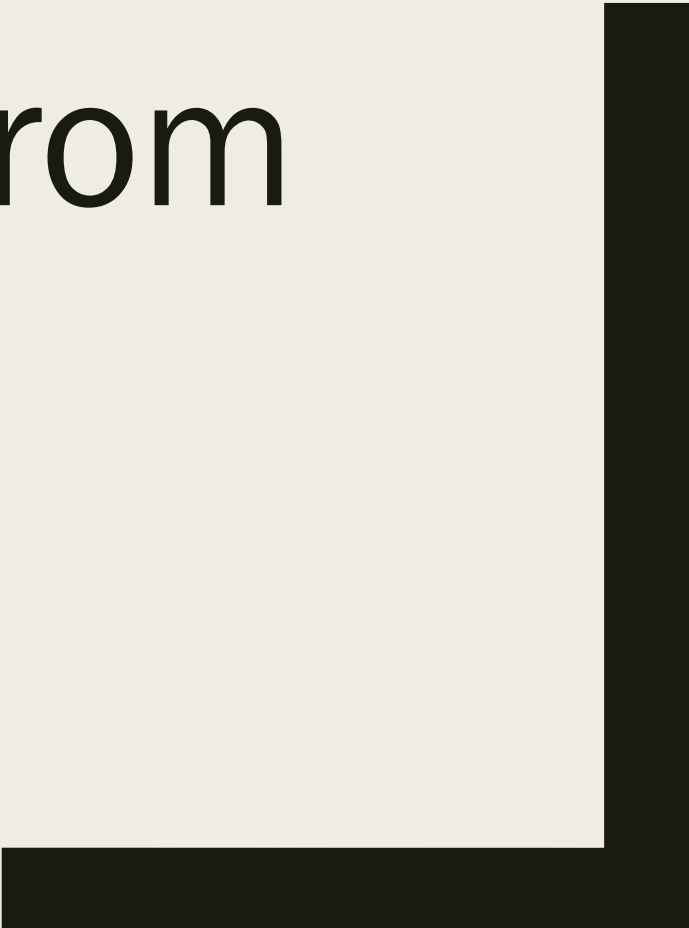




Deriving Alpha from News: GDELT

Ty Painter
Capstone Update #5
3/9/22



GDELT Data

- Storage issues resolved
 - *Removed columns to only contain date, website, URL*
 - *~15MB per month*

S&P 500 Data

- Gathered yearly market cap size
- Add a label for small, mid, large cap sizes for S&P 500
 - *Large:* $\geq \$100B$
 - *Mid:* $\$50B - \$100B$
 - *Small:* $< \$50B$

NLP

- Computation issues
 - *Google Colab*
 - *Amazon EC2*
 - *ACCRE*
- Web-scraping template issue
 - *Reuters.com*
- Sentiment scores
 - *Positive, Neutral, Negative*
 - Min, Max, Median, Mean
- Tag articles based on company name/stock ticker
 - *Fuzzywuzzy package*



Microsoft Excel
Worksheet

Next Steps

- Find computation environments
- Resolve reuters.com web scraping issues
- Create templates to use NLP for readability metrics
- Use fuzzywuzzy package to tag articles