Floods

Joe Acanfora, Myron Su, David Keimig and Marc Evangelista

Introduction

1. Objective
2. Discussion of Corpora
3. Final Results
4. Tools we used for cleaning the Data
5. Tools we used for language processing
6. Tools we did not use
7. What we learned
8. Conclusion

Objective

Generate summaries of flooding events based on collections of news articles.

Flood Data

- ClassEvent - Islip_Flood
  - 11 Files
- YourSmall - China_Flood
  - 537 files
- YourBig - Pakistan_Flood
  - 20,416 files

Issues with having unclean data

U9 Results

In June 2011 a flood spanning 9.94 miles caused by heavy rain affected the yangtze river in China. The total rainfall was 170.0 millimeters and the total cost of damages was 760 million dollars. The flood killed 255 people, left 87 injured, and approximately 4 million people were affected. In addition 168 people are still missing. The cities of Wuhan Beijing and Lancing were affected most by flooding, in the provinces of Zhejiang Hubei and Hunan. Finally nearly all of the flood damage occurred in the state of China.

U9 Results

In August 2010 a flood spanning 600 miles caused by heavy monsoon affected the indus river in Pakistan. The total rainfall was 200.0 millimeters and the total cost of damages was 250 million dollars. The flood killed 3000 people, left 809 injured, and approximately 15 million people were affected. In addition 1300 people are still missing.

The cities of Nasirabad Badheen and Irvine were affected most by flooding, in the provinces of Sindh Mandalay and Punjab. Finally nearly all of the flood damage occurred in the state of Pakistan.

Tools We Used...

Cleaning the data

1. Removed files less than 5kb
2. Machine Learning
    a. **DecisionTreeClassifier = 90%**
    b. NaiveBayesClassifier = 80%
    c. MaxEntropyClassifier= 73%
    d. SklearnClassifier = 92%
3. Picked top Paragraphs from Corpora
    a. Used WordNet on 20 words
    b. Tokenized by Paragraph
    c. Picked paragraphs with at least 2 WordNet results

Cleaned Data

Collection	Pre-clean size	Post-clean size	% bytes reduced
YourSmall	2.0 MiB	288 KiB	86%
YourBig	136.7 MiB	3.7 MiB	98%

Collection	Pre-clean #docs	Post-clean #docs
YourSmall	537	1
YourBig	20,416	1

Classifier

Machine learning

-Using decision tree

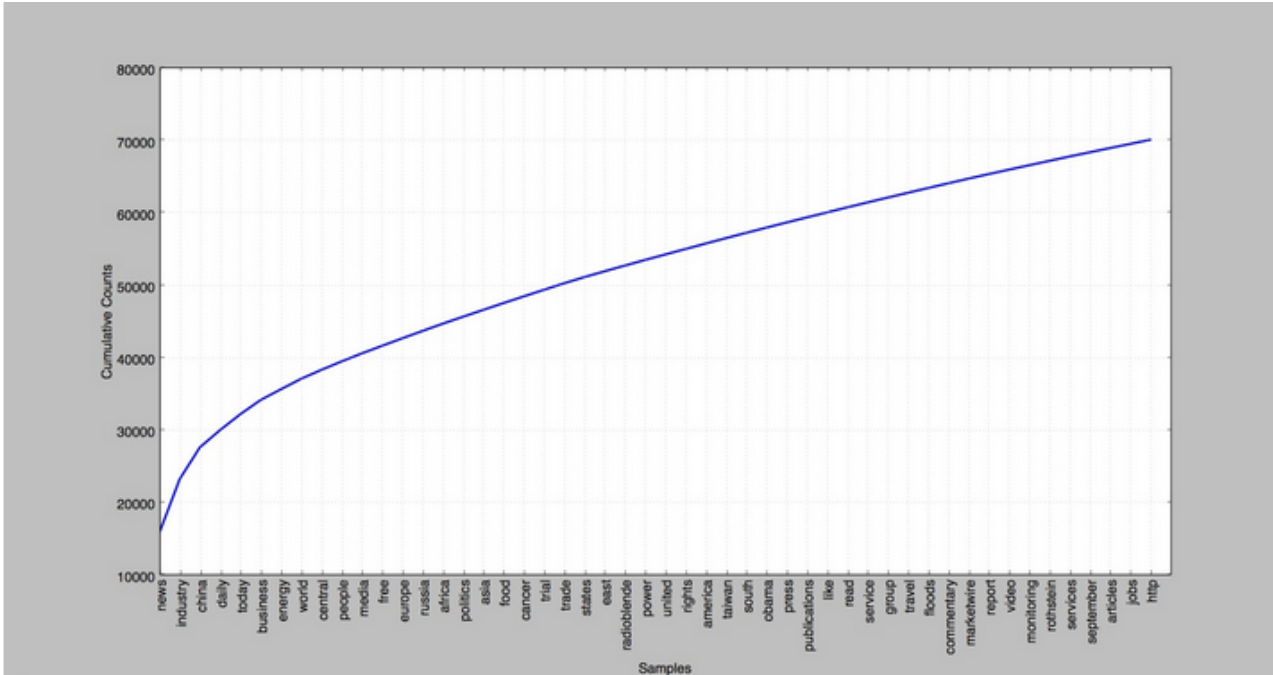
	Accurate	**Inaccurate**	**Percentage**
YourSmall	90	10	90%
YourBig	83	17	83%

Frequency Analysis

We used a frequency analysis for:

-Cleaning data

-Getting summary



POS Tagging

Used the POS tagger for our regular expression “cause” string

Checked to see if the cause string returned by the regular expression contained some subject (noun)

In June 2011 a flood spanning 9.94 miles caused by heavy rain affected the yangtze river in China.

Regex

- best used on cleaned data
- works well because:
  - patterns prevalent in news reports
  - corpora about the same event use same words

Regex examples

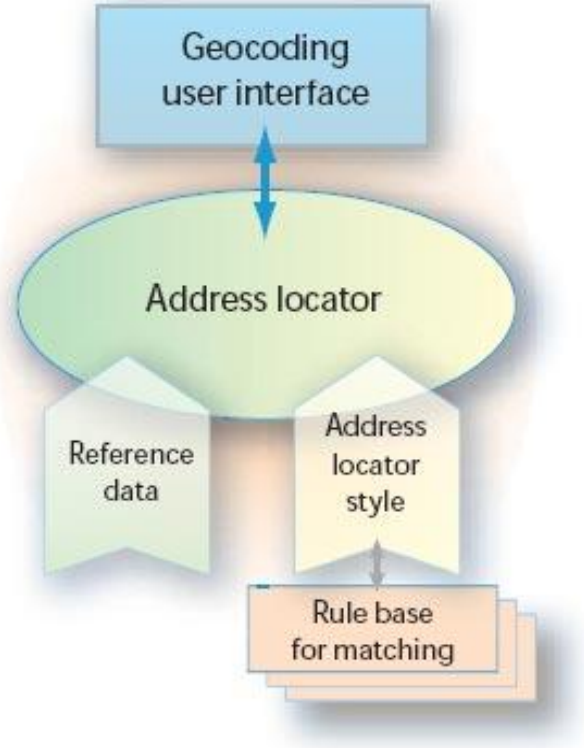
- "affected by ____", "result of ____", "caused by _____", "by ____"
- months
- ___ people killed/missing/injured
- ____ (m|tr)illions dollars
- ____ miles/km

NER Tagger

Rather than using the NER tagger for tagging locations we decided to use a google maps api...

Contextualizing Locations

Google Geocoder API with pygeocoder python package



Tools We Did Not Use...

Bi-Grams & N-grams

- not used extensively
- bigrams were good, but already in YourWords
- operations we used were based on single words
- did help with regex

Useful bigrams	YourWords
flash flooding heavy rains inches rain rain fell	**flood** **rain** overflow dam storm severe water damage submerge washed collapsed river discharge downpour **flash** sweep torrential runoff

Useful bigrams	Some regexes
flash flooding heavy rains inches rain rain fell	(\d+.\d+\smillimeters)|(\d+.\d+**smm**))|(\d+.\d+\s(**inches**|**inch**) **due\sto**(\s[A-Za-z]{3,}){1,3}|**result\sof**(\s[A-Za-z]{3,}){1,3}|**caused\sby**(\s[A-Za-z]{3,}){1,3}|by\s([A-Za-z]{4,}){1,2})|heavy\s([A-Za-z]{3,}

Clustering & Mahout

- documents similar enough that clusters would be indistinguishable
- wanted data from all good sources
- clean data was good enough

Chunking

- finds multitoken sequences
- already knew the data we were searching for
  - brainstormed our own chunks, which was good enough
  - would be helpful if we didn’t know patterns
- regular expressions alone did the job well on clean data

Conclusion

U9 Results

In June 2011 a flood spanning 9.94 miles caused by heavy rain affected the yangtze river in China. The total rainfall was 170.0 millimeters and the total cost of damages was 760 million dollars. The flood killed 255 people, left 87 injured, and approximately 4 million people were affected. In addition 168 people are still missing. The cities of Wuhan Beijing and Lancing were affected most by flooding, in the provinces of Zhejiang Hubei and Hunan. Finally nearly all of the flood damage occurred in the state of China.

U9 Results

In August 2010 a flood spanning 600 miles caused by heavy monsoon affected the indus river in Pakistan. The total rainfall was 200.0 millimeters and the total cost of damages was 250 million dollars. The flood killed 3000 people, left 809 injured, and approximately 15 million people were affected. In addition 1300 people are still missing.

The cities of Nasirabad Badheen and Irvine were affected most by flooding, in the provinces of Sindh Mandalay and Punjab. Finally nearly all of the flood damage occurred in the state of Pakistan.

Wrap Up - Challenges

- New Technologies
  - Hadoop - Map/Reduce
  - NLTK Library
- Group Logistics
  - Times
  - Work Distribution

Wrap Up - Strengths

- Python
- LaTeX
- Synergy
- Willing to learn

Conclusion - Improvements

- Under Estimates
  - Deaths
  - Damages
- Spatial Locations
  - Mean Distances

Citations and Thanks

Dr. Fox

GTA Tarek

GTA Mohamed

GTA Xuan

Questions