Abstract

The goal of this project is to visualize how the topics of articles on the front page of the New York Times changes over time. This project is useful because it will allow us to see the lifespan of specific news topics and allow us to gauge which topic are becoming more frequent and which topics are becoming less frequent. In order to build this model, I scraped the front page of The New York Times for the last month using the archive pages and then placed the text of all articles into a data frame. After that I created a gensim DTM model.

Design

This project would be beneficial to all news outlets that are looking to measure how news topics change over time. They will be able to see the average lifespan of each story. Also writers at the paper will be able to see if the article that they are writing is about a topic that is on the rise in frequency or declining in frequency, which will allow them to make better decisions about what they want to include in the article. The general topic modeling analysis also allows the user to see what is most common during the entire timeframe, which will give them a better sense of what is being reported commonly.

Data

The dataset contains 1,028 articles that were scraped from the New York Times archive pages. Each article has an average of 1,175 words. I decided to break down each row into articles instead of paragraphs, because many articles are written with thirty to forty very small paragraphs that do not have enough information in them for them to stand by themselves. The only columns in the data frame are the text of the article and the day the article was written.

Algorithms

The initial investigation of the dataset was done using a NMF model to generate the topics that were present when grouping the entire dataset as a singular entity. This reduced the data to only ten topics because during EDA, I looked at an NMF model with 5, 10, 15, and 20 topics and qualitatively decided that 10 topics would yield the best results. I then used this same number of topics to produce the dynamic topic model.

Tools

In order to acquire the data, I scraped the New York Times website using BeautifulSoup. Once I had the raw data, I preprocessed the data using NLTK’s WordNetLemmatizer, along with general custom function with the assistance of RegEx and Pandas. I used sklearn for both the count vectorizer as well as the NMF decomposition. For the dynamic topic model I used gensim’s ldaseqmodel.

Communication

I have included the visualizations created for this project on the presentation slides.