Computational Linguistic Analysis of Earthquake Collections

Team F: Kevin Kokal, Kwamina Orleans-Pobee, Chris Wakeley, Reed Bialousz

CS4984 - 12/08/2014
Virginia Tech, Blacksburg, VA 24061

Outline

- Problem Statement
- Acclimation (Units 1-3)
- Exploration (Units 4-7)
- Culmination (Units 8-9)
- Results
- Lessons Learned

Problem Statement

- Objective: Construct a meaningful summary of an earthquake event using a combination of computational linguistic techniques.

Collection	**Location**	**Year**	**# of Files**
Small	Lushan, China	2013	454
Big	Virginia, U.S.A.	2011	4599

Acclimation: Overview

1. Familiarize with NLTK/Python
2. Examine Value of Word Frequency
3. Refine
    a. Lemmatization
    b. Word Stemming
4. Familiarize with Hadoop/MapReduce
5. Examine Part of Speech Tagging
6. Utilize Cluster

Acclimation

- Unit 1: Word Frequencies with ClassEvent
  - With restrictions, decent results
  - Collocations improved results

```
Best small set of word pairs:
Long Island; New York; Weather Service; National Weather;
Bay Shore; heavy rains; Suffolk County; rain fell;
Associated Press
```

Acclimation (cont.)

- Unit 2: Improve on Unit 1 Solution
  - Utilize WordNet
  - Lemmatize
  - Explore Hadoop/MapReduce

- Noise in Small Collection
  - After Simple Cleaning, Results Improved

Acclimation (cont.)

- Unit 3: Incorporate Part of Speech Tagging
  - Constructed a Trigram Part of Speech Tagger
    - Back-Off Model
  - Lemmatized Top Nouns and a Few Top Verbs
  - Mixed Results
    - More Cleaning Required (See Below)

Small: ['earthquake', '2717'], ['quake', '1965'], ['sichuan', '1473']

Big: [' earthquake', ' 36399'], [' comment', ' 31146'], [' video', ' 31131']

Exploration: Overview

1. Clean Collections
2. Incorporate Named Entity Recognition
3. Summarize by Topics Using Mahout/LDA
4. Explore Clustering to Summarize by Indicative Sentences

Exploration

- Unit 4: Focus on Cleaning Collections
  - Using Bigram Features
  - Internet Stop Words List
  - Decision Tree Classifier (> 90% Accuracy)
  - Rerunning Past Units Produced Improved Results

The powerful earthquake that rocked China has claimed 165 lives so far and injured more than 6,700, according to state media reports.
The quake that southwest China on Saturday wrecked homes and triggered landslides in an area devastated by a major tremor in 2008.

initialComments:true! pubdate:08/25/2011 06:54 EDT! commentPeriod:3! commentEndDate: 8/28/11 6:54 EDT! currentDate:9/5/11 8:51 EDT! allowComments:false!

Exploration (cont.)

- Unit 5: Examining Named Entities
  - Locations = Most Important
  - Stanford NER > NLTK.ne_chunk
  - Unhelpful results for “People” for both technologies
  - Solr Queries: More Locations = Better Results

```
Small: [('China’, 450), ('Sikkim', 244), ('Sichuan', 171)]

Big: [('Virginia', 46), ('East', 15), ('Washington', 14), ('Coast', 12)]
```

Exploration (cont.)

- Unit 6: Summarizing by Topic
  - Explored LDA using Mahout on Hadoop Cluster
  - A Few Meaningful Results:

```
{sichuan:0.034,china:0.033,strikes:0.025,deadly:0.021,saturday:0.014,cnn:
0.011,quake:0.010,april:0.009,chinese:0.009,sunday:0.008}
```

- Solr Results for Top 5 Topics: Great for Small, Mediocre for Big
  - Big Collection Needed Cleaning

Exploration (cont.)

- Unit 7: Sentence Clustering
  - Cleaned Big Collection
  - Experimented with Number of Clusters
  - Generally not meaningful, with Exceptions

```
180 people have died, at least another 24 are missing.
```

Culmination: Overview

1. Construct a Template of Relevant Information for Earthquakes
2. Fill in Template for Small and Big Collections
3. Construct English Report Using Filled in Template

Culmination

- Unit 8: Construct and Fill “Earthquake Template”
  - Researched Earthquakes to Construct Template
  - Used Regular Expressions to Fill
  - Corpus Language Influences Results (i.e. Magnitude)
  - Mode = Most Accurate
  - Great Results

Culmination (cont.)

Big Corpus Earthquake Analysis:

Date: August 23, 2011
Location: Virginia
Magnitude: 5.8
Epicenter: Louisa
Deaths: 140
Time: 1:51 p.m.

Small Corpus Earthquake Analysis:

Date: April 20, 2013
Location: Lushan
Magnitude: 6.6
Epicenter: Lushan
Deaths: 186
Time: 8:02 a.m.

Culmination (cont.)

- Unit 9: Using Filled Templates to Construct English Reports
  - Improved Template: Added booleans for aftershocks, tsunamis, etc.
  - Created a report structure for any earthquake

Results

YourBig:
On 23 August, 2011 at 1:51 p.m., a 5.8 magnitude earthquake struck Virginia. The epicenter of the quake was located at Louisa. There were aftershocks that followed the earthquake and no tsunami was caused by the earthquake. There were no reports of landslides due to this earthquake. A total of 140 deaths occurred.

YourSmall:
On 20 April, 2013 at 8:02 a.m., a 6.6 magnitude earthquake struck Lushan. The epicenter of the quake was located at Lushan. There were aftershocks that followed the earthquake and no tsunami was caused by the earthquake. There were reports of landslides due to this earthquake. A total of 186 deaths occurred.

Lessons Learned

- Cleanliness is Key
- Familiarize -> Explore -> Practice -> Utilize
- Parallelize
- Document and Reuse
- Implement Using Multiple Strategies

Acknowledgements

- CS4984 Instructors
  - Dr. Edward Alan Fox (fox@vt.edu)
  - Tarek Kanan (tarekk@vt.edu)
  - Xuan Zhang (xuancs@vt.edu)
  - Mohamed Magdy (mmagdy@vt.edu)
- CS4984 Classmates
- Sponsors
  - NSF DUE-1141209
  - IIS-1319578

References

- Bird, Steven, and Ewan Klein. Natural Language Processing with Python. Beijing: O'Reilly, 2009. Print.
- See “References” Page of CS4984 Wiki
- http://www.nltk.org/
- https://docs.python.org/2/

Questions

http://www.cardiomyopathy.
org/assets/images/news/question.jpg