Topic Recognition on New York Times Articles-Executive summary

Team 20: Ravi Tripathi, Touseef Haider, Ping Wan, Schinella D'Souza, Alessandro Malusà, Craig Franze

GitHub: https://github.com/rptripathi/NYTimes

Overview and Goals

In this project, we analyse New York Times articles and build two models to identify topics from the collection of articles, assign each article one or more topics (weighted by relevance), and given an article we build a recommendation system to suggest similar articles. Topic recognition is based solely on abstracts of the articles. The models we used were Latent Dirichlet Allocation (LDA) and BERTopic.

In the absence of user interaction data, implementing unsupervised methods like LDA or BERTopic is useful to tackle a cold start problem. Our stakeholders include the New York Times and possible other news platforms.

The Dataset

The dataset consists of data on New York Times articles from May 15, 2023 till May 14, 2024. The data was scraped through the official New York Times API, and consists of 42623 articles. The data associated with each article includes abstract, lead paragraph, URL, headline, published date, snippets.

Data Exploration and Cleaning

Based on exploration of abstracts, snippets, and lead paragraphs, we used abstracts as our corpus of documents to feed to the LDA model. Abstracts were converted to lowercase, punctuation was removed, dates and acronyms kept in while dropping words with at most three letters, and words brought to their base form. Certain articles contained puzzles and quotations so we dropped these as they accounted for only 3.38% of total abstracts. Abstracts were then tokenized (i.e., converted to a vector of words).

Models

The models we used were LDA and BERTopic. A brief summary of each is given below but a further discussion can be found in the readme file on GitHub.

- The LDA model is a generative statistical model that takes in a corpus of documents which, in our project, were New York Times article abstracts. After specifying a number of keywords and a number of topics, the model then generates a list of top keywords associated with each topic. We used cross validation and log-likelihood scores to determine that 10 topics and 30 keywords were optimal.
- BERTopic is a topic recognition model that uses sentence transformers to embed our documents into a high dimensional space. Then UMAP reduces the dimensionality for clustering, and HDBSCAN performs hierarchical clustering on the data. The representative documents from the identified clusters, or topics, from BERTopic were then fed into a large language model, Ollama, to generate short sensible labels. The model identified 420 emergent topics over the past year and produces time series for those topics to show their popularity over time.

Future Directions

- Include additional data, for e.g. lead paragraphs and keywords, into our analysis.
- Enhance model performance with hyperparameter tuning.
- Build a UI design which takes user inputs as keywords and recommends articles based on a similarity metric. Ask users about the quality of the recommendation.
- Incorporate user feedback data, either existing ones, or from their reviews on our UI recommendations. Use this data to suggest more personalised recommendations.
- Analyse temporal trends.

Acknowledgements

We thank the Erdős Institute for the opportunity and training, our mentor Matthew Graham for providing us with valuable insights and encouragement, and to the New York Times for access to their articles.