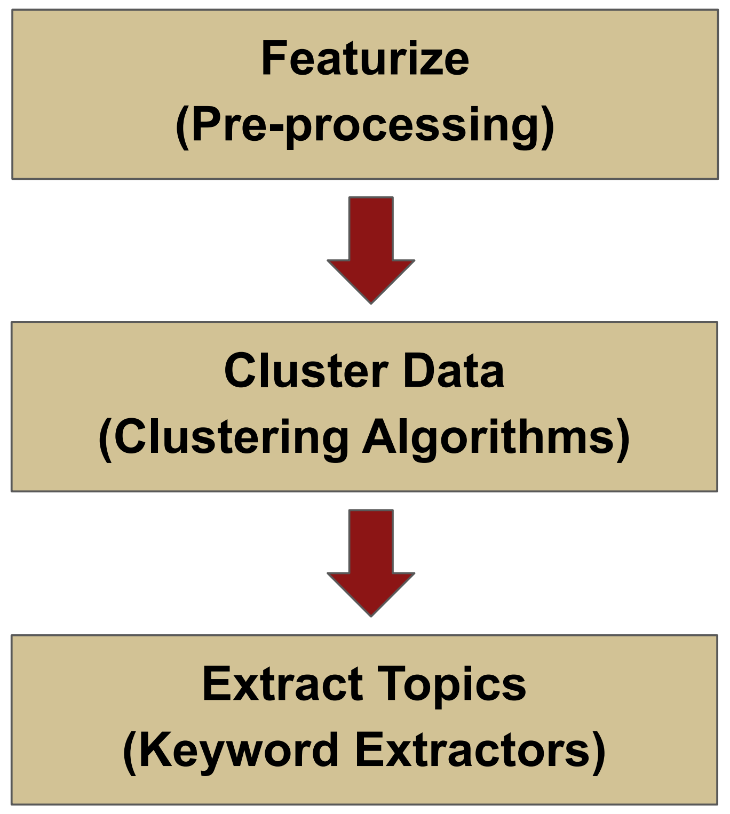
Latent Dirichlet Allocation (LDA)	**K-Means**	**Topic Keyword Model (TKM)**
• Most common approach - baseline • Creates clusters based on topic and word distributions	• Given featurized data, optimize clusters' means and cluster assignments for each articles	• Scores for each potential keyword w/ joint probabilities • Finds probability of topic given sum of keyword scores

*TKM and LDA performed poorly, so K-Means was utilized as clustering algorithm

Cluster keyword extractors

Count-based	**Centroid-based**
• Simplest, and baseline - disregards uniqueness • Ranked based on frequency of term within cluster	• Maximize uniqueness of words prior to relevance of words • Assign selected words to clusters based on relevance to cluster
Logistic regression-based	**Pseudo TF-IDF**
• logistic regression on features (label = cluster assignment) • Selects highest weight coefficients for each topic	• Attempts to factor in uniqueness of keywords in topic extraction • Uses inverse frequency across all clusters to normalize

Approach



Featurize (Pre-processing)

Cluster Data (Clustering Algorithms)

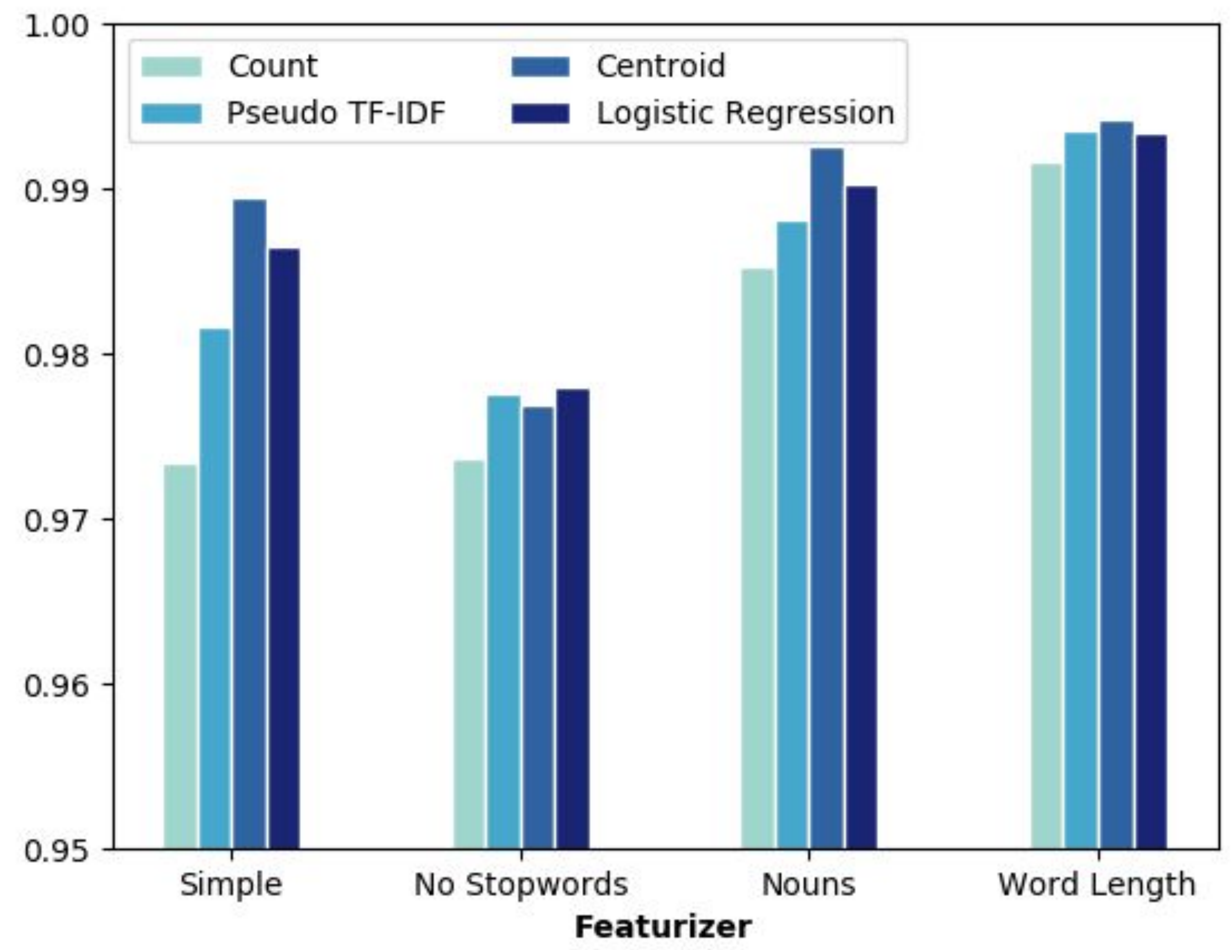
Extract Topics (Keyword Extractors)

Challenges

- Curse of dimensionality for K-Means
- High computation time for certain models and keyword extractors
- Clustering algorithms (and unsupervised learning in general) hard to evaluate due to subjectivity
  - Hence, qualitative analysis of topics becomes necessary

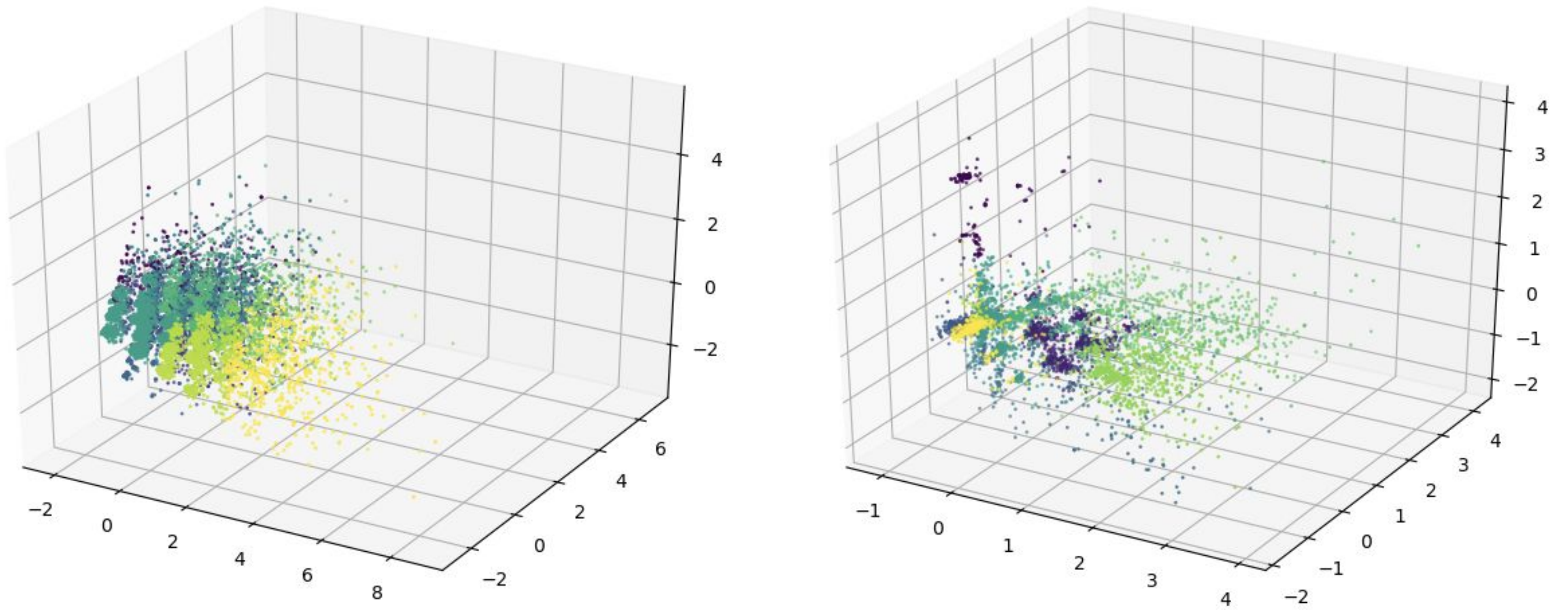
Quantitative Results

Ability to Re-classify Keyword-Only Texts



Accuracy of predicting clusters through keyword-only text vs. through full text (same classifier; 100% accuracy on full text)

Comparison of Best and Worst Featurizer
(in terms of clustering - Davies-Bouldin Index)



Simple (Worst)

Nouns (Best)

Qualitative Analysis

Examples of articles from good cluster [word length]

- Article on Cory Booker's promises for worker rights
- Beto O'Rourke's gun buy-back program
- Democrats targeting Gen Z for the upcoming election

Examples of articles from bad cluster [simple]

- Trump and Graham clashing on Iran policy
- Iowa poll on Democratic candidates

Good examples of keywords

- "whistleblower", "investigate", "president", "zelensky", "ukraine"

Bad examples of keywords

- "in", "the", "a", "to", "of", "and", "for", "on", "as", "is", "his", "trump"

Key Insights

- Good accuracy does not mean good clustering or vice versa (*Word Length* had the best accuracy but *Nouns* had the best clustering)
- Word Length is the best featurizer in terms of accuracy
- Centroid is generally the best extractor, but best extractor can vary by featurizer (*Logistic Regression* was best on *No Stopwords*)
- *K-Means* performed far better for clustering than *TKM* and *LDA*

Conclusion

Ability to automatically determine key topics in news with
>99% accuracy

Social Impact

- Identify current and relevant issues
- Understand specific topics, especially critical issues like climate change, without being inundated by unrelated articles

Acknowledgements

- NewsAPI
- scikit-learn
- Special thanks to our mentor Chuma Kabaghe

References

- https://arxiv.org/abs/1710.02650
- http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- https://www.pewresearch.org/fact-tank/2018/06/05/almost-seven-in-ten-americans-have-news-fatigue-more-among-republicans/