News/Blog Analysis for the Social Sciences

Charles B. Ward
Dept. of Computer Science
Stony Brook University
http://cs.sunysb.edu/~charles
Large-Scale Text Analysis

- Our Lydia text analysis system processes daily 1000+ English-language sources including newspapers, blogs, twitter, and more.
- We automatically detect and track tens of millions of people, places, and things in more than a terabyte of text, and provide visualizations of temporal, spatial, relational, and sentiment analyses.
Who wants to be President?
“Osama bin Laden's threats on a newly released audiotape should be taken very seriously, terrorism experts warned today.”

“There was a splendid turnaround at the Olympics in Athens Tuesday, with American swimmer Michael Phelps winning a pair of gold medals, opening the night by setting an Olympic record of 1:54.04 in the 200-meter event.”
Heatmaps

Eliot Spitzer → Disgraced Former Gov. of New York

Don Siegelman → Disgraced Former Gov. of Alabama

Juxtapositions

- We compute the most statistically significant co-occurrences between entities.

\[
\left( \frac{e^{n_a n_b} - 1}{e^{n_a n_b}} \right)^{n_a n_b} \]

<table>
<thead>
<tr>
<th>Rank</th>
<th>Newspapers</th>
<th>Count</th>
<th>Blogs</th>
<th>Count</th>
<th>Twitter</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Republican</td>
<td>47254</td>
<td>Republican</td>
<td>3236</td>
<td>GOP</td>
<td>5239</td>
</tr>
<tr>
<td>2</td>
<td>tax</td>
<td>24075</td>
<td>GOP</td>
<td>2383</td>
<td>Homosexuality</td>
<td>4211</td>
</tr>
<tr>
<td>3</td>
<td>Mitt Romney</td>
<td>20194</td>
<td>tax</td>
<td>1860</td>
<td>@RalphGarman</td>
<td>4075</td>
</tr>
<tr>
<td>4</td>
<td>GOP</td>
<td>19202</td>
<td>Mitt Romney</td>
<td>1489</td>
<td>tax</td>
<td>3061</td>
</tr>
<tr>
<td>5</td>
<td>Florida</td>
<td>12749</td>
<td>Obama</td>
<td>992</td>
<td>Republican</td>
<td>2527</td>
</tr>
<tr>
<td>6</td>
<td>Rick Perry</td>
<td>9179</td>
<td>black</td>
<td>848</td>
<td>black</td>
<td>2252</td>
</tr>
<tr>
<td>7</td>
<td>Sharon Bialek</td>
<td>9143</td>
<td>Sharon Bialek</td>
<td>846</td>
<td>Cain</td>
<td>2220</td>
</tr>
<tr>
<td>8</td>
<td>Obama</td>
<td>8646</td>
<td>Romney</td>
<td>784</td>
<td>#tcot</td>
<td>2058</td>
</tr>
<tr>
<td>9</td>
<td>Romney</td>
<td>6980</td>
<td>other</td>
<td>769</td>
<td>Mitt Romney</td>
<td>1739</td>
</tr>
<tr>
<td>10</td>
<td>taxes</td>
<td>5841</td>
<td>women</td>
<td>764</td>
<td>Obama</td>
<td>1699</td>
</tr>
</tbody>
</table>

Table 1: Juxtapositions with “Herman Cain”, October 1, 2011 – November 15, 2011
What can we do with this?

- Text-driven Forecasting:
  - Movie grosses [Zhang and Skiena, 2009]
  - Stock market [Zhang and Skiena, 2010]
  - Football spreads [Hong and Skiena, 2010]

- Social Sciences
  - Sociology of Fame [van de Rijt, et. Al, 2011]
  - Trends in gender representation [Shor, et. al 2011]
Name Ethnicity Classification

- Family and given names convey considerable information about ethnicity.
- We developed a HMM / decision tree classifier to assign people in the news to the most likely cultural/ethnic group.
- Group trends become clear when applied to the 20 million distinct names of our news data.

A. Ambekar, C. Ward, J. Mohammed, S. Male, and S. Skiena, Name-Ethnicity Classification from Open Sources, KDD 2009
Ethnic Biases in News Coverage

- **Muslim Group References 1987-2004 (New York Times)**

Frequency of coverage of entities with Hispanic names in the U.S. news, 2004-2008

Percentage of population self-reporting as Hispanic in the 2000 census. Courtesy of Wikipedia.
Hispanic Sentiment in the U.S.

We discovered a negative sentiment against Hispanics in regions where they are heavily represented, unlike other groups in our study.
Ethnic Biases in News Coverage

- (a) African
- (b) Hispanic
- (c) East Asian
- (d) Indian
- (e) Eastern European
- (f) Muslim
Future Work

- Quantifying historical significance
- Improving the quality of our sentiment analysis.
- Non-English language NLP.
- Bigger scale – “Whole Web” analysis
- Better visualization tools.
- We have lots of data and are eager for collaboration!

<table>
<thead>
<tr>
<th>Person</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jesus</td>
<td>7.53</td>
</tr>
<tr>
<td>Napoleon</td>
<td>7.18</td>
</tr>
<tr>
<td>William Shakespeare</td>
<td>7.12</td>
</tr>
<tr>
<td>Muhammad</td>
<td>7.09</td>
</tr>
<tr>
<td>Abraham Lincoln</td>
<td>6.97</td>
</tr>
<tr>
<td>George Washington</td>
<td>6.96</td>
</tr>
<tr>
<td>Adolf Hitler</td>
<td>6.73</td>
</tr>
<tr>
<td>Aristotle</td>
<td>6.68</td>
</tr>
<tr>
<td>Alexander the Great</td>
<td>6.63</td>
</tr>
<tr>
<td>Thomas Jefferson</td>
<td>6.58</td>
</tr>
<tr>
<td>Henry VIII</td>
<td>6.54</td>
</tr>
<tr>
<td>Elizabeth I</td>
<td>6.51</td>
</tr>
<tr>
<td>Julius Caesar</td>
<td>6.51</td>
</tr>
<tr>
<td>Charles Darwin</td>
<td>6.51</td>
</tr>
<tr>
<td>Karl Marx</td>
<td>6.50</td>
</tr>
<tr>
<td>Martin Luther</td>
<td>6.47</td>
</tr>
<tr>
<td>Queen Victoria</td>
<td>6.46</td>
</tr>
<tr>
<td>Joseph Stalin</td>
<td>6.44</td>
</tr>
<tr>
<td>Theodore Roosevelt</td>
<td>6.40</td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>6.40</td>
</tr>
</tbody>
</table>
Thanks!

- Questions?