DERIVING ALPHA FROM NEWS

Ty Painter

Objective

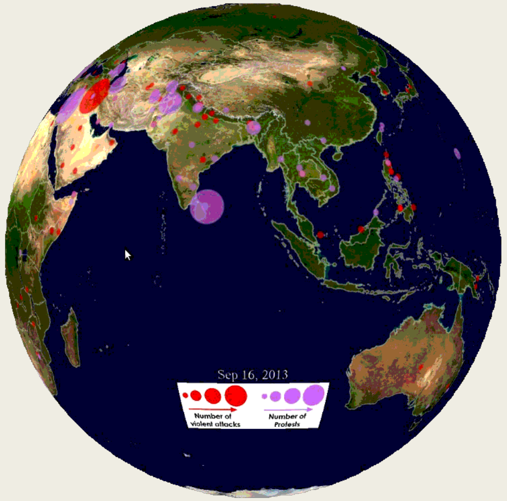
- Run NLP sentiment analysis on financial news to predict the alpha of stocks
  - *Collect news data from database*
  - *Web-scrape news article contents*
  - *Apply NLP sentiment analysis to retrieve sentiment scores*
  - *Web-scrape individual S&P 500 stock historical data*

****Alpha** refers to excess returns earned on an investment above the benchmark return**

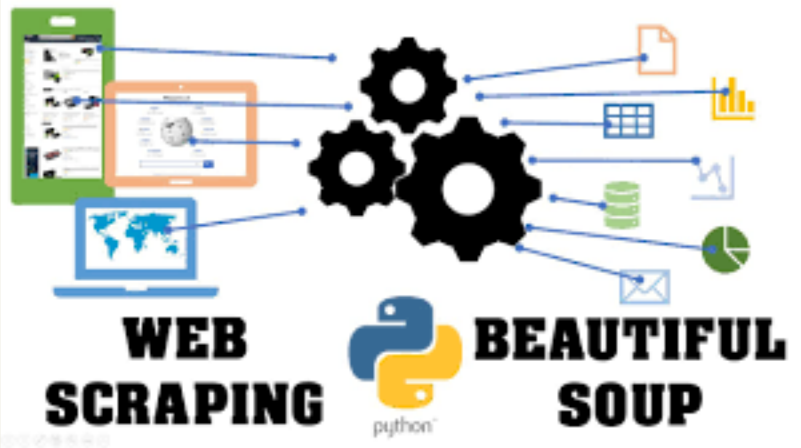
Data

- GDELT 2.0 Event Database
  - *Gather 15-minute files for all of 2021*
  - *Filtered for 3 columns*
    - Date
    - Website
    - URL
  - *Filtered on websites*
    - Yahoo.com
    - Reuters.com
    - Marketwatch.com
    - Prnewswire.com
- Individual S&P 500 stocks
  - *Historical daily open & close prices from 2021*
  - *Market cap*

The GDELT Project



Web-Scraping



- BeautifulSoup
  - *Navigated to URL and scraped article contents*
  - *Company quarterly market caps*
- ChromeDriver
  - *Yahoo finance for historical data*

NLP

- HuggingFace API (distilRoberta-financial-sentiment)
  - *Sentence by sentence*
  - *Outputs*
    - Positive
    - Negative
    - Neutral
  - *Collect summary statistics of all sentences for each output score*
    - Minimum
    - Maximum
    - Average
    - Mean

Next Steps

- Readability scores
  - *Smjsindustry or TextDescriptives*
- Tag articles with companies/stocks
  - *FuzzyWuzzy*
- Merge historical stock price and NLP sentiment scores
- Develop predictive model using sentiment scores, readability scores, and market cap as features to predict stock price movement

Challenges

- Dirty data
  - *Different formats*
  - *Dead URLs*
- Web-scraping & HuggingFace API
- Scalability