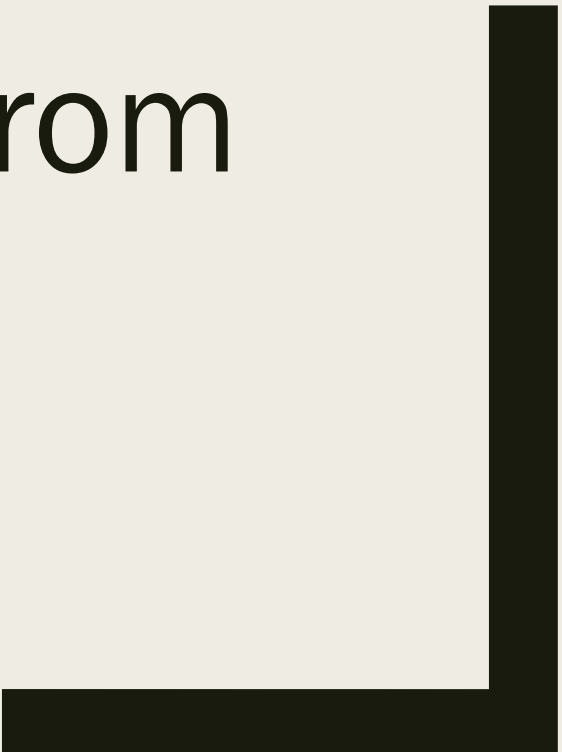# Deriving Alpha from News: GDELT

Ty Painter
Capstone Update #6
3/31/22

NLP

- Sentiment scores
  - *Added net scores (positive – negative)*
  - *Roberta is producing many neutral scores*
    - Single sentence could skew results
    - Focus on mean/median rather than max/min?
- Article Tags
  - *Fuzzywuzzy package*
  - *Token_set_ratio()*
    - split the tokens into two groups: intersection and remainder; use those sets to build a comparison string.
  - *Filter based on what value?*
- Readability scores
  - *Readability package*
  - *What metrics are most useful?*

Microsoft Excel
Worksheet

Fuzzywuzzy


```python
tickers = ["AAPL", "TSLA", "AMZN", "GOOG"]
companies = ["Apple", "Tesla", "Amazon", "Alphabet", "Facebook"]
ticks_and_comps = tickers + companies
scores = process.extract(article,
                         ticks_and_comps,
                         scorer=fuzz.token_set_ratio)
print(scores)
tags = [scores[i][0] for i in range(len(scores)) if scores[i][1]>85]
print(tags)

[('AAPL', 100), ('Apple', 100), ('Alphabet', 2), ('Facebook', 2), ('TSLA', 1)]
['AAPL', 'Apple']
```

Readability


```
results = readability.getmeasures(article, lang='en')
results


OrderedDict([('readability grades',
              OrderedDict([('Kincaid', 92.63419913419914),
                           ('ARI', 118.35402597402597),
                           ('Coleman-Liau', 14.386707952380956),
                           ('FleschReadingEase', -157.64298701298696),
                           ('GunningFogIndex', 99.15324675324676),
                           ('LIX', 261.73593073593076),
                           ('SMOGIndex', 37.20526275297414),
                           ('RIX', 71.0),
                           ('DaleChallIndex', 21.724532900432898)])),
             ('sentence info',
              OrderedDict([('characters_per_word', 5.1558441558441555),
                           ('syll_per_word', 1.5367965367965368),
                           ('words_per_sentence', 231.0),
                           ('sentences_per_paragraph', 1.0),
                           ('type_token_ratio', 0.6536796536796536),
                           ('characters', 1191),
                           ('syllables', 355),
                           ('words', 231),
                           ('wordtypes', 151),
                           ('sentences', 1),
                           ('paragraphs', 1),
                           ('long_words', 71),
                           ('complex_words', 39),
                           ('complex_words_dc', 97)])),
             ('word usage',
              OrderedDict([('tobeverb', 12),
                           ('auxverb', 1),
                           ('conjunction', 6),
                           ('pronoun', 3),
                           ('preposition', 38),
                           ('nominalization', 8)])),
             ('sentence beginnings',
              OrderedDict([('pronoun', 0),
                           ('interrogative', 0),
                           ('article', 0),
                           ('subordination', 0),
                           ('conjunction', 0),
                           ('preposition', 0)]))])
```

Next Steps

- Find computation environments

- Apply article tags and readability scores to the rest of the data

- Merge historical stock price and market cap data with sentiment scores, article tags, and readability scores

- Begin modeling