Event Trend Detector

Ryan Ward, Skylar Edwards, Jun Lee, Stuart Beard, Spencer Su
CS 4624 Multimedia, Hypertext, and Information Access
Instructor: Edward A. Fox
May 4, 2018
Virginia Tech, Blacksburg VA 24061

Table of Contents

1. Project Overview
2. Current Status
3. Trend Detection
4. Clustering
5. Challenges
6. What’s Left
7. Acknowledgements

Project Introduction

Collects news articles from Reddit and Google and identifies trends in frequency of mentioned entities.

Builds on a previous CS4624 project which identifies similarities (clusters) in top Reddit news stories.

Tasked with improving clustering algorithm and UI and implementing trend detection.

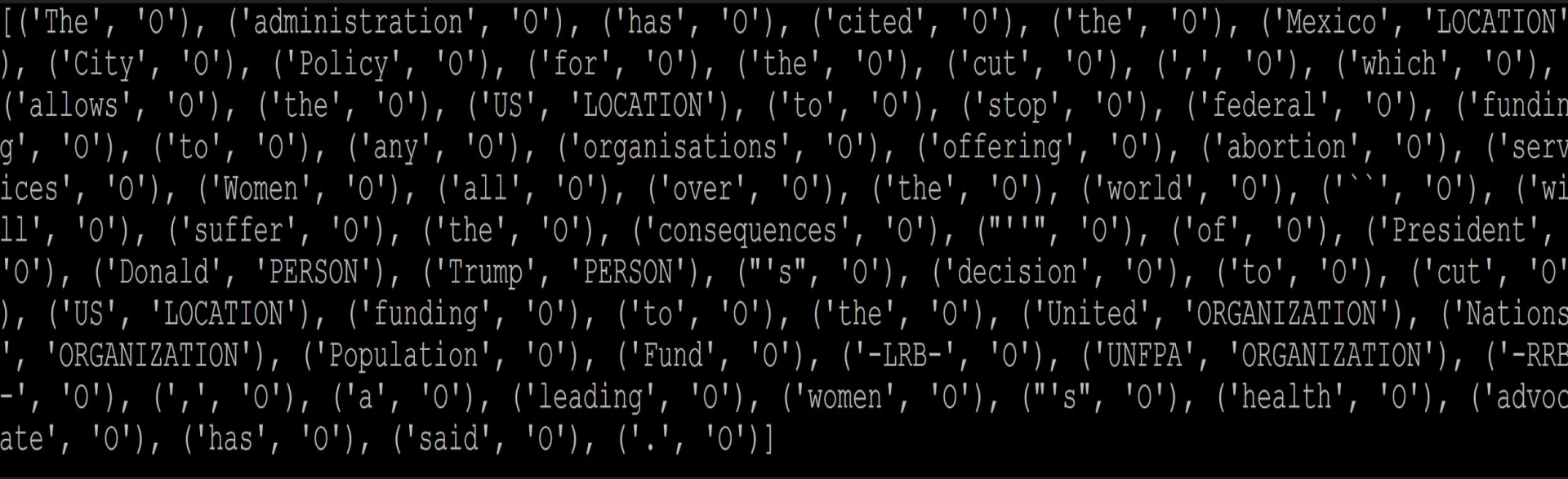
The project is viewable outside Torgersen 2030.

Work Completed

- Clustering algorithm
- Trend detection
- Google News article collection
- Updated UI

Cluster Display

Tagged Entities before cleaning



Cleaned Tagged Entities

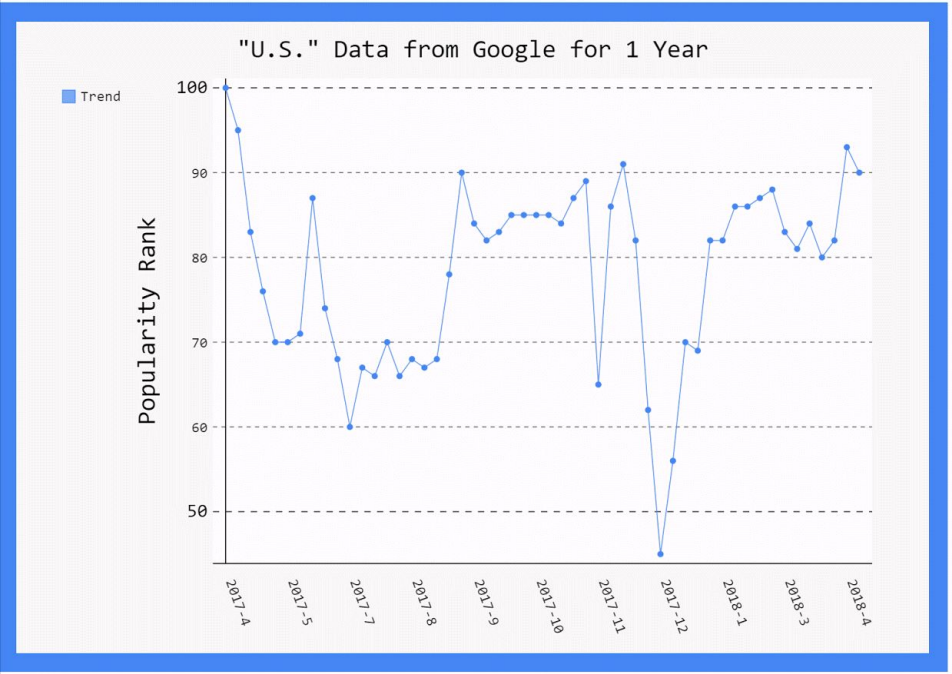
```
Counter({'LOCATION: US': 2, 'ORGANIZATION: United Nations': 1,
'PERSON: Donald Trump': 1, 'ORGANIZATION: UNFPA': 1, 'LOCATION:
 Mexico': 1})
```

Tagged Entity Database Table Example

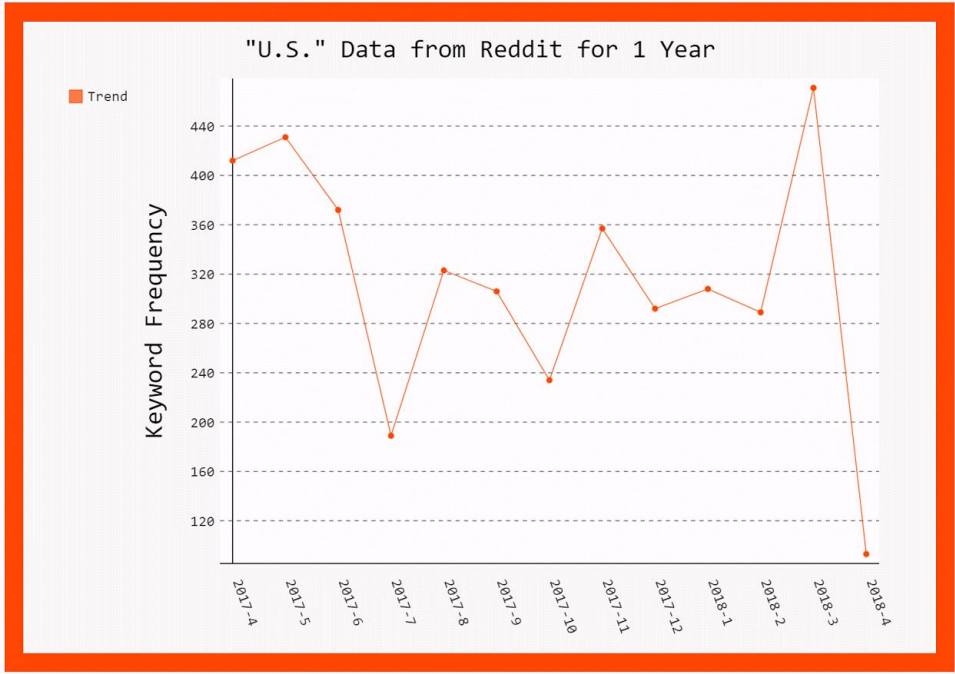
	name	tag	frequency	month	year	date
Edit Copy Delete	Facebook	LOCATION	1	3	2018	2018-03-24 19:26:14
Edit Copy Delete	Facebook	LOCATION	1	4	2018	2018-04-03 11:23:52
Edit Copy Delete	Facebook	LOCATION	1	11	2017	2017-11-05 13:13:33
Edit Copy Delete	Facebook	ORGANIZATION	120	1	2018	2018-01-01 04:44:42
Edit Copy Delete	Facebook	ORGANIZATION	109	2	2018	2018-02-03 14:24:52
Edit Copy Delete	Facebook	ORGANIZATION	523	3	2018	2018-03-05 05:21:09
Edit Copy Delete	Facebook	ORGANIZATION	42	4	2017	2017-04-04 04:57:16
Edit Copy Delete	Facebook	ORGANIZATION	124	4	2018	2018-04-01 00:22:13
Edit Copy Delete	Facebook	ORGANIZATION	74	5	2017	2017-05-01 02:42:16
Edit Copy Delete	Facebook	ORGANIZATION	57	6	2017	2017-06-01 16:13:38
Edit Copy Delete	Facebook	ORGANIZATION	25	7	2017	2017-07-04 11:51:20
Edit Copy Delete	Facebook	ORGANIZATION	31	8	2017	2017-08-01 21:19:40
Edit Copy Delete	Facebook	ORGANIZATION	105	9	2017	2017-09-05 12:12:28
Edit Copy Delete	Facebook	ORGANIZATION	94	10	2017	2017-10-02 08:06:03
Edit Copy Delete	Facebook	ORGANIZATION	175	11	2017	2017-11-01 05:00:54
Edit Copy Delete	Facebook	ORGANIZATION	59	12	2017	2017-12-01 04:24:34

Top Trends of the Week: Facebook, U.S., Russia, China, US, Trump, Brickenden, Brit

Google Trends



Reddit Trends



Clustering Implementation

- Document Similarity Matrix
  - Determines subgraph connectivity
  - Subgraphs are recalculated for dynamic similarity threshold
- Threshold filtering
  - Sizes of the subgraphs change based on different similarity threshold settings
  - Decrease threshold in each iteration to decrease the number of clusters
  - Subsequently, number of centroids also decreases
  - Goal is to create ‘the most acceptable’ number of clusters with highest similarities

Changes to algorithm

- TensorFlow -> Scikit-learn
  - Tools performs K-means clustering
  - Hardship in manipulating data for cluster representation
- Creating subgraphs with iterations
- Testing various threshold percentages (High -> Low)
- Using clique as representative
- New articles will be…
  - Added in clusters -OR-
  - Used to create new clusters

Challenges Faced

- Apache configuration/version issues
- Matching the x-axis for trend graphs
- Using pre-built libraries - sometimes not so compatible
- Deciding number of clusters for display system

Acknowledgements

Client: Liuqing Li

Supported by NSF (IIS-1619028 and 1619371)

References:

- Google Trends: https://trends.google.com/trends/story/US_cu_J7SG6GEBAADA3M_en

  https://trends.google.com/trends/explore?q=trend

- Cluster Methods:

  http://www.sthda.com/english/articles/25-cluster-analysis-in-r-practical-guide/111-types-of-clustering-methods-overview-and-quick-start-r-code/

Questions?