Jay Kwon | Metis Data Science and Engineering (flex program) | Module 5 – NLP (May, 2022)

**Financial News Topic Modeling and Recommender**

**Abstract**

The main goal of this project was to define different topics of financial news articles and to gain useful insights about those topics for the readers. A recommender was built that displays similar articles for any given article. The types of topics revealed and the proportions of those topics based on different segmentations of the data would be useful to anyone interested in finance including investors, financial services, rating agencies, and financial news agencies. Financial news articles published online from January to March of 2018 were obtained from a [Kaggle dataset](https://www.kaggle.com/datasets/jeet2016/us-financial-news-articles?datasetId=49948). Preprocessing included using the spaCy library to lemmatize and filter the lemma’s to nouns only. Scikit-learn’s Latent Dirichlet Allocation (LDA) module was used to model 20 topics. The most popular lemma’s for each topic and sample articles from that topic were observed to define what the topic was about. A recommender that uses cosine similarity proved to be successful in recommending similar articles.

**Design**

Anyone interested in finance, including investors and financial services, must pay close attention to the financial news. The ability to categorize the articles into different topics will not only give an overview of what was going on at the time but will enable the reader to pull up articles based on topics. Given any article of interest, a recommender will be able to pull up articles that are most similar to it.

**Data**

The dataset of articles was obtained from Kaggle:

<https://www.kaggle.com/datasets/jeet2016/us-financial-news-articles?datasetId=49948>

The dataset contains 306,242 JSON files, each representing a financial news article. Additional information about the article is included in the JSON files, including date published, title, and url. The articles are from Jan to May, 2018 and were taken from Bloomberg.com, CNBC.com, reuters.com, wsj.com, and fortune.com. The article text from the JSON files were imported into a DF, each row representing an article/document. After LDA was performed, each row represented one of 20 topics and each column represented a lemma or a bigram.

**Algorithm**

- lemmatization and assigning parts of speech with spaCy

- Latent Dirichlet Allocation (LDA) for topic modeling

* 20 topics based on top terms, many of which were clearly definable, select topics include:
  + Topic 0: Tech businesses and their management
  + Topic 2: Oil
  + Topic 4: Lawsuits
  + Topic 5: US Government
  + Topic 7: Health
  + Topic 8: Trade (export/import)
  + Topic 11: Investment Management
  + Topic 15: Sports
  + Topic 16: Business performance (financial statement metrics)
  + Topic 18: Tech
  + Topic 19: Currency and Rates

\* see presentation and notebook for top words and more topics

**Recommender**

- cosign similarity for recommender

**Tools**

spaCy for lemmatization and assigning parts

nltk to obtain a base of English stopwords

Sklearn for vectorization and LDA

Pandas to filter and manipulate data

Matplotlib for visualizations

**Communication**

- 5 minute presentation with slides

- GitHub repo: <https://github.com/jaykwon2/NLP_Financial_News_Recommender>