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

Global Event Trend Detection

- A Syria decision hasn't been made when Trump tweeted missiles will be coming
- Mexican Drug Cartels Warn Politicians Drop Out or Be Killed As Presidential Election Nears
- Israeli intelligence reportedly says Trump's Syria strike failed, didn't take out much of anything
- A sperm whale that washed up on a beach in Spain had 64 pounds of plastic and waste in its stomach
- North and South Korea reportedly set to announce official end to war
- Facebook's Tracking Of Non-Users Sparks Broader Privacy Concerns - Zuckerberg said that, for security reasons, the company collects data of people who have not signed up for Facebook.
- International chemical weapons watchdog confirms UK analysis of type of nerve agent used in Russian ex-spy poisoning
- All of Puerto Rico is without power
Tagged Entities before cleaning

Counter({'LOCATION: US': 2, 'ORGANIZATION: United Nations': 1, 'PERSON: Donald Trump': 1, 'ORGANIZATION: UNFPA': 1, 'LOCATION: Mexico': 1})
### Tagged Entity Database Table Example

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

"U.S." Data from Google for 1 Year

ReddIt Trends

"U.S." Data from Reddit for 1 Year
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:

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

- Cluster Methods:
Questions?