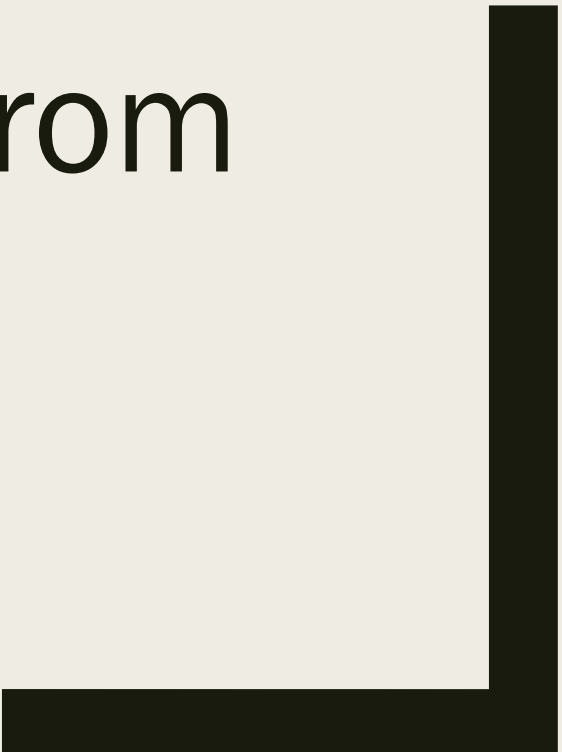




Deriving Alpha from News: GDELT

Ty Painter
Capstone Update #4
2/24/21



GDELT Data

- Storage issues
 - *Amazon S3*
- Filter rows by website
 - *Additional filters?*
- How to assign sentiment scores?
 - *Group by tags/keywords & date?*
 - *Extract company names/tickers from articles?*

S&P 500 Data

- Gathered quarterly market cap size
- Add a label for small, mid, large cap sizes for S&P 500
 - *Hard cutoffs? (\$100B, \$50B)*
 - *Thirds?*
 - *Quantiles?*

Web Scraping

- Added websites in addition to yahoo.com
 - *reuters.com*
 - *prnewswire.com*
 - *marketwatch.com*

	website	counts
0	iheart.com	161187
1	msn.com	81981
2	medium.com	48707
3	reuters.com	37931
4	yahoo.com	28885
5	indiatimes.com	26354
6	prnewswire.com	22544
7	dailymail.co.uk	18859
8	texasguardian.com	18031
9	apnews.com	17731
10	sandiegosun.com	16727
11	newyorktelegraph.com	16668
12	ianslive.in	13846
13	sfgate.com	12787
14	chron.com	12676
15	marketwatch.com	12166
16	washingtontimes.com	11970
17	onenewspage.com	11206
18	thenews.com.pk	10850
19	thehindu.com	10523
20	tmcnet.com	10394
21	finanznachrichten.de	9590
22	business-standard.com	9512
23	upstrack.com	9483
24	investigate.co.uk	9434
25	chinadaily.com.cn	9295
26	express.co.uk	8922
27	cnet.com	8545
28	freerepublic.com	8469
29	lmtonline.com	8129
30	prokerala.com	8111

NLP

- Template for [ProsusAI/finbert](#)
 - [distilRoberta-financial-sentiment](#)
- Pass in maximum capacity of characters (~500 limit)
 - *Pass in maximum number of sentences staying below character limit*
- Slice article into minimum number of sections
- Gather max, min, mean, median sentiment scores for all sections

Next Steps

- Find storage
- Filter and tag articles for sentiment analysis
- Add market cap size labels
- Create sentiment score template