**Update #1 (1/13/21)**

* Filter by most popular sites
* Decimal columns are GDELT NLP version, random letters are mapped
* Code on dashboard to run searches through dashboard
* Tone = Polarity
* Start from 2021; but go back as far as possible hopefully Mar 2020 (COVID)
* Run all data/features for today and then project for tomorrow
* Start with distil Roberta model
* Opening and closing, opening prices are important because of changes over night
* Readability
  + Hard to read 🡪 hide something
  + Easy to read 🡪 good performance
  + Look for opensource packages
  + 4th grade reading level, for news not necessarily
  + Word count, paragraph count, sentence length

**Update #2 (1/27/22)**

* Add sector and market cap to stock tickers
* Filter columns and websites
* Read in a day or week at a time
* Scrape 5-10 financial/popular websites

**Update #3 (2/10/22)**

* Additional web scraping sites

|  |
| --- |
| reuters.com |
| yahoo.com |
| prnewswire.com |
| marketwatch.com |

* Add back in GDELT NLP Scores
* Use quarterly or monthly market cap for stock history
  + Label small/mid/large market caps for S&P 500
* Pass in sentence/paragraphs to NLP API then take avg. score

**Meeting with Dr. Blocher (2/23/21)**

* File storage (couple 100s GB, Box?)
  + Try and put filtering and and getting text and then store NLP scores
  + Try using scratch instead data
  + Use Accre compute and Amazon S3 storage (Stewart)
  + Amazon EC2 computing
  + Jobarray in accre
  + Tail-f in unix might add logging capability, ssh to server log file, os package write to somewhere besides .out file
  + Log for j for logging in python
* How to pass text into NLP (sentence take days)
  + Find maximum capacity of Hugging Face then cut into minimum slices
  + Min, max, mean, median
* How to use HuggingFace NLP API (PyTorch with tensors)
  + Ask Ashely

**Update # 4 (2/24/21)**

* How to tag article
  + String match
  + NLP methods (fuzzy matching – fuzzywuzzy)
  + Named entity recognition
* S&P 500 market cap labels
  + Hard cutoffs
    - Large: $100B
    - Mid: $50B
    - Small: <$50B
    - Look more into distribution
* Web scraping
  + Business insider
  + PR sites
* Only need date, website, URL when reading in data