CS5984:Big Data Text Summarization

Instructor: Dr. Edward A. Fox

Virginia Tech
Blacksburg, VA 24061

Dataset : Hurricane Irma

Team 9

- Raja Venkata Satya Phanindra Chava
- Siddharth Dhar
- Yamini Gaur
- Pranavi Rambhakta
- Sourabh Shetty

About the Dataset

- The dataset provided to us was on Hurricane Irma.
- The dataset consisted of WARC and CDX files to be processed on the DLRL cluster.
- Sentences were extracted by running an Archive Spark Scala script and Solr indexes were created after copying the JSON file with the sentences on the Solr server.
- The dataset consisted of 15,305 documents.

Technologies Used

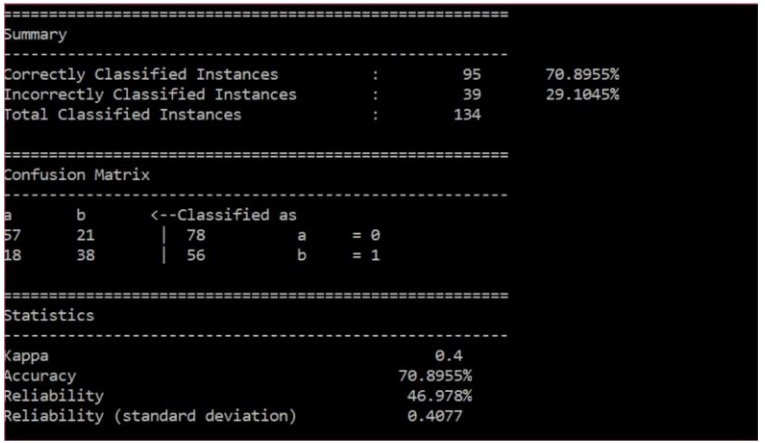
- Python
- NLTK
- Scala
- PySpark
- TensorFlow

Data Preprocessing

- Conversion of WARC file to JSON format.
- Extracted sentences of each document from JSON.
- Noise removal like boilerplate using jusText .
- Normalization.
- Tokenization.
- Stop word removal.
- Lemmatization using POS tags.

Classification Model

- Mahout CBayes classifier to classify documents as relevant and irrelevant - about **70% accuracy** on 300 pre-labelled documents.
- Bad results on the whole dataset - 2 documents classified as irrelevant.
- Hurricane Harvey and Hurricane Irma almost occurred at the same duration of time, which is what made the classification more complicated.



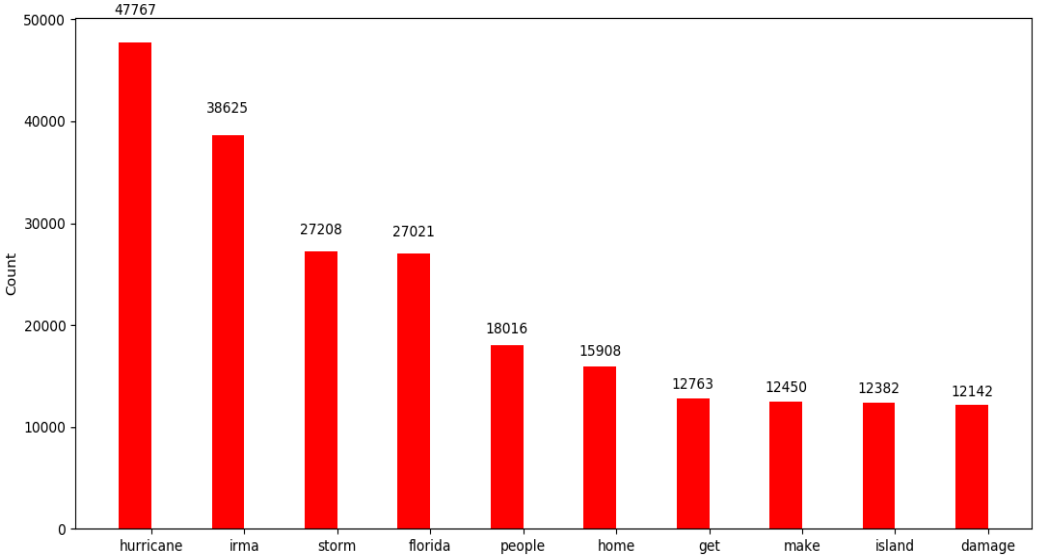
Decision Rules Classifier

Simpler Approach (**Occam's razor)**

- Removed irrelevant documents by using word filters like ‘Irma’ or ‘irma’.
- Removed duplicate documents.
- The dataset now contains 7000 documents (about 50% of the initial dataset).

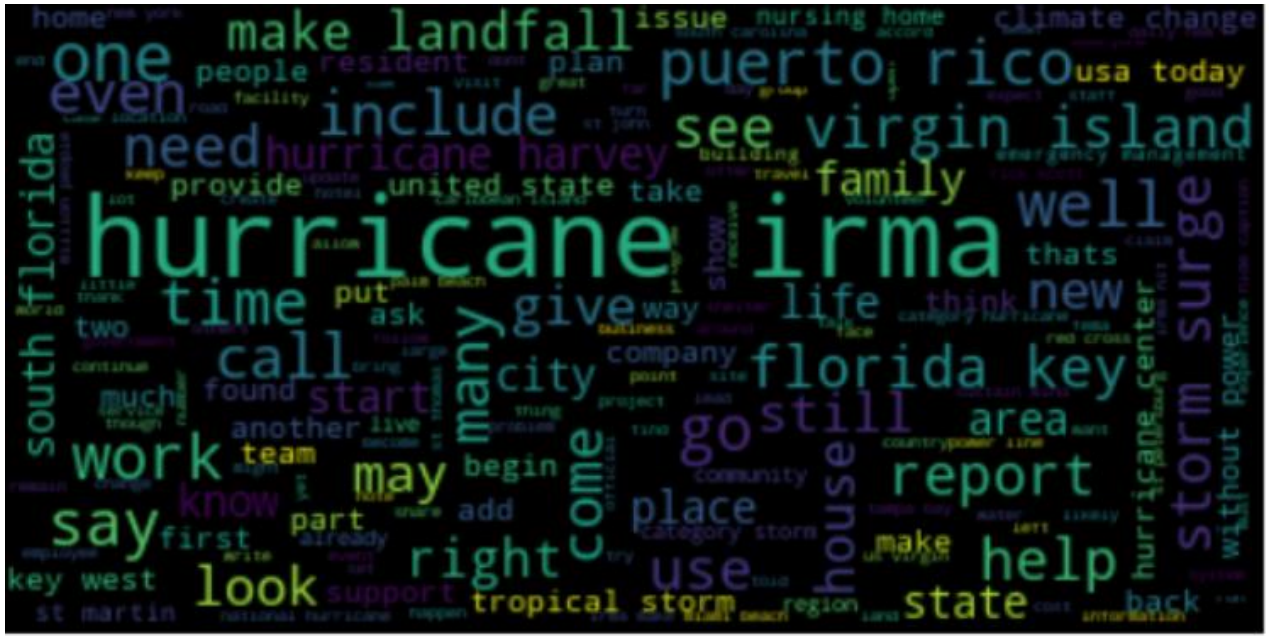
Exploratory Analysis

- Most Frequent Words



Exploratory Analysis

- Most Frequent Words



Exploratory Analysis

- Bigrams
- Named Entities
- LDA Topic Modeling

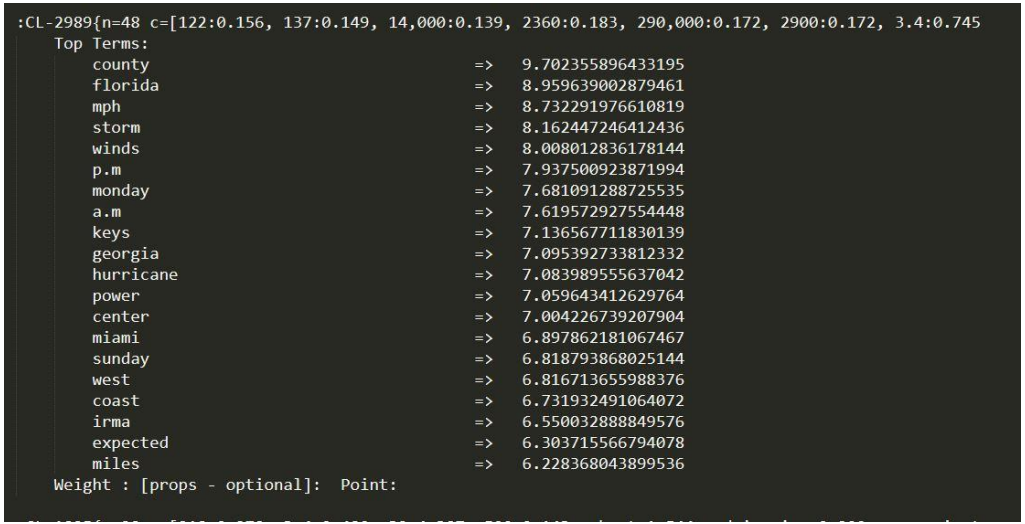
```
(0, u'0.028*"irma" + 0.028*"hurricane" + 0.020*"2017" + 0.015*"florida"')
(1, u'0.016*"florida" + 0.013*"power" + 0.013*"county" + 0.011*"irma"')
(2, u'0.011*"usgs" + 0.010*"coastal" + 0.010*"science" + 0.009*"research"')
(3, u'0.010*"get" + 0.010*"go" + 0.008*"like" + 0.008*"one"')
(4, u'0.024*"year" + 0.018*"name" + 0.008*"school" + 0.007*"newspress"')
(5, u'0.011*"hurricane" + 0.007*"disaster" + 0.005*"need" + 0.005*"help"')
(6, u'0.035*"hurricane" + 0.029*"storm" + 0.024*"irma" + 0.016*"wind"')
(7, u'0.027*"island" + 0.016*"caribbean" + 0.014*"st" + 0.013*"irma"')
(8, u'0.015*"help" + 0.013*"hurricane" + 0.012*"shelter" + 0.010*"need"')
(9, u'0.008*"home" + 0.007*"nursing" + 0.007*"facility" + 0.006*"damage"')
```

Clustering

- Used Mahout's K-means clustering to cluster similar documents together, for possible generating summaries for each cluster and then combining them to get a complete summary.
- Ran the K-means clustering for 10 clusters and 20 iterations.
- While attempting to extract the clustering results using Mahout's 'clusterdump' command, the Hadoop cluster continuously ran into memory errors.

Clustering

- The partial results that were generated showed well formed clusters.



```
:CL-2989{n=48 c=[122:0.156, 137:0.149, 14,000:0.139, 2360:0.183, 290,000:0.172, 2900:0.172, 3.4:0.745
    Top Terms:
        county                                  =>   9.702355896433195
        florida                                 =>   8.959639002879461
        mph                                     =>   8.732291976610819
        storm                                   =>   8.162447246412436
        winds                                   =>   8.008012836178144
        p.m                                     =>   7.937500923871994
        monday                                  =>   7.681091288725535
        a.m                                     =>   7.619572927554448
        keys                                    =>   7.136567711830139
        georgia                                 =>   7.095392733812332
        hurricane                               =>   7.083989555637042
        power                                   =>   7.059643412629764
        center                                  =>   7.004226739207904
        miami                                   =>   6.897862181067467
        sunday                                  =>   6.818793868025144
        west                                    =>   6.816713655988376
        coast                                   =>   6.731932491064072
        irma                                    =>   6.550032888849576
        expected                                =>   6.303715566794078
        miles                                   =>   6.228368043899536
    Weight : [props - optional]:  Point:
```

- Ultimately decided to drop the idea due to lack of time.

Deep Learning Model

- Two deep learning approaches : TensorFlow and PyTorch.
- After classification and noise removal, the number of relevant documents were reduced to 7157.
- Concatenated the data from all articles post classification into a single file.
- Preprocessed the data into binary (bin) files.
- Used the Pointer Generator Network (PGN), which implements Recurrent Neural Network (RNN) to obtain summaries.
- Generated vocab file and checkpoints from train mode.
- Generated summary from decode mode of PGN which takes vocab file test binary files and checkpoints as input and generates abstractive summary.

Data Post-processing

- Unwanted words from the vocab file were included in the summary.
- Added custom stopwords based on analysis of the summary and filtered them from the summary.
- During data pre-processing, for lemmatization, stemming and POS tagging we had converted all the words in the dataset to lowercase and trained the pointer generator on it. We wrote a Python script to capitalize the first letter of all POS tagged proper nouns, the first alphabetic character after every period, etc.
- Since the generated deep learning abstractive summary was longer than two pages, we did manual post-processing to cut down the summary to two pages.

Snippet of the Generated Summary

“ Hurricane Irma made landfall on September 10 as a category 4 hurricane at 9:10 a.m. on Cudjoe Key, with wind gusts reaching 130 mph. States of emergency were also issued in Alabama , Georgia , North Carolina and South Carolina. Hurricane Irma made another landfall in Naples, Florida. Irma, one of the strongest hurricanes on record in the Atlantic basin, made landfall a total of seven times . The storm gradually lost strength, weakening to a category 1 hurricane by the morning of September 11, 2017. At 5 a.m. ET, Irma was carrying maximum sustained winds of nearly 75 mph.”

ROUGE Evaluation

- Rouge_para

Rouge-1	Rouge-2	Rouge-L	Rouge-SU4
0.16667	0.0	0.11111	0.025

- Rouge_sent

  Max ROUGE-1 score among sentences: 0.84615
  Max ROUGE-2 score among sentences: 0.41667

- Cove_entity

  Entity Coverage: 12.28%

Gold Standard for Team 12

- We were tasked with creating the gold standard for Team 12.
- The dataset was on Hurricane Florence.
- Since the hurricane was recent, the dataset had some outdated data.

Conclusion

- Generated an abstractive summary, given a data corpus of over 15,000 articles on Hurricane Irma.
- Performed data preprocessing like noise removal, normalization and tokenization so that the deep learning model could learn efficiently from the data.
- Classification and clustering was performed on the data to filter out the irrelevant documents.
- Used the Pointer-Generator network to generate an abstractive summary of the data.

Future Work

- Create bigrams and trigrams to include countries and cities.
- Implement a better classification model.
- Complete the clustering of documents and generate summaries for each cluster.
- One of the other challenges we faced in post-processing was that the data listed the days as days of the week instead of absolute dates. As future work, we would like to convert relative dates to absolute dates to present a more cohesive timeline of events.

Acknowledgement

- We would like to extend our sincere thanks to Dr. Edward Fox and the teaching assistant for this course, Liuqing Li. We would also like to thank Prashant Chandrasekhar for his valuable inputs in helping us generate Gold Standard summaries for Team 12. We would also like to thank Chreston Miller for providing scripts for data preprocessing for the Pointer Generator Network.

- NSF grant IIS-1619028: Collaborative Research: Global Event and Trend Archive Research(GETAR).

Questions?