CS4984/CS598: Big Data Text Summarization

Final Presentation (11/29/2018)

Team 3 - Hurricane Matthew
Professor: Edward A. Fox
Areeb Asif, Michael Goldsworthy, Brendan Gregos, Thoang Tran

Virginia Tech, Blacksburg, VA 24061

Contents

- **Introduction**
  - Tools used
  - Cleaning
- **Important Results**
  - Word count and POS tagging
  - LDA filtering
  - Regex and Template Summary
- **Conclusions**
  - Analysis of tools
  - Takeaways and future work

Introduction

Tools and Environments

- Zeppelin
- Hadoop
- Spark
- Python
- NLTK
- Git
- SpaCy
- Scikit learn
- Gensim

Cleaning

- Remove HTML code
- Remove empty articles
- Remove duplicate articles
- Remove articles that have less than 3 sentences
- Remove articles that have high strength in bad LDA topics

Important Results

Word Frequency

- Remove punctuation
- Remove stop words and custom stop words

Words	Frequency	Important Words	TF-IDF
hurricane	18581	hurricane	11727.4
matthew	9754	storm	9277.9
storm	9253	matthew	9069.8
people	7242	october	8320
said	4925	people	8014
florida	4516	haiti	6973
haiti	4313	florida	6674
south	4294	destruction	6419
october	4058	south	6225
north	4019	winds	5705
one	4006	water	5155
winds	3748	carolina	5152
new	3672	beach	5143

POS Noun and Verb Frequency

- Initially: tagged words from word count list in unit 1
- Better: NLTK POS tagger on words in context

Noun (frequency)	Verb
'matthew', 22260	'said', 7889
'hurricane' (proper noun), 20227	'expected', 3206
'storm', 13294	'flooding', 1596
'haiti', 10445	'say', 1504
'people', 9029	'according', 1466
'florida', 8771	'including', 1419
'hurricane' (noun), 7105	'damaged', 1382
'carolina', 6963	'help', 1304
'wind', 5024	'evacuate', 1295
'october', 4569	'get', 1290

Word Lemmatization

- Use “Words Lemmatization” in NLTK
- Compare the results between words frequency with and without lemmatization

Words	Without Lemmatization	With Lemmatization	Difference
hurricane	18581	19439	858
storm	9253	9873	620
matthew	9754	9793	39
people	7242	7259	14
said	4925	4925	0
florida	4516	4519	3
haiti	4313	4314	1
south	4294	4294	0
october	4058	4058	0
north	4019	4019	0
one	4006	4146	140
winds	3748	5295	1547
new	3672	3672	0

LDA Topics

- Used Gensim to make multiple LDA topic models, pick model with highest coherence score
- Useful for cleaning documents with high strength in bad topics
- After cleaning, run again, sentences extracted based on topic strength

LDA Topic examples:

Good topics:

```
(4, u'0.033*"hurricane" + 0.027*"matthew"
+ 0.019*"carolina" + 0.019*"county" +
0.014*"north" + 0.013*"home" +
0.012*"photo" + 0.010*"power" +
0.010*"flooding" + 0.010*"october"')

(6, u'0.033*"hurricane" + 0.028*"storm" +
0.021*"matthew" + 0.019*"florida" +
0.015*"wind" + 0.011*"coast" +
0.008*"beach" +0.008*"south" +
0.007*"friday" + 0.007*"area"')
```

Bad topics:

```
(1, u'0.014*"trump" + 0.011*"election" +
0.010*"state" + 0.010*"clinton" +
0.008*"voter" + 0.008*"court" +
0.007*"vote" + 0.006*"president" +
0.006*"political" + 0.006*"white"')

(7, u'0.039*"news" + 0.033*"health" +
0.014*"disease" + 0.011*"blog" +
0.010*"google" + 0.010*"report" +
0.009*"zika" + 0.008*"french" +
0.008*"update" + 0.007*"site"')
```

LDA Sentences:

Good sentence:

After passing Jamaica and Haiti, Matthew's centre was expected to pass about 50 miles (80 kilometres) east of the US Navy base at Guantanamo Bay, Cuba, where authorities evacuated about 700 spouses and children of service members on military transport planes to Florida.

Bad sentences:

Share Share Content brought to you by Mike Moss Sept.

Hide Caption 23 of 80 Photos: Hurricane Matthew's path of destruction Adam and Alec Selent watch waves crash over a retainer wall at the Ocean Club condominiums in Isle of Palms, South Carolina, on October 7

Extractive Summary

- Uses TextRank Summariser in Gensim package
- Divides the collection into groups of 300 documents
- Performs the summarization on groups of documents
- Combines the results into into a text file
- Summarizes the text file to obtain extractive summary

Extractive Summary Result

Residents across four states on the east coast are bracing for impact as Hurricane Matthew continues to make its way toward the U.S. The storm headed toward Florida on Wednesday after causing severe devastation across Haiti and bringing heavy rainfall and strong winds to the entire Caribbean region.
Hurricane Matthew pummeled the Florida coast this morning with powerful winds, potentially devastating storm surges and torrential rain.
United States: Downgraded to a Category 3 storm, Hurricane Matthew made landfall as evacuation orders were issued in Florida, Georgia and South Carolina.
More than two million people have been evacuated from Florida, Georgia, North and South Carolina as the United States prepares for Hurricane Matthew to hit.
At 5 p.m. EDT Friday, the National Hurricane Center said Matthew had sustained winds of 110 mph, making it a very powerful Category 2 storm.
Piers in Florida and South Carolina were damaged by storm surge and heavy surf from Hurricane Matthew. Hurricane Matthew has moved on to South Carolina where it made landfall as a Category 1 storm with sustained winds of 75 mph.
Update: Hurricane Matthew has weakened slightly to a Category 2 storm with 110 mph winds off Florida and Georgia coasts.

Named Entity

- Used spaCy to extract named entities from each article
- Created a file with each word followed by entity type or POS
- Used Python’s multiprocessing library to speed up creation

Named Entity Results

Almost	RB
all	DT
mobile	JJ
homes	NNS
in	IN
its	PRP$
path	NN
were	VBD
obliterated	VBN
.	.
At least 40	CARDINAL
people	NNS
were	VBD
killed	VBN
in	IN
Florida	GPE
.	.

The	DT
hurricane	NN
rolled	VBD
across	IN
the	DT
sparsely	RB
populated	JJ
tip	NN
of	IN
Cuba	GPE
overnight	TIME

Template Summary

- Groups named entity results into sentences
- Wrote a method that searches sentences with an input regex and containing a given named entity or POS type.
- Inserts most common result(s) into template.

Template Summary Result

Hurricane Matthew, a category 4 hurricane, made landfall on Tuesday in Haiti. The wind speeds were measured at 140mph, and the hurricane caused 11 deaths as it traveled onto land. The largest amount of damage took place in Haiti, Florida, Fla., Cuba, Roseboro, Baracoa, and Nassau. millions homes lost power. AP, UN, Reuters, Facebook, FEMA, U.N., and NOAA were working to document the damage and assist the victims. There were also reports of looting in Haiti. 175,000 people were evacuated from their homes by authorities in anticipation of the storm.

Conclusion and Takeaways

Analysis of Tools

Most Useful:

- Python
- NLTK
- Git
- SpaCy
- Gensim

Takeaways

- Cleaning data and filtering irrelevant documents is key
- Explore and use multiple tools and libraries
- Don’t waste time on methods that are not working, adapt and work around

Future Work

- Try to classify the collection using supervised machine learning
- Improved template with more slots
- Deep Learning

Thank you!