Computational Linguistics

Team Hurricane

Senior Capstone CS 4984
Virginia Polytechnic Institute and State University, Blacksburg, VA
December 12 2014
Team Members: Nick Crowder, Szu-Kai “Andy” Hsu, Will Mecklenburg, Jeff Morris, David Nguyen
Computational Linguistics

Team Hurricane

Senior Capstone CS 4984
Virginia Polytechnic Institute and State University, Blacksburg, VA
December 12 2014
Team Members: Nick Crowder, Szu-Kai “Andy” Hsu, Will Mecklenburg, Jeff Morris, David Nguyen
Contents

- Introduction
- Major topics
- Hurricanes
- Project Management
- Cleaning
- Summarization of Corpora
  - Early Efforts
  - Intermediate Efforts
  - Final Efforts
- Conclusion
- Moving Forward
Introduction
Major Topics

- NLTK + NLP tools
- Python
- Big Data + Hadoop
- Project Management
- Mahout
- LDA
- K-means Clustering
- Regular Expressions
- Templates
- Grammars
Hurricanes

- Attributes of Hurricanes
- YourSmall (Typhoon Haiyan)
- YourBig (Hurricane Sandy)
Project Management

- Difficult Schedules
- Largest Group
- Version Control (Drive vs. Git)
- Trello + GroupMe
# Cleaning

## File Deletion:

<table>
<thead>
<tr>
<th></th>
<th>Start</th>
<th>File Sizes</th>
<th>Duplicates</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>ourBig</td>
<td>77,000</td>
<td>30,000</td>
<td>30,000</td>
<td>2,000</td>
</tr>
<tr>
<td>ourSmall</td>
<td>2,000</td>
<td>1,500</td>
<td>112</td>
<td>112</td>
</tr>
</tbody>
</table>

- **Runtime Filtering**
  - Unique stopword list
  - Restricted the allowed length of the word (3 < length < 21)
Summarizing

Early Efforts:

- Wordnet
- Most frequent words
- Most frequent nouns
  - Basic trainer
  - POSTagging with trigrams and backoff
- Ngrams
Summarizing

Intermediate Efforts:

- Classifier, Flood files as negative
- NER (Stanford NER vs NLTK)
- Topics using Mahout
- Clustering using Kmeans
## K-means Results

**Top Terms:**

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>oceanic</td>
<td>7.50</td>
</tr>
<tr>
<td>surface</td>
<td>7.50</td>
</tr>
<tr>
<td>minute</td>
<td>6.69</td>
</tr>
<tr>
<td>atmospheric</td>
<td>6.69</td>
</tr>
<tr>
<td>speeds</td>
<td>6.54</td>
</tr>
<tr>
<td>administration</td>
<td>6.49</td>
</tr>
<tr>
<td>kph</td>
<td>5.45</td>
</tr>
<tr>
<td>national</td>
<td>5.36</td>
</tr>
</tbody>
</table>
Final Efforts:

- Using pre-written templates
- Cascading RegEx's
- Generating a template from a grammar

The storm, Typhoon Haiyan, hit in Tacloban on November 2013. Haiyan was a Category 5. Additionally, the typhoon formed in the Pacific. Furthermore, Haiyan caused 2,000 to be missing. Typhoon Haiyan had a size of 600 kilometres wide. Typhoon Haiyan caused 15 inches of rain. For more information, search for Typhoon Haiyan.
<table>
<thead>
<tr>
<th>Rule</th>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>summary</td>
<td>intro details close</td>
</tr>
<tr>
<td>intro</td>
<td>&quot;boilerplate required details&quot;</td>
</tr>
<tr>
<td>details</td>
<td>detail nextdetail</td>
</tr>
<tr>
<td>detail</td>
<td>start &quot;detail sentence&quot;</td>
</tr>
<tr>
<td>nextdetail</td>
<td>trans detail nextdetail</td>
</tr>
<tr>
<td>close</td>
<td>&quot;boilerplate close details&quot;</td>
</tr>
<tr>
<td>trans</td>
<td>&quot;Additionally&quot;, &quot;Also&quot;, &quot;Furthermore&quot;, &quot;&quot;</td>
</tr>
<tr>
<td>start</td>
<td>&quot;The typhoon&quot;, &quot;Typhoon Haiyan&quot;, &quot;Haiyan&quot;</td>
</tr>
</tbody>
</table>
Conclusion

- Cleaning is critical
- Early efforts are effective for summarizing
- Further efforts more about details and presentation
- Lots of potential improvement, for example try passing output from k-means clustering to more specialized set of regular expressions
Moving Forward

- Pursue the field of text generation
- Go back and rerun early effort summarization techniques on cleanest data set
- Use smarter extraction techniques on future corpora
Contact Information

If you have any questions or comments contact us:

Nicholas Crowder: crowdern@vt.edu
Szu-Kai “Andy” Hsu: skhsu91@vt.edu
Will Mecklenburg: willmeck@vt.edu
Jeff Morris: jeffm14@vt.edu
David Nguyen: dnguy06@vt.edu
Computational Linguistics

Team Hurricane

Senior Capstone CS 4984
Virginia Polytechnic Institute and State University, Blacksburg, VA
December 12 2014
Team Members: Nick Crowder, Szu-Kai “Andy” Hsu, Will Mecklenburg, Jeff Morris, David Nguyen