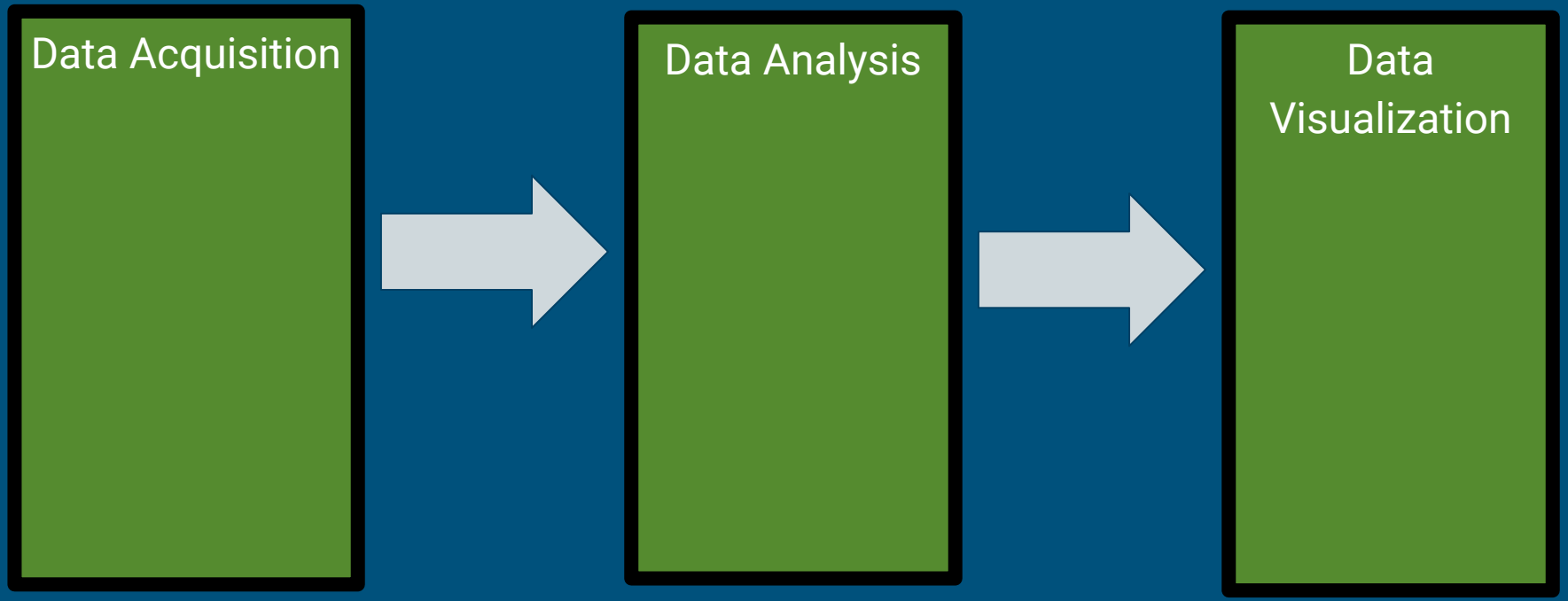
Global Event Detector

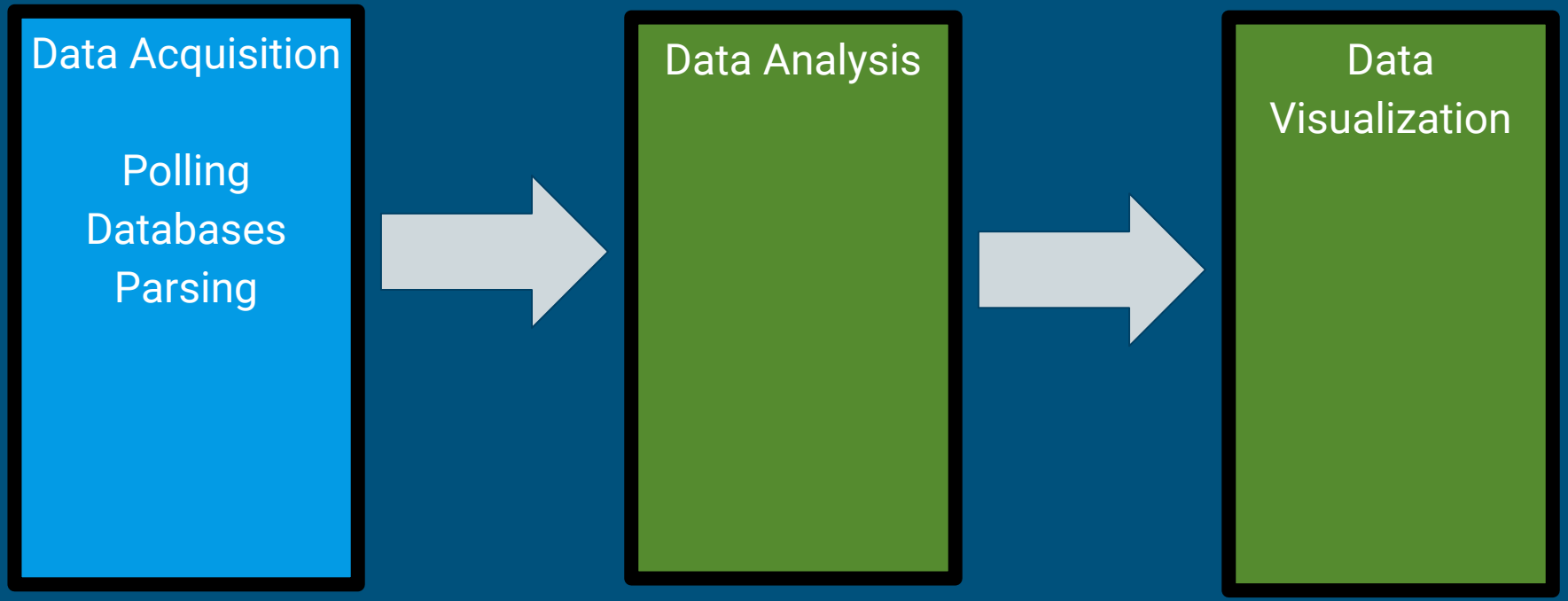
Final Project Presentation
Multimedia, Hypertext, and Information Access
4/27/2017
Blacksburg, VA 24061

Emma Manchester, Alec Masterson, Ravi Srinivasan, Harrison Grinnan, Sean Patrick Crenshaw

Project Systems



Project Systems



Polling

Collects data from Reddit WorldNews Subreddit

Driver script runs every 12 hours

Processes news articles

Stores articles in database

Databases

Stores data into raw, processed, cluster, and clusterPast tables

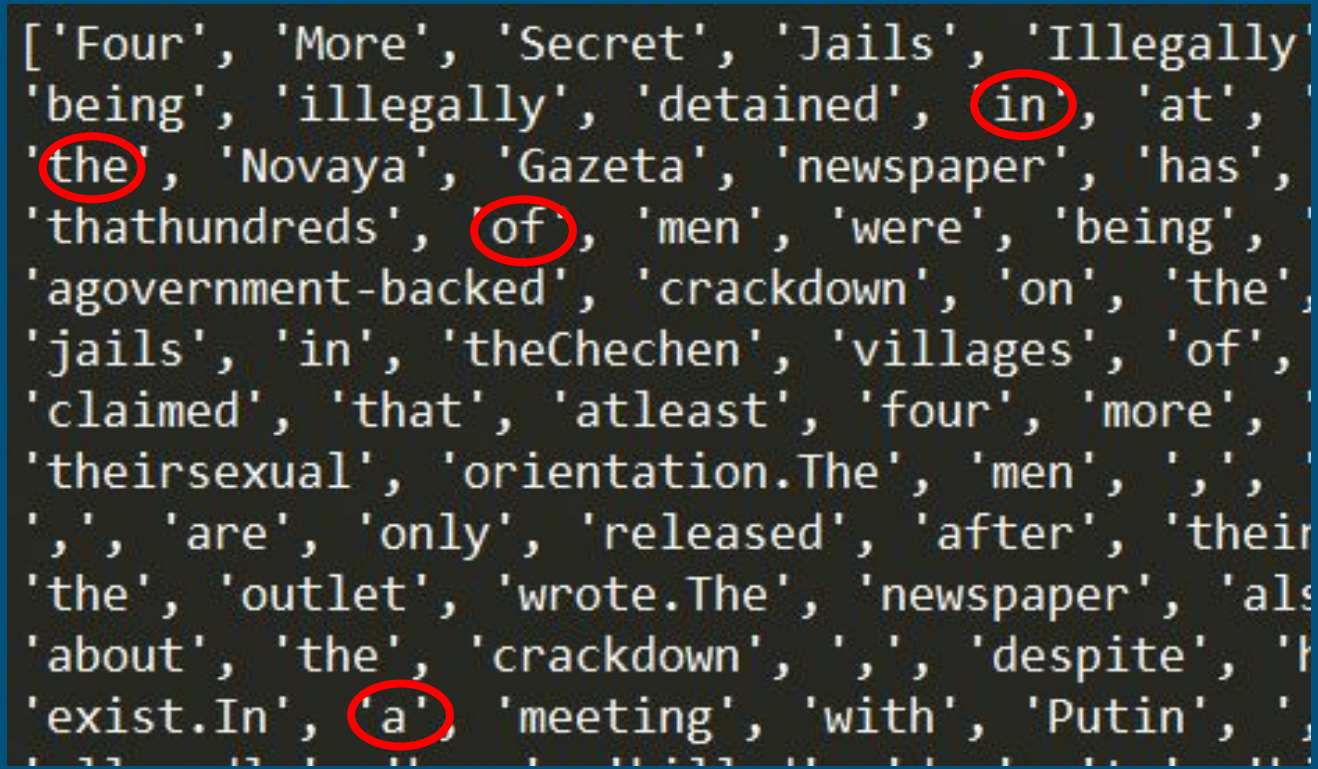
Retrieving articles more than once

Data used for visualizations

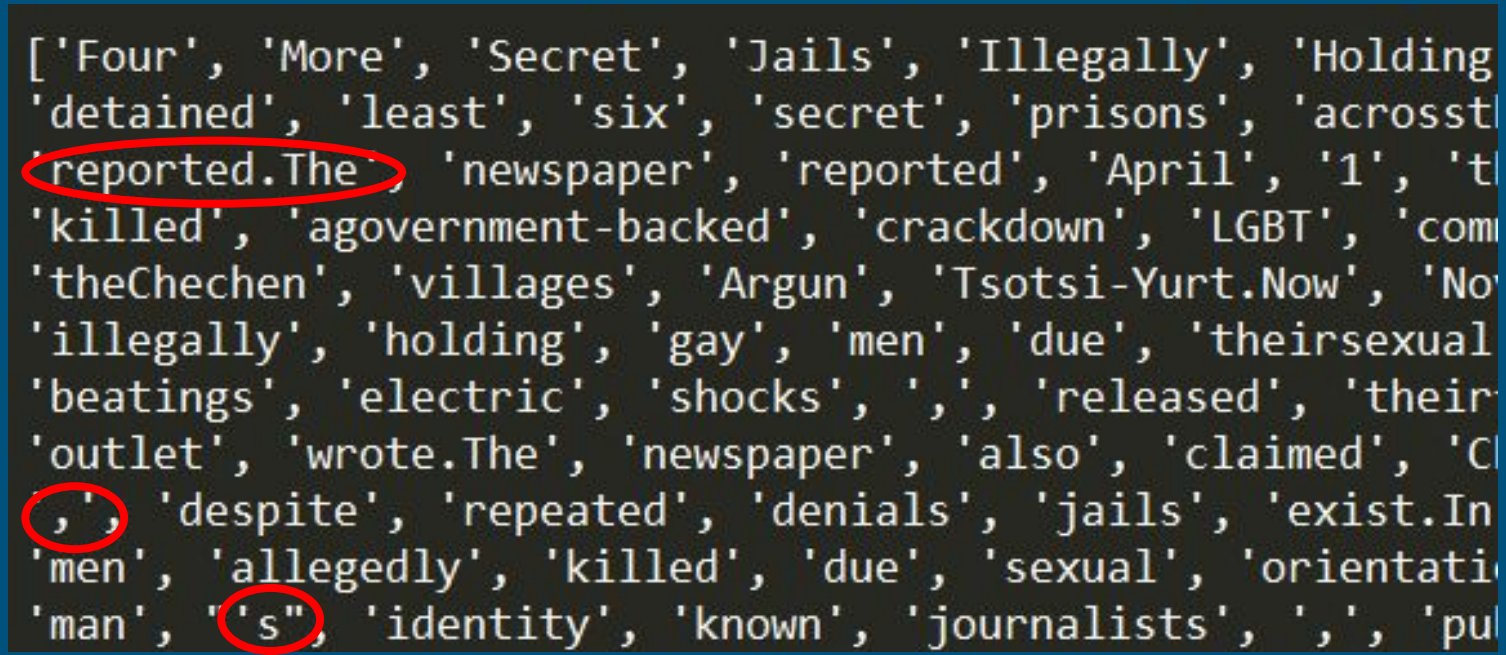
Raw Database Table

rawId	url	title	content	datePosted
63hgm5	http://www.cnn.com/2017/0…	White House: 'The clock has now run…	By Jeremy Diamond, CNNUp…	2017-04-04 18:15:32
63hj1t	http://www.reuters.com/artic…	North Korea fires projectile into sea of…	SEOUL North Korea test-fire…	2017-04-04 18:27:05
63f13y	http://www.independent.co.…	Donald Trump administration blames…	US President says the attack…	2017-04-04 12:05:27
63cpx1	http://www.bbc.com/news/w…	At least 18 people killed in suspected…	Share this withEmailFaceboo…	2017-04-04 03:33:59
63e5a5	http://www.dw.com/en/eu-to…	EU to cut gas dependency on Russia…	Israel and several EU nations…	2017-04-04 09:38:06
63cz8z	http://www.independent.co.…	Junior doctor whose Facebook post o…	Police say Rebecca Ovenden…	2017-04-04 04:57:16
63cz5i	https://www.yahoo.com/new…	Hero dog dies tackling female suicide…	A dog has died after attacking…	2017-04-04 04:56:25

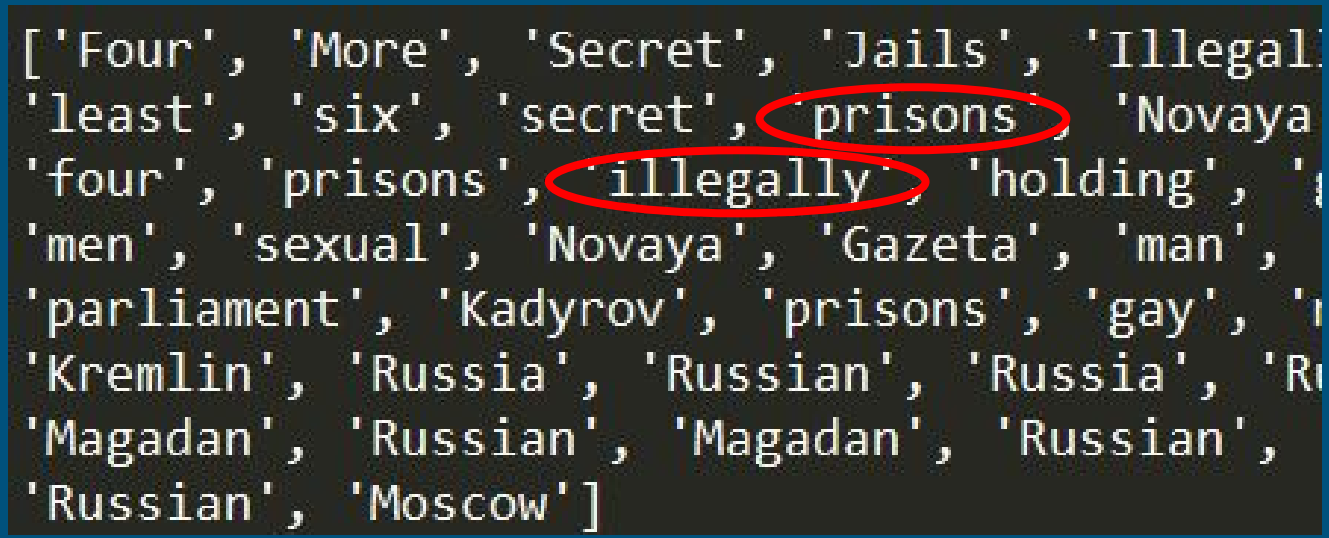
Parsing - BeautifulSoup



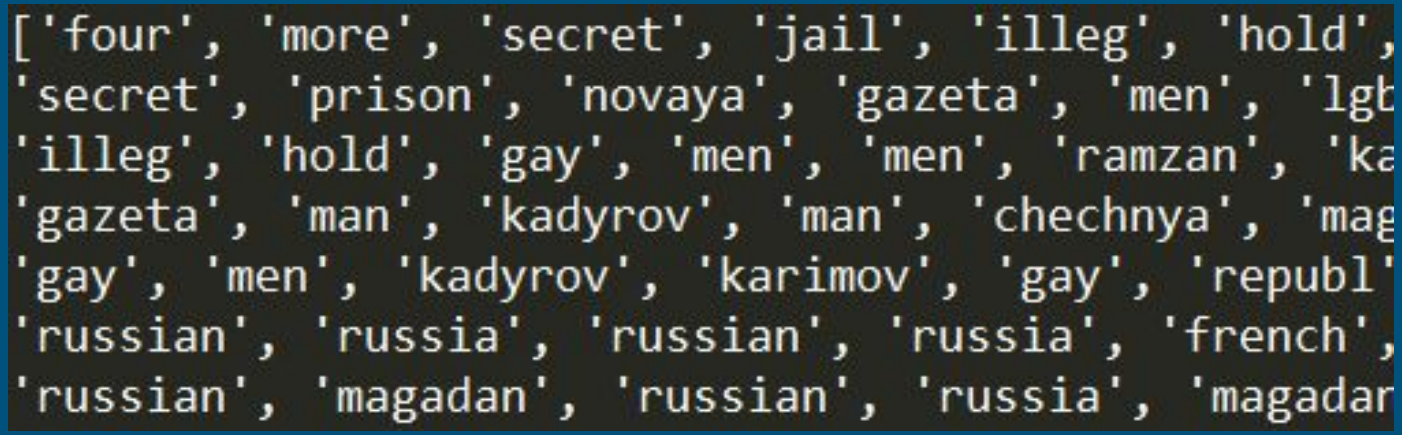
Parsing - Removing Stopwords



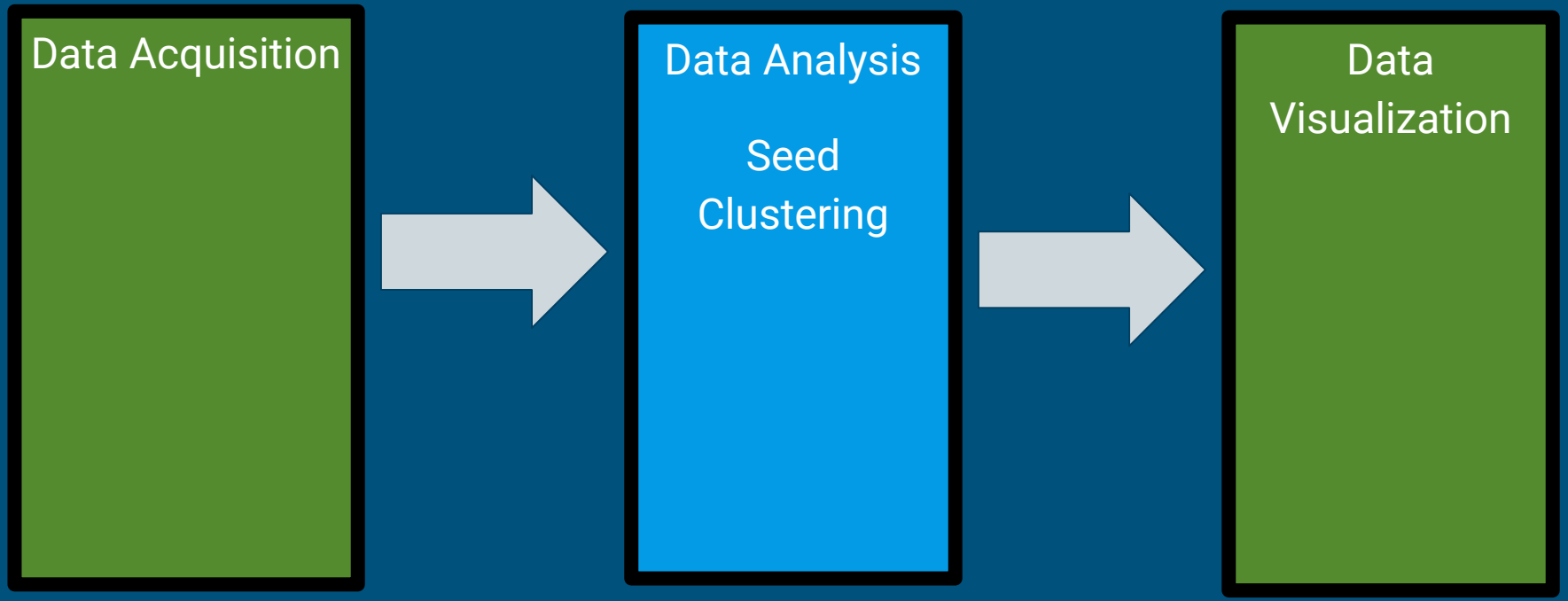
Parsing - Filtering Content



Parsing - Stemming Content



Project Systems



Stanford Named Entity Recognizer

3 Class Models used

- Location
- Person
- Organization

SNER demo

[1]

Seed Extraction

Seed - a series of entities

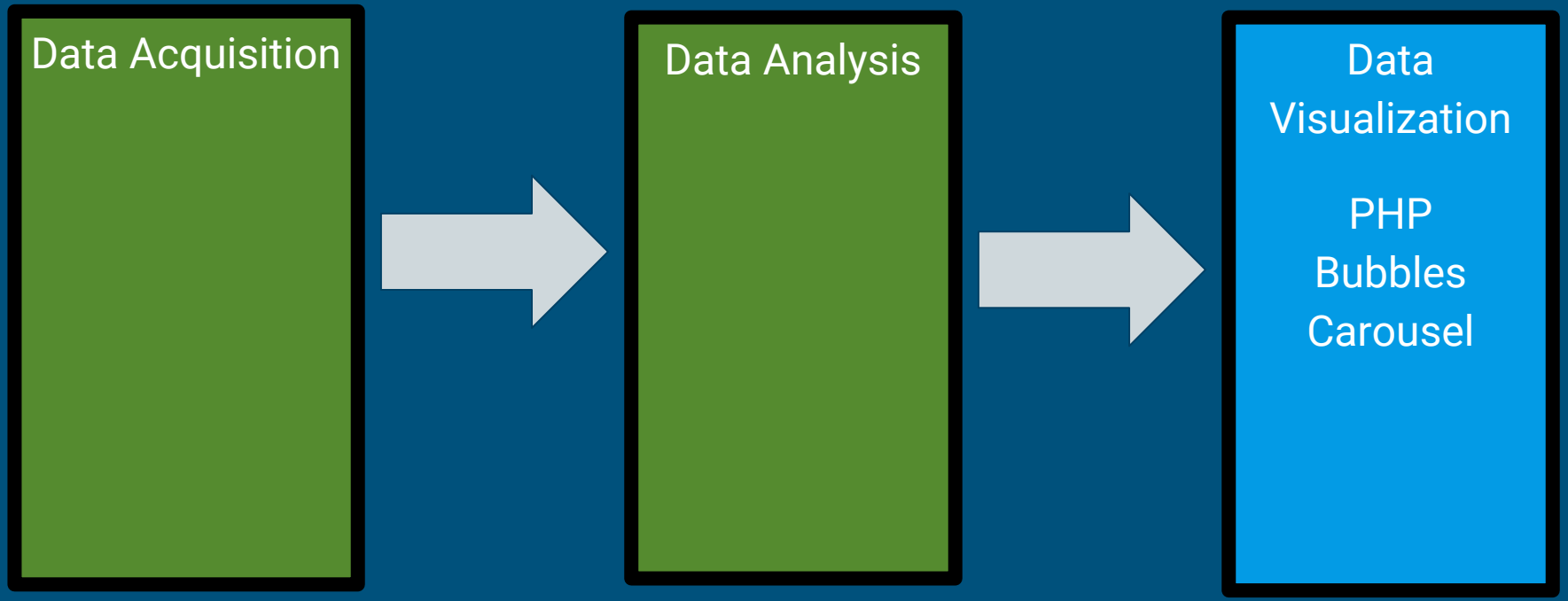
Attached to article object

Inserted into Database

Clustering

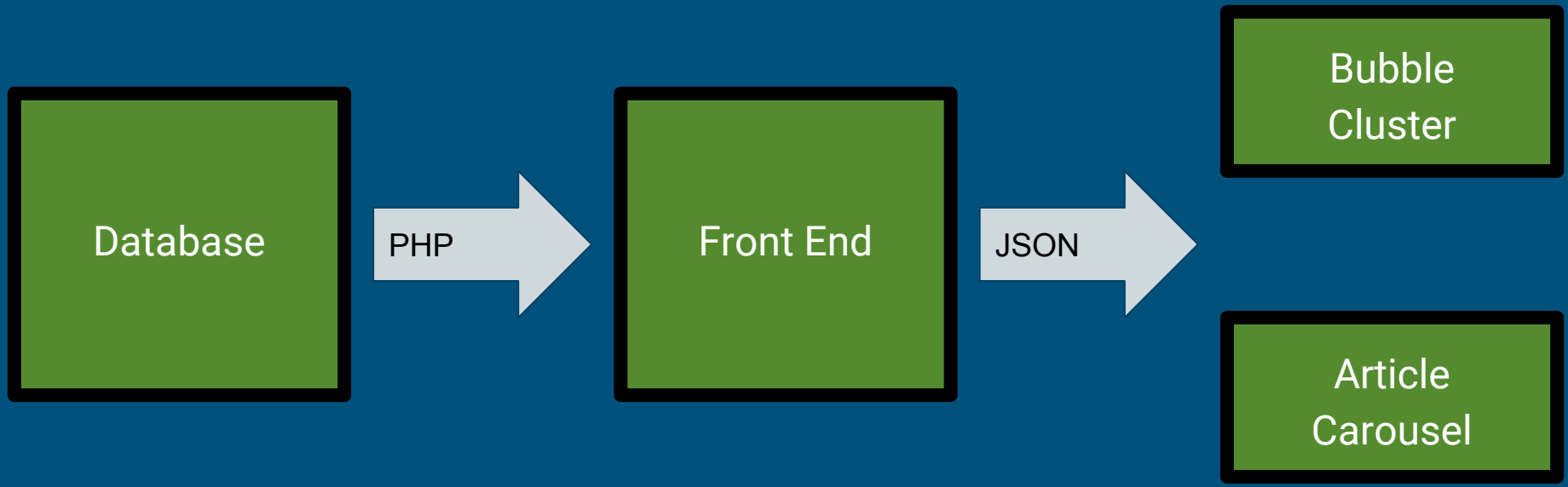
```
North Korea says ready to strike U.S. aircraft carrier
North Korea 'will test missiles weekly', senior official tells BBC
North Korea: 'US has now gone seriously mad' -- US strike group to
the North is ramping up for a sixth nuclear test.
US military considers shooting down North Korea missile tests
Putin sends troops to Russia's border with North Korea
Chances of imminent war with North Korea ‘wildly overblown,’ U.S. e
North Korea warns Australia of 'blindly toeing US line', warns of r
Kim Jong-un is starting to get 'very paranoid', UN ambassador warns
South Korea Tells Trump It's Actually Never Been a Part of China
North Korea warns it will 'wipe America off face of the Earth' afte
```

Project Systems

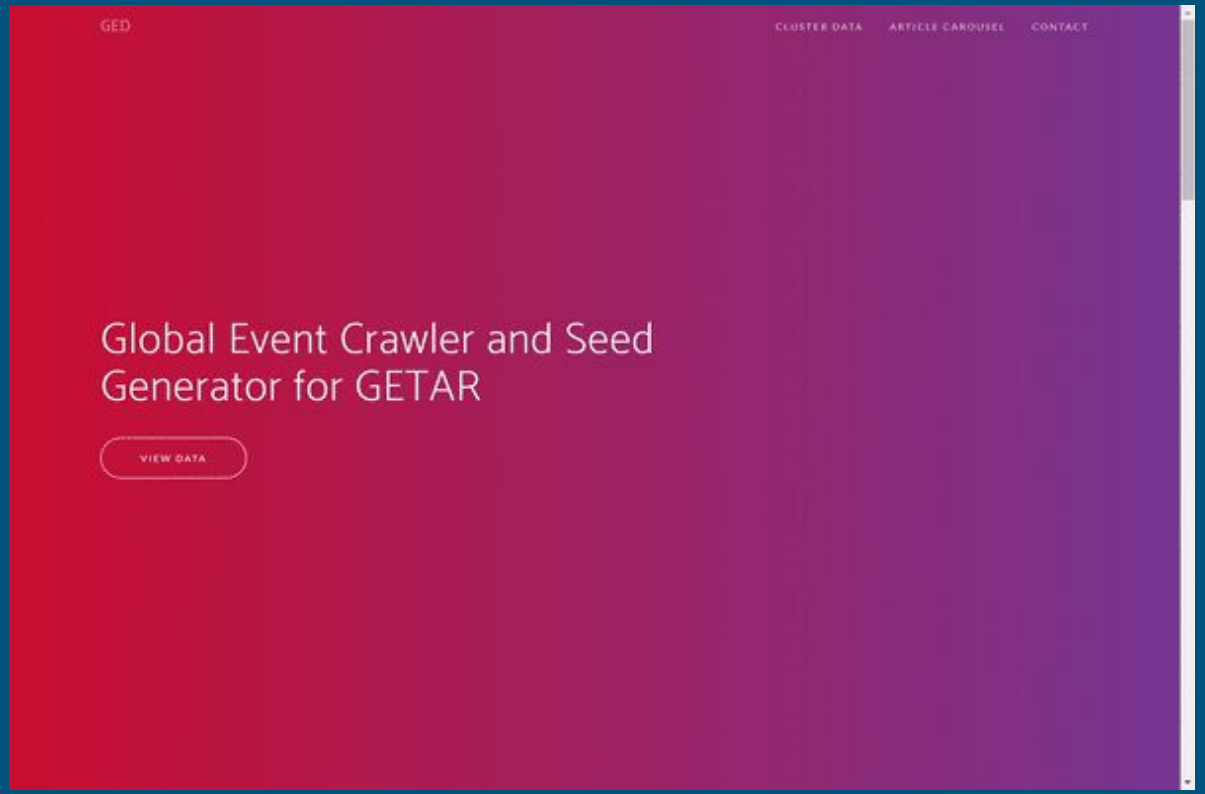


Data Visualization

- A PHP-based single-page Web-Application
- Used to display Cluster Data



Navigation


GED
CLUSTER DATA
ARTICLE CAROUSEL
CONTACT
Global Event Crawler and Seed Generator for GETAR
VIEW DATA

Article Carousel


Venezuela crisis: Teenager and woman shot dead at anti-government protests
[1]
[2]
2017-04-19 21:22:59
Venezuela crisis: Teenager and woman shot dead at anti-government protests
[3]
Seeds: Venezuela, Venezuela Nicolas, Caracas, San Cristobal, Colombian
[4]
2017-04-20 07:09:30
Venezuelan president Nicolas Maduro donates 500,000 to Trump fund despite economic woes
Seeds: Venezuelan, Nicolas Maduro, Trump, Venezuela, Venezuela Donald Trump

Cluster Data

Deliverables

Effective Clustering of Articles from Reddit

Storage of Entities in Database

Website

- Visualization of Clusters
- Article Carousel

Lessons Learned

- Ask for help early
- Project specifications change over time
- Most things have a built-in Python library - save time!
- Frequent, regularly scheduled meetings keep things moving

Questions?

References

1. Stanford University at Twitter. Retrieved April 24, 2017, from https://twitter.com/stanford
2. Collaborative Research: Global Event and Trend Archive Research (GETAR) (NSF Grant No. 1619028). URL: http://www.eventsarchive.org/sites/default/files/GETARsummaryWeb.pdf
3. Stop Words with NLTK. (2016). Retrieved March 16, 2017, from https://pythonprogramming.net/stop-words-nltk-tutorial/
4. Natural Language Toolkit. (2017, January 02). Retrieved March 16, 2017, from http://www.nltk.org/
5. Using word2vec with NLTK. (2014, December 29). Retrieved March 16, 2017, from http://streamhacker.com/2014/12/29/word2vec-nltk/
6. Rehurek, R. (2017, January 11). Gensim: topic modelling for humans. Retrieved February 17, 2017, from https://radimrehurek.com/gensim/
7. Stanford Named Entity Recognizer (NER). (2016, October 31). Retrieved March 16, 2017, from http://nlp.stanford.edu/software/CRF-NER.shtml
8. Jhlau/doc2vec. (2016, September 19). Retrieved March 16, 2017, from https://github.com/jhlau/doc2vec
9. Python Software Foundation. "20.5. urllib — Open arbitrary resources by URL." *20.5. urllib - Open arbitrary resources by URL — Python 2.7.13 documentation*. 27 Mar. 2017. Web. 28 Mar. 2017.

References

1. Rehurek, R. (2017, March 8). Models.doc2vec – Deep learning with paragraph2vec. Retrieved March 16, 2017, from https://radimrehurek.com/gensim/models/doc2vec.html
2. Richardson, L. (2015). Beautiful Soup Documentation. Retrieved March 16, 2017, from https://www.crummy.com/software/BeautifulSoup/bs4/doc/
3. Boe, B. (2016). PRAW: The Python Reddit API Wrapper. Retrieved March 16, 2017, from http://praw.readthedocs.io/en/latest/
4. Bostock, M. Data-Driven Documents. Retrieved March 17, 2017, from https://d3js.org/
5. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf
6. Mikolov, T. GoogleNews-vectors-negative300.bin.gz. Retrieved March 16, 2017, from https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
7. Schmidt, T. (2016, December 7). Named Entity Recognition with Regular Expression: NLTK. Retrieved March 16, 2017, from http://stackoverflow.com/questions/24398536/named-entity-recognition-with-regular-expression-nltk
8. Project, NLTK. "Nltk.stem package." Nltk.stem package — NLTK 3.0 documentation. 2015. Web. 16 Mar. 2017.