Through the Fire and Flames

Michael Zamani, Hayden Lee, Michael Trujillo, Jordan Plahn

CS 4984: Computational Linguistics
Virginia Tech, Blacksburg, VA
December 8, 2014

Outline

- Goals
- Driving Question
- Corpus Details
- Summary of Results
- Lessons learned
- Deliverables
- Acknowledgements

Team Goals

- Natural Language Processing
- Hadoop
- Solving open ended problems

Driving Question

What is the best summary that can be automatically generated for a document collection about a fire?

Corpus Details

- Small
  - Texas Wildfire
  - ~19,500 files
- Large
  - Brazil Nightclub Fire
  - ~690,000 files
- Duplicates
- File composition

Corpus	Avg. # lines per file pre-cleanup	Avg. # lines per file post-cleanup	% Duplicate Files
Small	1001	14	78
Large	3846	56	85

Cleaning the Collections

- Many duplicate documents due to the web crawler matching “forest park contact” sites.
- Each document contained sentences that had been scraped from irrelevant sections such as navigation menus and advertisements.
- These duplicate documents and sections were deleted.

Summary of Results

Feature Set Extraction

- Frequency with experimental filters
  - Stopwords
  - Word length
- Synonyms via NLTK WordNet
- Part of speech
  - Nouns and verbs
- N-grams

NLTK Most Informative Features

Feature	Odds (True : False)
flames	21.3 : 1.0
fires	20.1 : 1.0
burned	20.0 : 1.0
firefighting	16.4 : 1.0
acres	15.3 : 1.0
drought	15.2 : 1.0
evacuate	12.8 : 1.0
evacuated	12.6 : 1.0
burn	12.3 : 1.0
wildfires	12.0 : 1.0

Classification

- Small: 17.1% of ~7,200 classified positive.
- Large: 9.87% of ~30,000 classified positive.
- Chose ME classifier over DT, because training set was relatively small compared to our overall corpus size, to mitigate risks of over-fitting training data using the DT classifier.

5-Fold Validation Results

Fold #	Maximum Entropy	Decision Tree	Naive Bayes
1	0.78	0.80	0.68
2	0.98	0.95	0.95
3	0.87	0.88	0.87
4	0.97	0.93	0.90
5	0.77	0.80	0.77
Overall	~0.874	~0.872	~0.834

Topic Summarization

- Gensim
  - Latent Dirichlet Allocation (LDA)
  - Discover semantic structure of document

Corpus	Topics
Small	**Topic 1 :** 0.006*texas + 0.005*news + 0.005*fire + 0.005*2011 + 0.003*ago + 0.003*us + 0.003*new + 0.003*people + 0.002*1 + 0.002*said
	Topic 2 : 0.017*fire + 0.006*2011 + 0.005*september + 0.004*texas + 0.004*news + 0.003*us + 0.003*2010 + 0.003*firefighter + 0.003*firefighters + 0.003*new
Large	**Topic 1 :** 0.018*fire + 0.016*nightclub + 0.010*brazil + 0.007*people + 0.006*santa + 0.005*club + 0.005*said + 0.004*sign + 0.004*news + 0.004*maria,
	Topic 2 : 0.010*fire + 0.006*brazil + 0.006*sign + 0.006*people + 0.006*nightclub + 0.005*news + 0.004*santa + 0.004*youtube + 0.004*club + 0.003*ago

k-means Clustering

- Apache Mahout
- Grouping of object sets similar to each other based on a feature
- Results:

Word	k-means distance
state	3.629
counties	4.476
bastrop	2.630
erupted	4.476
wildfires	3.629
largest	3.965

Extracting and Refining Results

- Used Regular Expressions to extract data for each attribute.
- Narrowed data to top 10 most frequent results for each attribute.
- Used parts of speech tagging to ignore inappropriate results (such as a verb when an adjective was expected)
- Built a basic grammatical model to adjust our template based on the best result (e.g., inserting ‘ended up’ if a verb ending in ‘ing’ was present, or ‘there were’ if a number followed by a noun was present)
- Conjugated present tense verbs to past tense in selected result (using the Python Pattern library)

Results

Small: In September 2011, there was a fire started by a historic drought in Bastrop. This fire, fueled by hot temperatures, strong winds, grew to encompass 33,000 acres, burned for several days, and ended up killing four. 400 firefighters responded to the wildfire. 700 homes were affected as a result of the fire.

Large: In January 2013 there was a fire started by indoor fireworks in Santa Maria. This fire, fueled by ignited foam, grew to the size of the building, engulfed the club and ended up killing 309. Firefighters worked to douse a fire at the Kiss Club. One exit was made unavailable for a period of time. Compared to previous fires in the city it was fast-moving.

Lessons Learned

- Spend the time learning the theory BEFORE attempting the problem
- Designate a domain ‘expert’ for each topic covered.
- Test algorithms (i.e., MapReduce) locally on small data sets before scaling to full collections.
- Garbage in = Garbage out

Acknowledgements

We would like to thank:

- Dr. Fox
- Tarek Kanan
- Xuan Zhang
- Mohamed Magdy
- National Science Foundation (for providing the grant for this class)

References

Rehurek, Radim. "Introduction." Gensim: Topic Modelling for Humans. Gensim, 17 Nov. 2014. Web. 08 Dec. 2014.

"Clustering - K-means." A Tutorial on Clustering Algorithms. Polytechnic University of Milan, Web. 08 Dec. 2014.

Questions?

Contact

Michael Zamani - mzamani1@vt.edu
Hayden Lee - hjl33@vt.edu
Michael Trujillo - mtruj@vt.edu
Jordan Plahn - jplahn@vt.edu