Programming: Text Classification

Tools Used

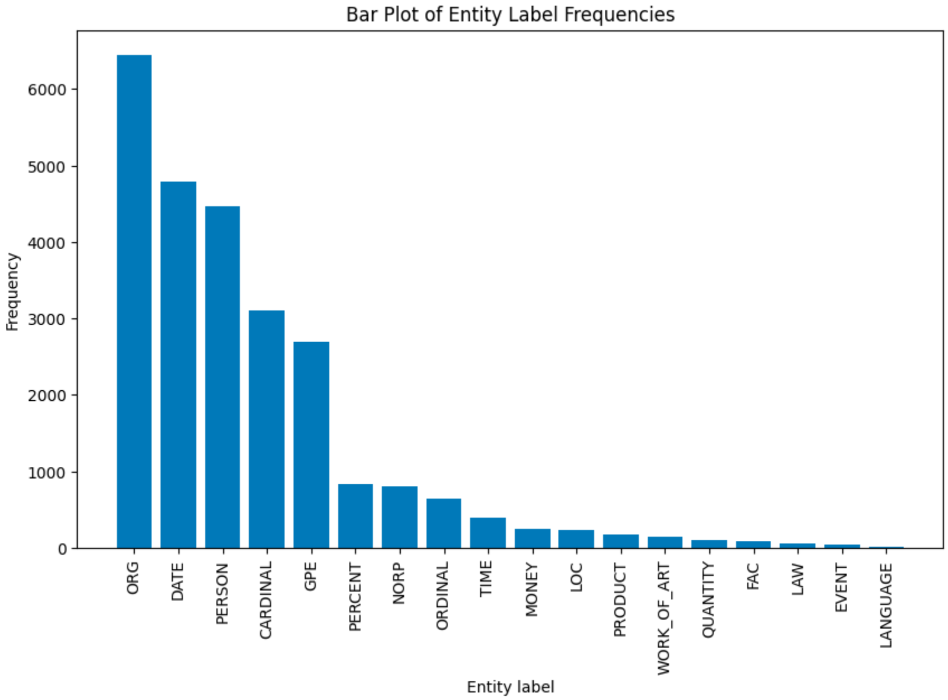
Web Scraping: BeautifulSoup

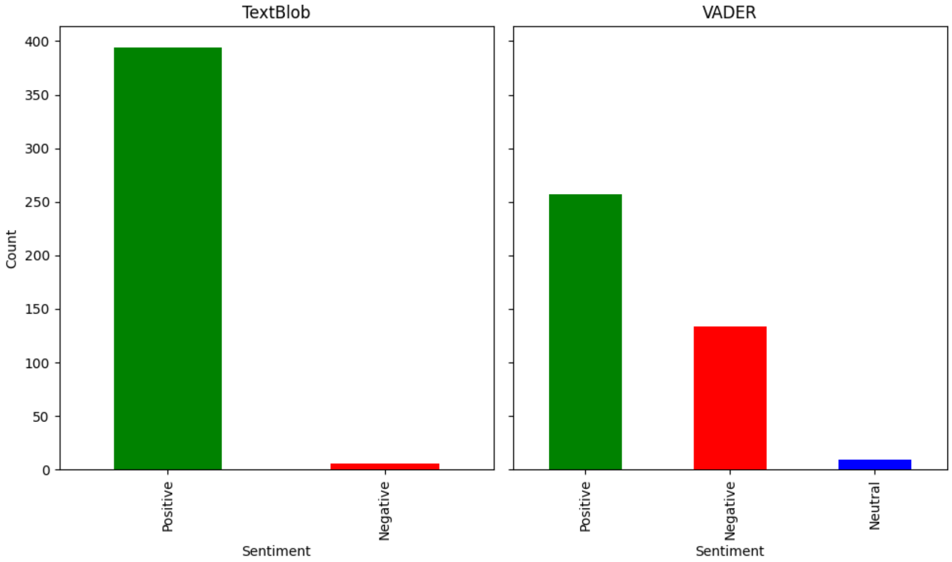
Named entity recognition: spaCy

Sentiment analysis: NLTK, TextBlob and VADER

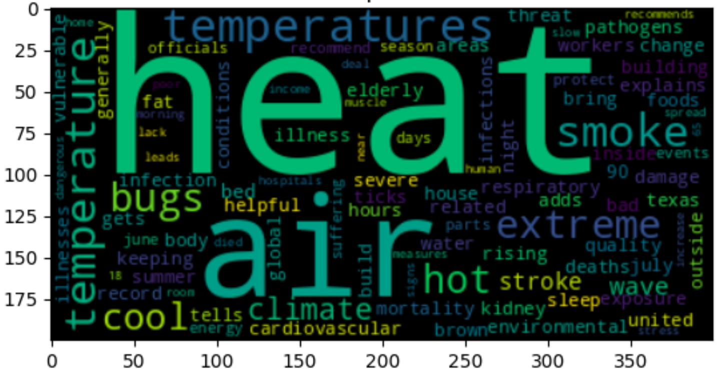
Topic modeling: sklearn and Gensim

Results

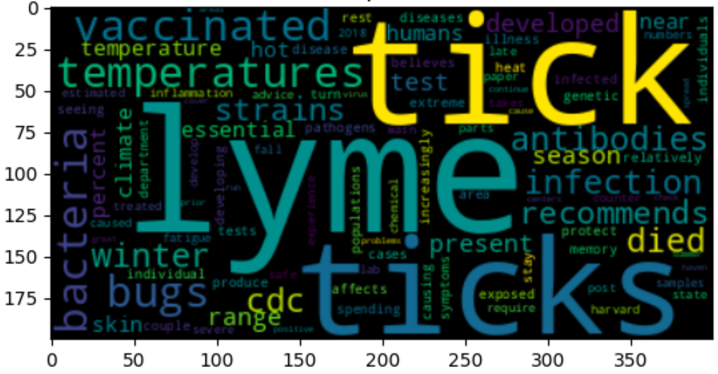




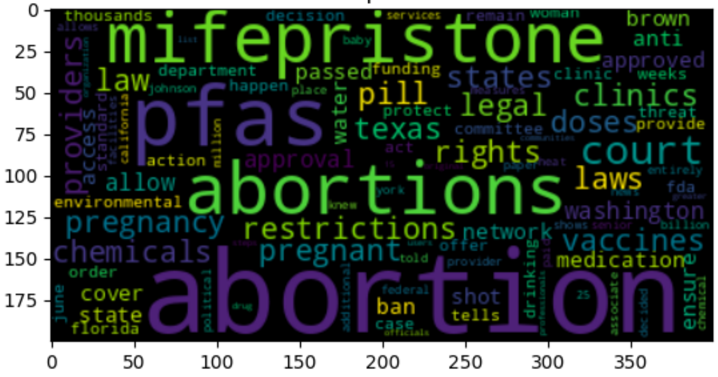
Topic 0



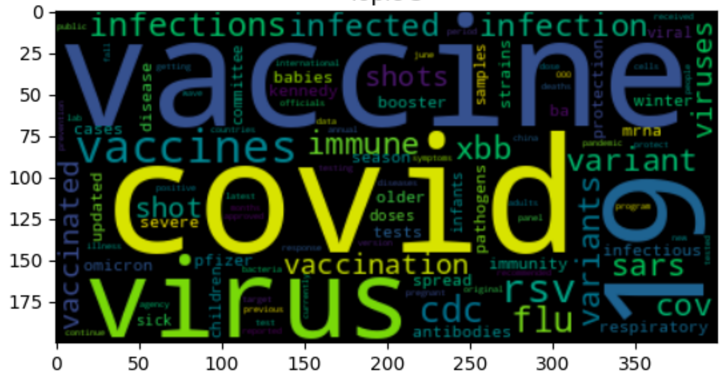
Topic 1



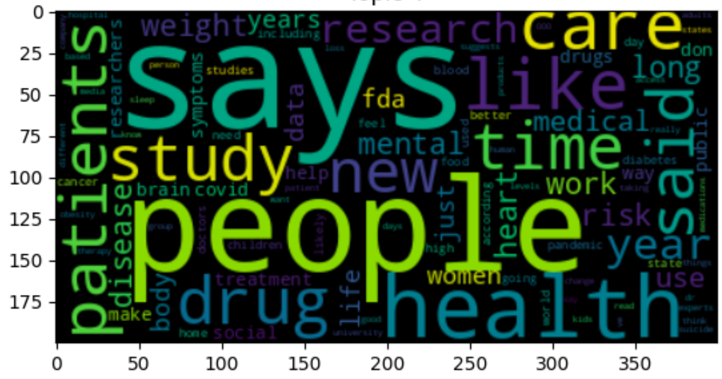
Topic 2



Topic 3



Topic 4



Analysis/Reflection

I extracted 400 health-related articles from the time.com news portal, spanning mid-April 2023 to mid-January 2024. Employing named entity recognition, I identified the predominant entities, which included organizations, dates, and persons. This aligns with the expected content of news articles, where mentions of entities such as covid-19, FDA, CDC, WHO, Pfizer, Moderna, etc., were prevalent. The articles extensively covered diverse aspects of the ongoing pandemic, encompassing topics like vaccine development, public health measures, and global initiatives to curb the virus's spread.

For sentiment analysis, I utilized both TextBlob and VADER for a comparative assessment, revealing that VADER provided more accurate results. The sentiment analysis indicated an overall prevalence of positive sentiments over negative ones. This outcome contrasts with the anticipation of a higher occurrence of negative sentiments given the context of the pandemic.

After performing topic modeling, the analysis identified five overarching topics: temperature/climate, tick/lyme, abortion, vaccine/covid, and people. The implementation of the scikit-learn package yielded more distinct topics (compared to the Gensim package), providing valuable insights into the diverse health-related themes covered in the articles.