Data Engineering with NY Times

Sandy Weng

Abstract

The goal of this project was to use Data Engineering tools to create a pipeline starting from data ingestion to data cleaning/aggregation and ending up with a model deployed in Streamlit. I worked with data obtained by New York Times API for topic modeling and model deployment.

Design

This project was centered around building a data pipeline with NLP and topic modeling with NY Times news articles using software tools like MongoDB and Streamlit. The app for topic modeling can inform writers on what are the latest topics that are being shared in the last week. They can utilize this tool for inspiration as well as looking at a topic from a different perspective, so all sides of the story are heard.

Data

The dataset contains 112,348 articles (after dropping duplicates and rows containing null values). The fields extracted from the database were the article's title, abstract, section, author, type of material, and date. The abstract field was used for analysis and modeling.

Algorithms

The data was cleaned using Pandas. NLP was used to remove punctuations and convert the letters to lowercase on the abstract column. TF-IDF was used to tokenize and vectorize the data. NMF with 10,15, and 20 components were used for topic modeling. Unit testing was done for the API call and to check for the 'abstract' field when retrieving data.

Tools

- MongoDB for data storage
- Pandas for data manipulation
- Scikit-learn for modeling

- Matplotlib and Seaborn for plotting
- Streamlit for model deployment

Communication

- Slides and presentation