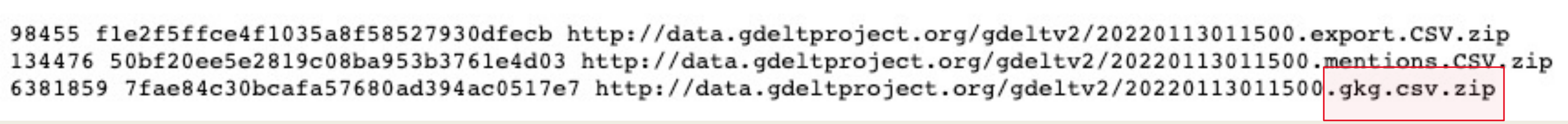
Deriving Alpha from News: GDELT

Ty Painter
Capstone Update #1
1/13/21

DATA SOURCES

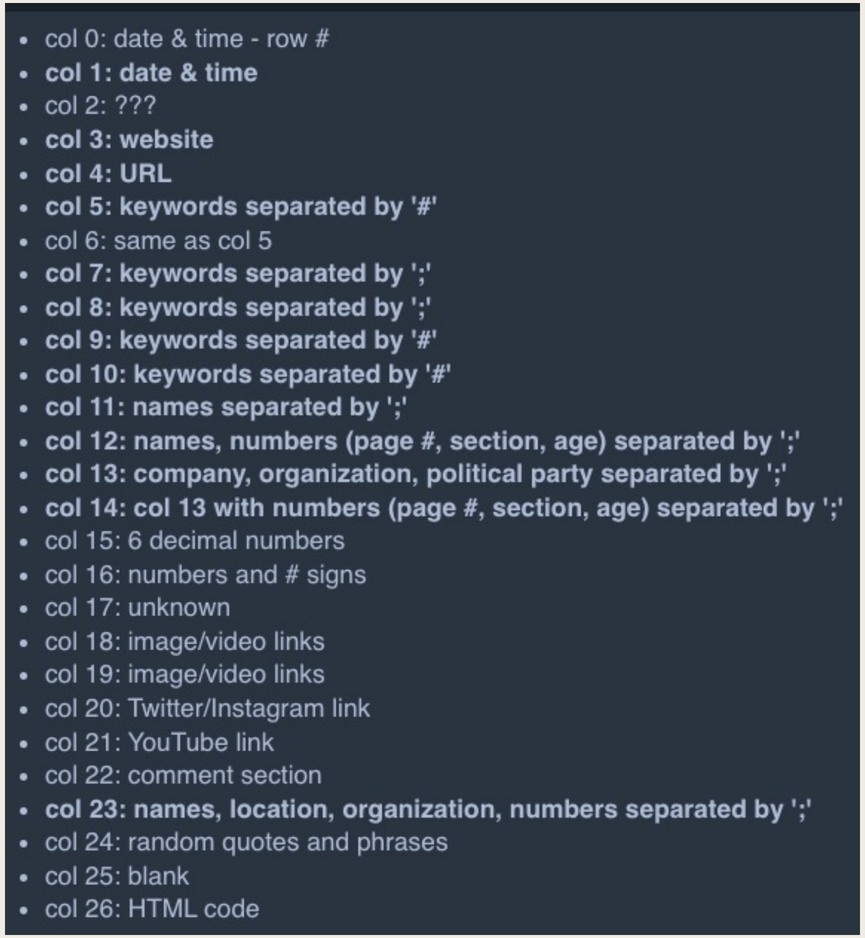
- GDELT 2.0 Database
- Online News Summary (API/Dashboard)
- Google Big Query

GDELT 2.0 Event Database



```
98455 f1e2f5ffce4f1035a8f58527930dfecb http://data.gdeltproject.org/gdeltv2/20220113011500.export.CSV.zip
134476 50bf20ee5e2819c08ba953b3761e4d03 http://data.gdeltproject.org/gdeltv2/20220113011500.mentions.CSV.zip
6381859 7fae84c30bcafa57680ad394ac0517e7 http://data.gdeltproject.org/gdeltv2/20220113011500.gkg.csv.zip
```

- Use the .gkg.csv.zip file
  - *Messy*
  - *No column names (manually labeled)*
  - *27 columns*
  - *Keywords, phrases, people, organizations, political party, etc.*
- The other ‘export’ and ’mentions’ files do provide a URL
  - *Only 5 columns*
- Run analysis on these columns and/or navigate to URL and run analysis?
- Filter for S&P 500 companies?

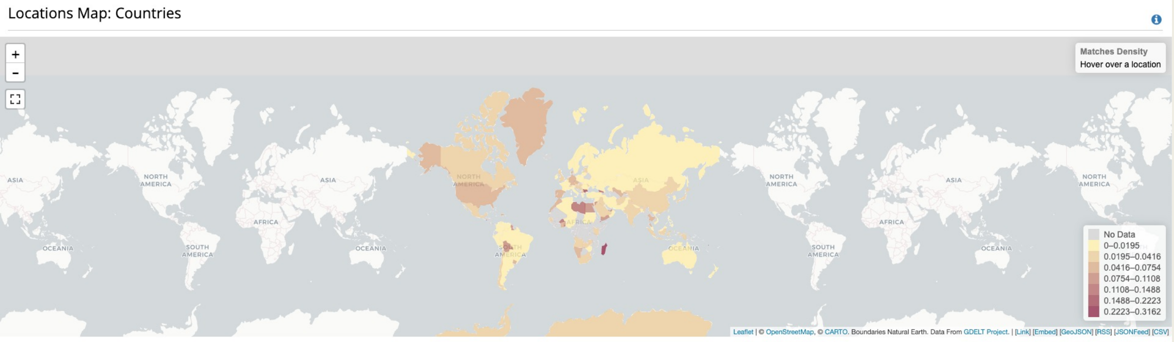


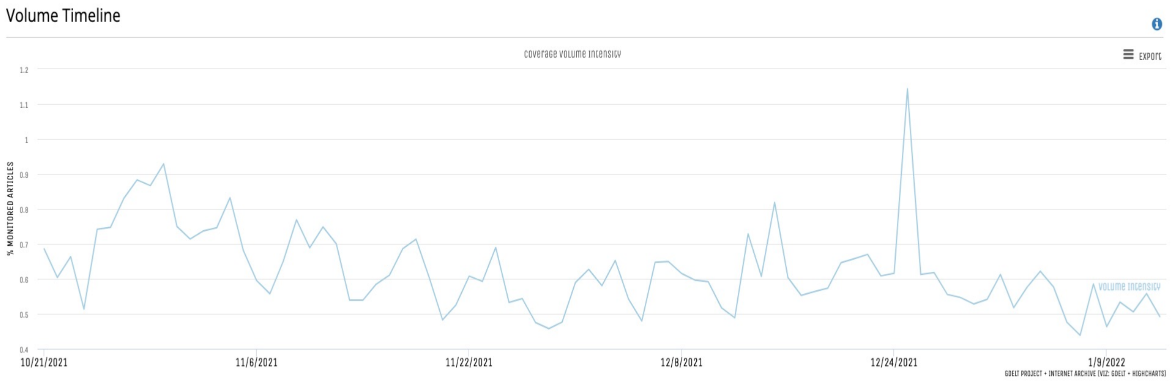
- col 0: date & time - row #
- **col 1: date & time**
- col 2: ???
- **col 3: website**
- **col 4: URL**
- **col 5: keywords separated by '#'**
- col 6: same as col 5
- **col 7: keywords separated by ';'**
- **col 8: keywords separated by ';'**
- **col 9: keywords separated by '#'**
- **col 10: keywords separated by '#'**
- **col 11: names separated by ';'**
- **col 12: names, numbers (page #, section, age) separated by ';'**
- **col 13: company, organization, political party separated by ';'**
- **col 14: col 13 with numbers (page #, section, age) separated by ';'**
- col 15: 6 decimal numbers
- col 16: numbers and # signs
- col 17: unknown
- col 18: image/video links
- col 19: image/video links
- col 20: Twitter/Instagram link
- col 21: YouTube link
- col 22: comment section
- **col 23: names, location, organization, numbers separated by ';'**
- col 24: random quotes and phrases
- col 25: blank
- col 26: HTML code

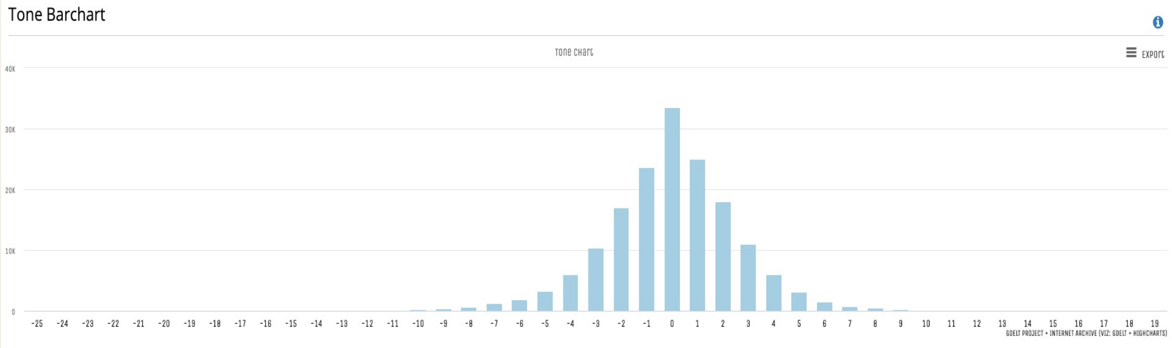
Online News Summary (API/Dashboard)

- MUST specify a keyword
  - *Can limit results*
- Can specify time period
- Provides a URL link

- Volume, tone, location charts
- Could be useful as a reference/comparison for specific companies/industries

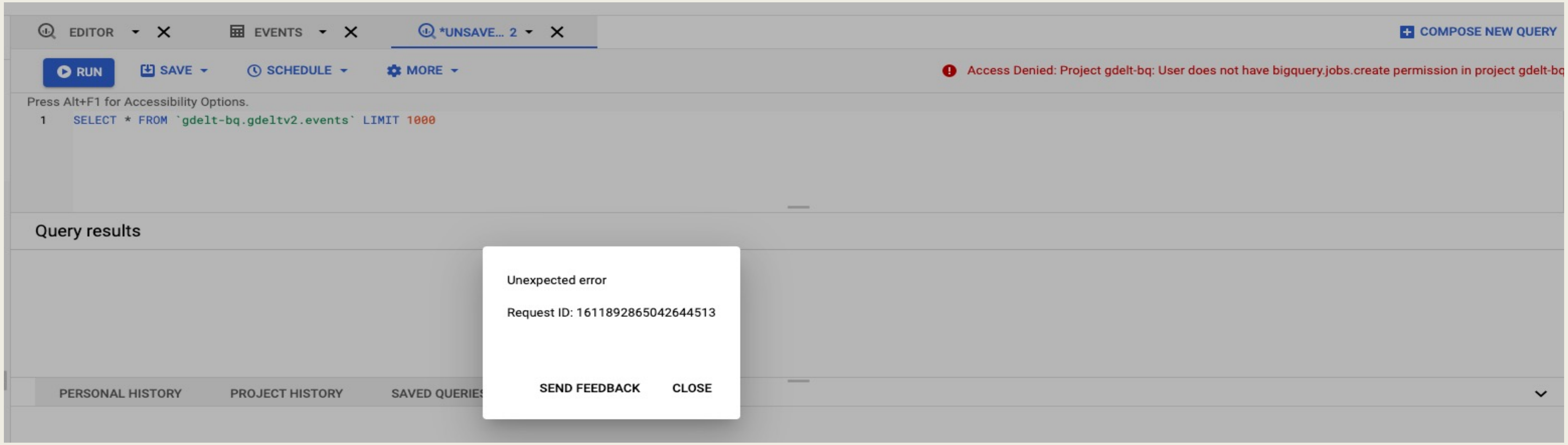






Google BigQuery

- 3 databases
  - *Events*
  - *Mentions*
  - *Global Knowledge Graph*
- Do not have permission to run queries



Access Denied: Project gdelt-bq: User does not have bigquery.jobs.create permission in project gdelt-bq

Press Alt+F1 for Accessibility Options.

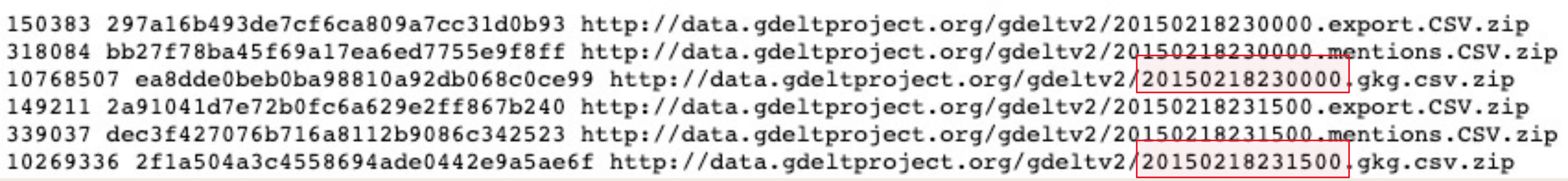
```
SELECT * FROM `gdelt-bq.gdeltv2.events` LIMIT 1000
```

Unexpected error

Request ID: 1611892865042644513

How to Collect Data?

- GDELT 2.0 Event Database
- Write a script to…
  - *Navigate to the ”Master” file (Feb 2015 – Current)*
    - Updated every 15 minutes
  - *Retrieve most recent files based on timestamp*
    - Run every day/week?
    - Append to a single master file of my own?



```
150383 297a16b493de7cf6ca809a7cc31d0b93 http://data.gdeltproject.org/gdeltv2/20150218230000.export.CSV.zip
318084 bb27f78ba45f69a17ea6ed7755e9f8ff http://data.gdeltproject.org/gdeltv2/20150218230000.mentions.CSV.zip
10768507 ea8dde0beb0ba98810a92db068c0ce99 http://data.gdeltproject.org/gdeltv2/20150218230000.gkg.csv.zip
149211 2a91041d7e72b0fc6a629e2ff867b240 http://data.gdeltproject.org/gdeltv2/20150218231500.export.CSV.zip
339037 dec3f427076b716a8112b9086c342523 http://data.gdeltproject.org/gdeltv2/20150218231500.mentions.CSV.zip
10269336 2f1a504a3c4558694ade0442e9a5ae6f http://data.gdeltproject.org/gdeltv2/20150218231500.gkg.csv.zip
```

NLP

- HuggingFace

HuggingFace

- *NLP library with various pre-trained models with hosted inference APIs*
- PEGASUS for Financial Summarization: Summarizes articles into 1-2 sentences
- distilRoberta-financial-sentiment: Produces a sentiment score for positive, neutral, and negative
- ProsusAI/finbert: Produces a sentiment score for positive, neutral, and negative
- Write a script to pass in either full article text or summary text (produced by PEGASUS) into one or both sentiment APIs to assign a rating.
- Potentially develop my own sentiment model.

FINANCIAL DATA

Financial Data

- Assess any correlation between the timestamped sentiment scores and the real-time stock performance
- Free data sources???

UPDATES

Updates

- Current work
  - *Import and parse data file to collect date and timestamp*
  - *Collect all data files dating back to ???*
  - *Initial EDA to gather volume count for specific sectors and companies*
  - *Connect HuggingFace APIs to run on keywords*
- Next steps
  - *Write a script to navigate to URL*
  - *Web scrape entire news article to run through sentiment analysis*
- Questions
  - *Readability (complexity, structure, characteristics)*
  - *GitHub*