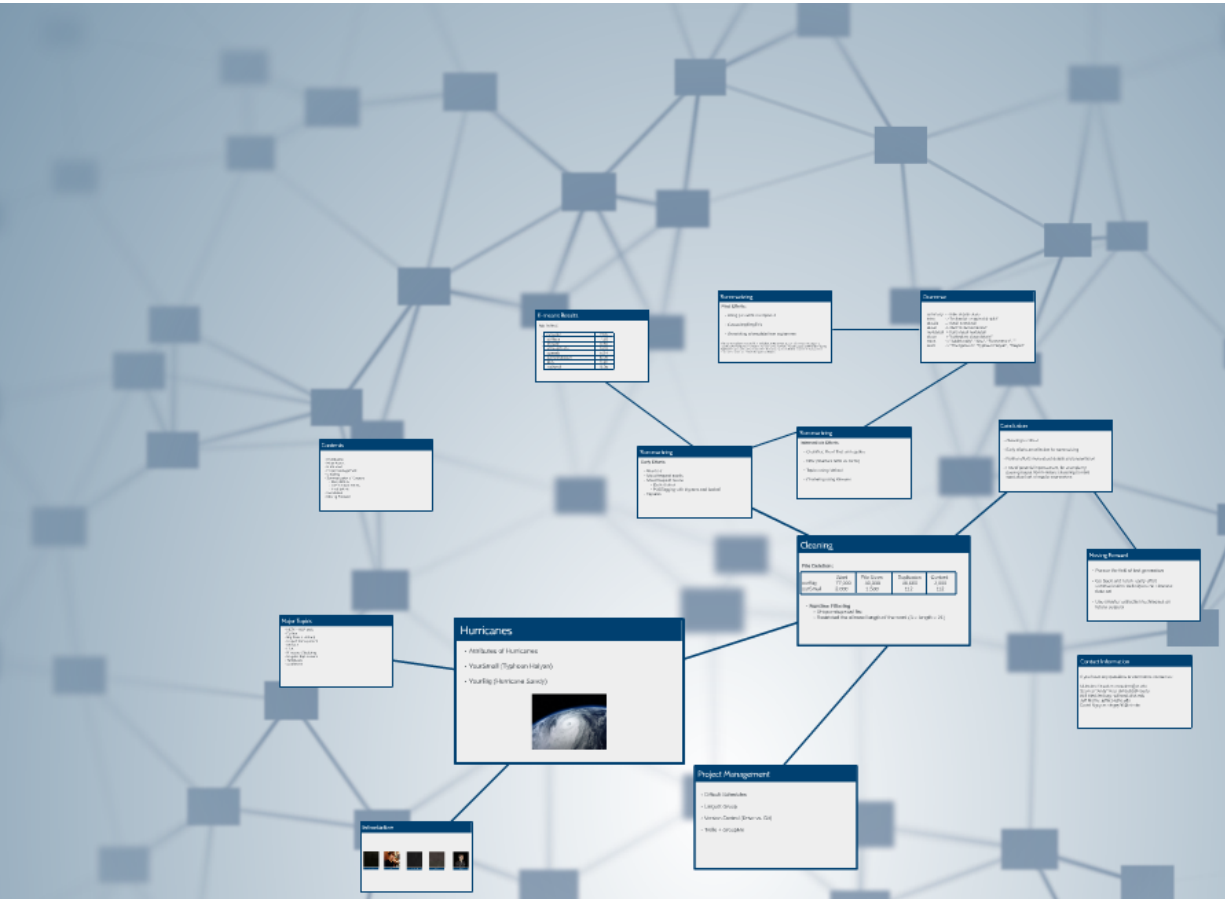
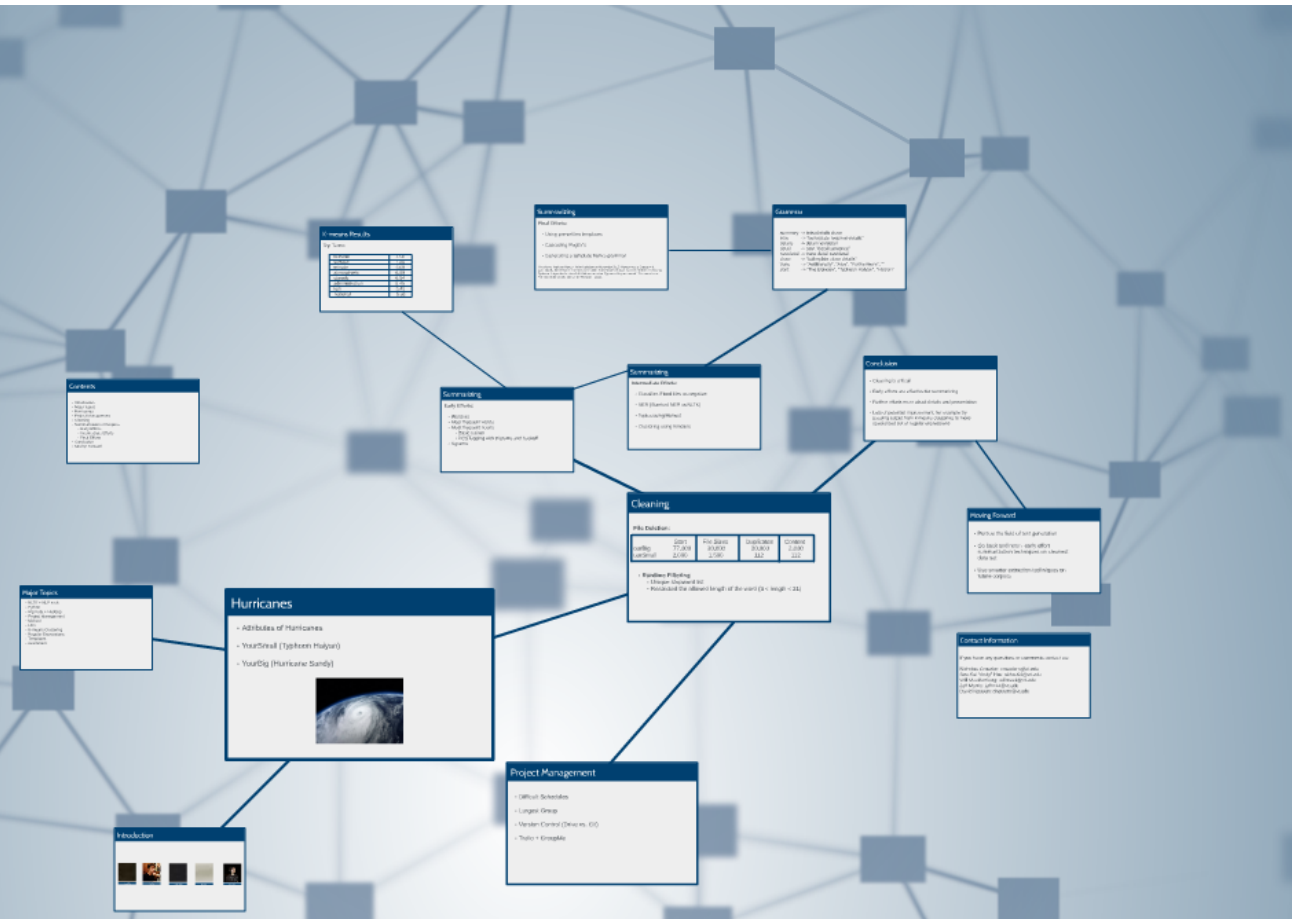


Computational Linguistics

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

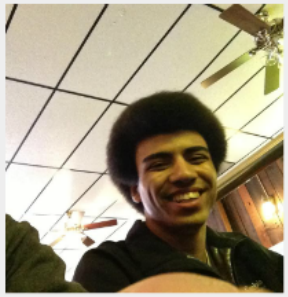
Team Hurricane

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

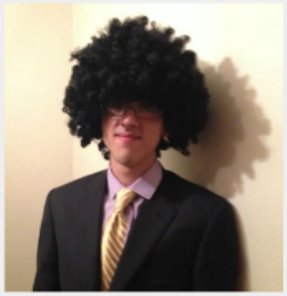
- Introduction
- Major topics
- Hurricanes
- Project Management
- Cleaning
- Summarization of Corpora
  - Early Efforts
  - Intermediate Efforts
  - Final Efforts
- Conclusion
- Moving Forward

Introduction



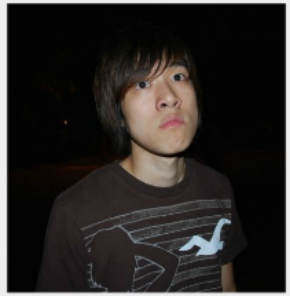
Nick Crowder

Andy Hsu



Will Mecklenburg

Jeff Morris



David Nguyen

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:

	Start	File Sizes	Duplicates	Content
ourBig	77,000	30,000	30,000	2,000
ourSmall	2,000	1,500	112	112

- **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:

oceanic	7.50
surface	7.50
minute	6.69
atmospheric	6.69
speeds	6.54
administration	6.49
kph	5.45
national	5.36

Summarizing

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.

Grammar

```
summary   -> intro details close
intro     -> "boilerplate required details"
details   -> detail nextdetail
detail    -> start "detail sentence"
nextdetail -> trans detail nextdetail
close     -> "boilerplate close details"
trans     -> "Additionally", "Also", "Furthermore", ""
start     -> "The typhoon", "Typhoon Haiyan", "Haiyan"
```

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

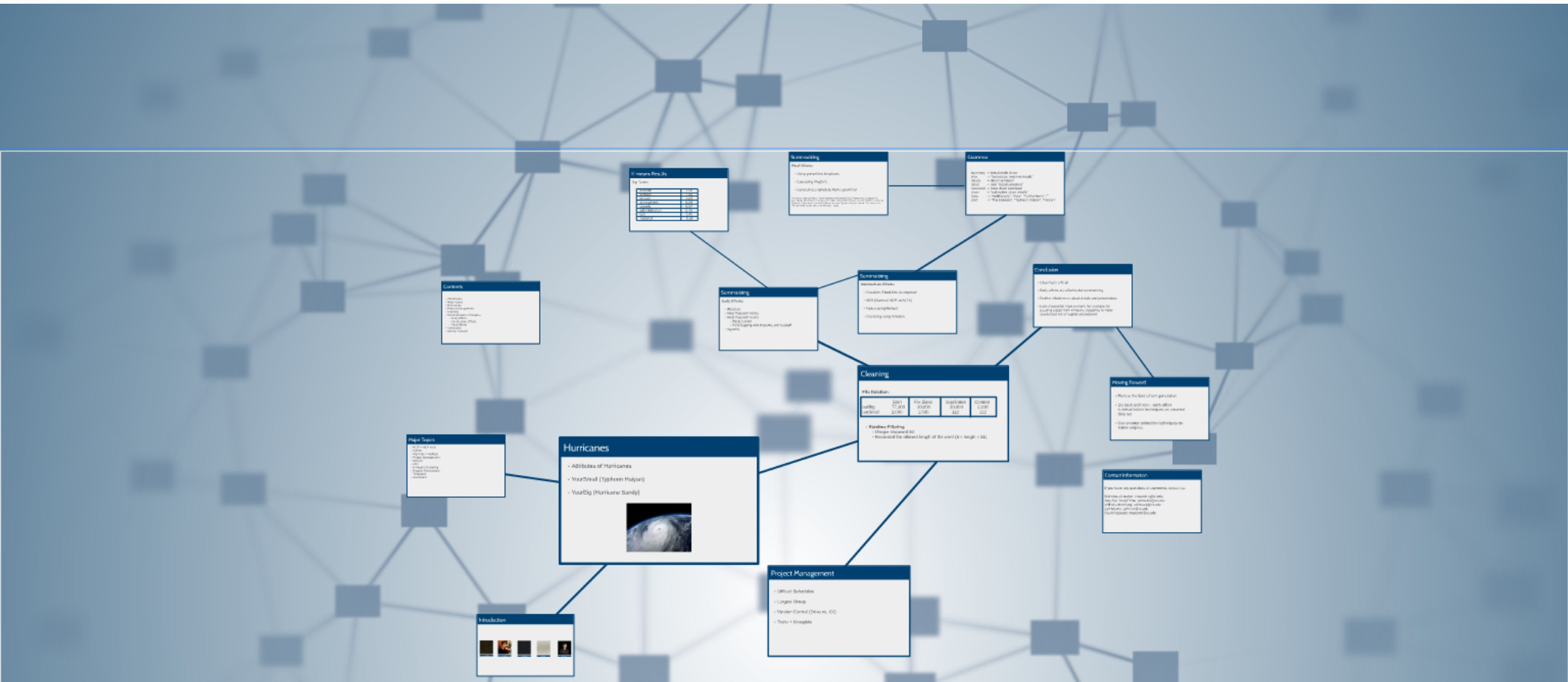
Moving Forward

- 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