Data Analysis on the GDELT Project

Global Database of Events, Language, and Tone

Gayathri Gude, Curry Bucher, Rohan Baishya, Benjamin King

Project Description

We strive to analyze and examine, in detail, the various trends that occur in human society. This data is being obtained from the GDELT project, in which our data archives a span of 215+ years worth of information. This data is essentially known to be a study of the **global human society.** The overarching trend genres within our dataset, include: situational awareness, influencers, risk assessment & global trends, policy reactions, and humanitarian & crisis responses.

Prior Work

- Founder of GDELT, Kalev Leetaru, won a $10,000 award for his proposal to the Wildlife Crime Tech Challenge. Leetaru plans to map trafficking in order to be able to predict events. Wildlife Crime
- Analysis of “Influencers” around the world in industry, politics, geographic location, and more. Influencers
- News Media Analysis of 72 hours of Fox coverage Vox: GDELT Fox News
- News Media Analysis of gun violence Gun Violence
- Cryptocurrency and blockchain Analysis Cryptocurrency & Blockchain

Datasets

- GDELT has a size of 2.5 **terabytes**
  - Subsets
  - Pruning
- Netflix Dataset with 480,000 users and over 100 million ratings.
  - Will the same methods used on GDELT work on the Netflix Dataset?
    - Check correlations to worldwide events
- GDELT Download:
  - Link: GDELT, downloaded on Ben King's PC
- Netflix Download:
  - Link: Netflix, not downloaded

Proposed Work

- Data Cleaning
  - Preprocessing
    - Sort data with necessary topics
    - Furtherly, only look at necessary attributes
    - Select specific sizes of data, where the proper span of information is being represented
      - **215 years worth of data is TOO much! **
- Integration
  - Cluster subsets of the dataset together to find more effective results and trends
  - Possible Netflix integration
- Mapping correlations in year by year gif or animation

Tools and Resources

- Python 3
  - Pandas and Numpy Libraries
- Matlab Plots
- Keras
- Tensorflow
- Git/GitHub
  - Link: https://github.com/cubu0178/WeLikeBigData
- Weka
  - Detailed data analytics:
    - Data Clustering and Classification
    - Visualization

Evaluation

- Statistical analysis of worldwide incidents
  - Examine clustering
  - Represent visuals through MatLab Plots and Pandas
  - Check for significance
- Finding correlations of the incidents between locations and dates
- Analyze risk assessments and crisis response
- Increase situational awareness
  - Statement of data to backup, scientifically
- Determine the effectivity of policy impacts
- Examine influences of various factors
  - Industry, geographical region, organization , etc