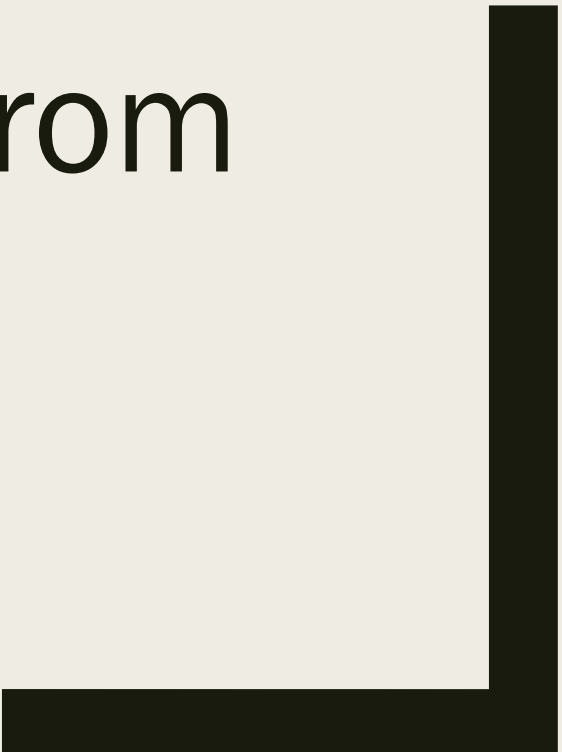




Deriving Alpha from News: GDELT

Ty Painter
Capstone Update #2
1/27/21



GDELTA Data

- Automated a script to download all the gkg.csv files for 2021
 - *Trouble handling data (~25+ days of data)*
- Parallel processing

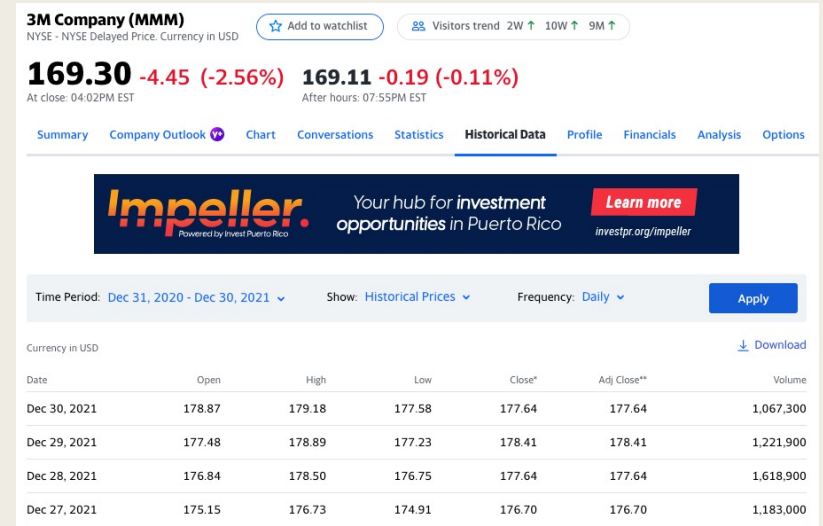
Stock Data

- Gathered S&P 500 companies ([Wikipedia](#))

S&P 500 component stocks [\[edit \]](#)

Symbol ↕	Security ↕	SEC filings ↕	GICS Sector ↕	GICS Sub-Industry ↕	Headquarters Location ↕	Date first added ↕	CIK ↕	Founded ↕
MMM ↗	3M	reports ↗	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1976-08-09	0000066740	1902
AOS ↗	A. O. Smith	reports ↗	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	0000091142	1916
ABT ↗	Abbott	reports ↗	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800	1888
ABBV ↗	AbbVie	reports ↗	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
ABMD ↗	Abiomed	reports ↗	Health Care	Health Care Equipment	Danvers, Massachusetts	2018-05-31	0000815094	1981
ACN ↗	Accenture	reports ↗	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989

- Pass ticker into Yahoo Finance
 - *Historical Data (12/31/20 – 12/30/21)*



NLP - Readability

- [smjsindustry](#)

- *Client library of Amazon SageMaker Jumpstart*
- *“process financial text documents with features such as summarization and scoring for sentiment, litigiousness, risk, and readability”*

- [TextDescriptives](#)

- Python library
- *“used to calculate several descriptive statistics, readability metrics, and metrics related to dependency distance”*

Next Steps

- Finished loading all GDELT data
- Initial EDA on GDELT data
 - *Filter most popular websites*
 - *Volume count for keywords, companies, sectors (Michael's code)*
- Start developing web scraping templates
 - *How many websites?*
- Connect HuggingFace API ([distilRoberta-financial-sentiment](#))
- [GitHub](#)