Computational Linguistics PJ
-Explore Correlation between Newswires and Twitter

Client: Mohamed Magdy Farag
by Tianyu Geng, Wei Huang, Ji Wang, and Xuan Zhang

March, 6th, 2014
Blacksburg, VA
Table of Contents

- Introduction
- Solution
- Progress
Problem to Solve

Motivation: Much news, much tweets, little connection...

- Key points in news?
- Relation between news & tweets?
- Major attitudes of audience?

Objects:

1. Summarize info in news and tweets
2. Explore correlation between news & tweets
3. Mine opinions in tweets
Solution Overview

1. Fetch text from news & tweets respectively
2. Preprocess texts: stemming, stop-word…
3. Extract events from news
   
   \textit{Event: \{Topic, Named entities(who, what, where, when)\}}

4. Map tweets to events (correlation model)
5. Mine major opinions around events
Solution: Link Tweets to Events

Pre-processor
LDA
NER

Pre-processor
NER
Correlation analyzer
Progress: Text Extraction

IN

HTML Source Code

Main Content

OUT

Pre-processing (remove tags etc.)

HTML Text (No Tags)

Body Detection based on line block distribution
We ran LDA on a sample of the #Iran collection from IDEAL

- 50,000 tweets
- Feb 13, 2013 23:58:30 ~ Feb 15, 2013 00:00:03
- 4 topics

<table>
<thead>
<tr>
<th>Topic 0</th>
<th>P</th>
<th>Keyword</th>
<th>Topic 1</th>
<th>P</th>
<th>Keyword</th>
<th>Topic 2</th>
<th>P</th>
<th>Keyword</th>
<th>Topic 3</th>
<th>P</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.028</td>
<td></td>
<td>nuclear</td>
<td>0.027</td>
<td></td>
<td>iran</td>
<td>0.018</td>
<td></td>
<td>iran</td>
<td>0.054</td>
<td></td>
<td>camp</td>
</tr>
<tr>
<td>0.018</td>
<td></td>
<td>weapons</td>
<td>0.019</td>
<td></td>
<td>time</td>
<td>0.017</td>
<td></td>
<td>via</td>
<td>0.053</td>
<td></td>
<td>liberty</td>
</tr>
<tr>
<td>0.015</td>
<td></td>
<td>iranian</td>
<td>0.015</td>
<td></td>
<td>now</td>
<td>0.012</td>
<td></td>
<td>2help</td>
<td>0.020</td>
<td></td>
<td>never</td>
</tr>
<tr>
<td>0.011</td>
<td></td>
<td>stop</td>
<td>0.013</td>
<td></td>
<td>opposition</td>
<td>0.011</td>
<td></td>
<td>killed</td>
<td>0.015</td>
<td></td>
<td>martin</td>
</tr>
<tr>
<td>0.010</td>
<td></td>
<td>menlu</td>
<td>0.012</td>
<td></td>
<td>u</td>
<td>0.010</td>
<td></td>
<td>sanctions</td>
<td>0.014</td>
<td></td>
<td>kobler</td>
</tr>
<tr>
<td>0.008</td>
<td></td>
<td>iran</td>
<td>0.012</td>
<td></td>
<td>feb</td>
<td>0.009</td>
<td></td>
<td>syria</td>
<td>0.013</td>
<td></td>
<td>mr</td>
</tr>
<tr>
<td>0.008</td>
<td></td>
<td>executions</td>
<td>0.011</td>
<td></td>
<td>seeking</td>
<td>0.009</td>
<td></td>
<td>guards</td>
<td>0.012</td>
<td></td>
<td>adequate</td>
</tr>
</tbody>
</table>
Progress: News Analysis

Dataset:
1) 2762 news about “Iran Election”.
   --Only news titles used for topic modeling
2) News articles from CTRnet PJ

Tools: GibbsLDA, Stanford NLP

Event 1
- Iran
- Sanction
- Obama
- warn
- Where: Iran
- Who: Obama
- What: N/A
- When: N/A

Event 2
- UN
- urge
- human
- right
- Iran
- Where: Iran
- Who: UN
- What: N/A
- When: N/A

News Articles

Steps:
- HTML Parser
- Text Extractor
- Pre-processor
- LDA
- NER
Future Work: Opinion Mining
Appendix
1. Analysis between Tweets and News Articles
   - Fact: News providers report events earlier, but Twitter contains more details
   - Algorithm: LDA, cosine similarity, sentiment analysis

1. Summary based on Templates
   - Systems: SUMMARIST, Artequakt, etc.
   - Topic signature is used for selecting summarizing sentences
   - Using Apple Pie Parser, GATE and WordNet for knowledge extraction
The Characteristics of HTML


Reference


