CS4984/5984 Big Data Text Summarization

Team 15: Maryland Shooting

Final Presentation

12/04/2018

Prapti Khawas
Bipasha Banerjee
Yoonjin Kim
Shuqi Zhao
Yiyang Fan

Instructor: Dr. Edward A. Fox
Virginia Tech
Blacksburg, VA - 24061

Methodology

- A set of most frequent important words
- A set of WordNet synsets that cover the words
- A set of words constrained by POS, e.g., nouns and/or verbs
- A set of words/word stems that are discriminating features (that also are helpful in a classifier for the relevant webpages)
- A set of frequent & important named entities
- A set of important topics, e.g., identified using LDA
- An extractive summary, as a set of important sentences, e.g., identified by clustering
- A set of values for each slot matching collection semantics
- A readable summary explaining the slots & values
- A readable abstractive summary, e.g., from deep learning

Preprocessing Data:

- Processed WARC and CDX files (initial dataset) on DLRL cluster by cleaning the content using OpenNLP's sentence detector tool.
- Converted cleaned content into JSON format and copied the resulting JSON file to our directory.
- Indexed the JSON file on Solr.
- Converted JSON file into text files each containing content from an article.
- Cleaned the articles by eliminating sentences containing unwanted words such as 'login', 'advertisement', 'video' etc.
- Segregated articles into two groups using keywords "Capital Gazette" and "Great Mills".

Most Frequent Words

[(u'school', 27314), (u'said', 24232), (u'shooting', 20658), (u'police', 20047), (u'high', 15357), (u'capital', 15130), (u'gazette', 12219), (u'news', 12088), (u'people', 12068), (u'maryland', 11254), (u'shot', 10680), (u'great', 10464), (u'mills', 8977), (u'students', 8918), (u'one', 8719), (u'two', 8703), (u'newspaper', 7434), (u'rollins', 7329), (u'also', 7093), (u'officer', 7078), (u'ramos', 7065), (u'county', 6856), (u'killed', 6617), (u'shooter', 6530), (u'student', 6428), (u'gunman', 6132), (u'gun', 6104), (u'told', 6055), (u'five', 5960), (u'baltimore', 5876), (u'authorities', 5858), (u'suspect', 5604), (u'according', 5425), (u'new', 5393), (u'cameron', 5038), (u'resource', 5009), (u'local', 4827), (u'media', 4500), (u'fired', 4470), (u'rights', 4343), (u'first', 4227), (u'parents', 4198), (u'office', 4195), (u'victims', 4171), (u'would', 4130), (u'identified', 4111), (u'building', 4110), (u'2018', 4053), (u'thursday', 3952), (u'june', 3940), (u'reporter', 3883), (u'annapolis', 3829), (u'law', 3815), (u'man', 3764), (u'heard', 3738), (u'officers', 3706), (u'newsroom', 3699), (u'inside', 3691), (u'tuesday', 3663), (u'injured', 3616), (u'could', 3472), (u'scene', 3457), (u'may', 3416), (u'anne', 3408), (u'court', 3381), (u'last', 3347), (u'died', 3344), (u'march', 3332), (u'fire', 3323), (u'say', 3308), (u'arundel', 3297), (u'three', 3282), (u'says', 3263), (u'video', 3187), (u'later', 3158), (u'trump', 3141), (u'get', 3140), (u'us', 3140), (u'journalists', 3114), (u'editor', 3066), (u'public', 3011), (u'several', 2969), (u'sheriff', 2950), (u'enforcement', 2946), (u'around', 2927), (u'male', 2865), (u'please', 2858), (u'press', 2790), (u'jarrod', 2783), (u'taken', 2755), (u'many', 2754), (u'sun', 2749), (u'female', 2742), (u'multiple', 2729), (u'officials', 2721), (u'home', 2631), (u'took', 2614), (u'like', 2556), (u'back', 2555), (u'opened', 2552)]

Most Frequent Collocations

[(u'Capital', u'Gazette'), (u'Great', u'Mills'), (u'Mills', u'High'), (u'High', u'School'), (u'school', u'resource'), (u'resource', u'officer'), (u'high', u'school'), (u'Anne', u'Arundel'), (u'Cameron', u'said'), (u'law', u'enforcement'), (u'Arundel', u'County'), (u'Baltimore', u'Sun'), (u'opened', u'fire'), (u'school', u'shooting'), (u'New', u'York'), (u'five', u'people'), (u'Leonardtown', u'High'), (u'Gazette', u'newspaper'), (u'said', u'Rollins'), (u'shooting', u'Capital'), (u'Police', u'said'), (u'Austin', u'Wyatt'), (u'BuzzFeed', u'News'), (u'Stoneman', u'Douglas'), (u'people', u'killed'), (u'June', u'2018'), (u'social', u'media'), (u'Maryland', u'high'), (u'shooting', u'Great'), (u'Rollins', u'shot'), (u'five', u'counts'), (u'Donald', u'Trump'), (u'gun', u'violence'), (u'Tim', u'Cameron'), (u'multiple', u'people'), (u'Gazette', u'newsroom'), (u'said', u'suspect'), (u'Maryland', u'school'), (u'male', u'student'), (u'glass', u'door'), (u'killing', u'five'), (u'thoughts', u'prayers'), (u'Sheriff', u'Tim'), (u'Jarrod', u'Ramos'), (u'Associated', u'Press'), (u'class', u'heard'), (u'people', u'shot'), (u'parents', u'told'), (u'President', u'Donald'), (u'Mason', u'said'), (u'facial', u'recognition'), (u'shooting', u'Maryland'), (u'mass', u'shooting'), (u'two', u'students'), (u'Marjory', u'Stoneman'), (u'University', u'Maryland'), (u'charged', u'five'), (u'County', u'Police'), (u'County', u'Sheriff'), (u'Douglas', u'High'), (u'police', u'officer'), (u'inside', u'Maryland'), (u'shooter', u'killed'), (u'editorial', u'page'), (u'female', u'student'), (u'prayers', u'victims'), (u'identified', u'Austin'), (u'police', u'said'), (u'sales', u'assistant'), (u'Glock', u'handgun'), (u'Mills', u'students'), (u'Police', u'Department'), (u'Tuesday', u'morning'), (u'critical', u'condition'), (u'nearby', u'Leonardtown'), (u'170', u'people'), (u'Bureau', u'Firearms'), (u'Privacy', u'Policy'), (u'identified', u'Jarrod'), (u'shot', u'glass'), (u'later', u'said'), (u'Police', u'Chief'), (u'newsroom', u'shooting'), (u'United', u'States'), (u'Firearms', u'Explosives'), (u'across', u'street'), (u'people', u'injured'), (u'mentioning', u'school'), (u'Police', u'say'), (u'AP', u'Fullscreen'), (u'good', u'condition'), (u'Southern', u'Maryland'), (u'story', u'must'), (u'John', u'McNamara'), (u'clear', u'whether'), (u'smoke', u'grenades'), (u'several', u'people'), (u'gun', u'control'), (u'shot', u'female'), (u'gunman', u'opened')]

POS tagging

[(u'school', 'NN'), (u'said', 'VBD'), (u'shooting', 'JJ'), (u'police', 'NN'), (u'high', 'JJ'), (u'capital', 'NN'), (u'gazette', 'NN'), (u'news', 'NN'), (u'people', 'NNS'), (u'maryland', 'VBP'), (u'shot', 'JJ'), (u'great', 'JJ'), (u'mills', 'NNS'), (u'students', 'NNS'), (u'one', 'CD'), (u'two', 'CD'), (u'newspaper', 'NN'), (u'rollins', 'NNS'), (u'also', 'RB'), (u'officer', 'NN'), (u'ramos', 'NNS'), (u'county', 'NN'), (u'killed', 'VBN'), (u'shooter', 'JJ'), (u'student', 'NN'), (u'gunman', 'NN'), (u'gun', 'NN'), (u'told', 'VBD'), (u'five', 'CD'), (u'baltimore', 'NN'), (u'authorities', 'NNS'), (u'suspect', 'VBP'), (u'according', 'VBG'), (u'new', 'JJ'), (u'cameron', 'NN'), (u'resource', 'NN'), (u'local', 'JJ'), (u'media', 'NNS'), (u'fired', 'VBD'), (u'rights', 'NNS'), (u'first', 'JJ'), (u'parents', 'NNS'), (u'office', 'NN'), (u'victims', 'NNS'), (u'would', 'MD'), (u'identified', 'VB'), (u'building', 'NN'), (u'2018', 'CD'), (u'thursday', 'NN'), (u'june', 'NN'), (u'reporter', 'NN'), (u'annapolis', 'NN'), (u'law', 'NN'), (u'man', 'NN'), (u'heard', 'VBD'), (u'officers', 'NNS'), (u'newsroom', 'JJ'), (u'inside', 'IN'), (u'tuesday', 'NN'), (u'injured', 'VBN'), (u'could', 'MD'), (u'scene', 'VB'), (u'may', 'MD'), (u'anne', 'VB'), (u'court', 'NN'), (u'last', 'JJ'), (u'died', 'VBD'), (u'march', 'JJ'), (u'fire', 'NN'), (u'say', 'VBP'), (u'arundel', 'IN'), (u'three', 'CD'), (u'says', 'VBZ'), (u'video', 'NN'), (u'later', 'RB'), (u'trump', 'VB'), (u'get', 'VB'), (u'us', 'PRP'), (u'journalists', 'NNS'), (u'editor', 'NN'), (u'public', 'JJ'), (u'several', 'JJ'), (u'sheriff', 'JJ'), (u'enforcement', 'NN'), (u'around', 'IN'), (u'male', 'JJ'), (u'please', 'NN'), (u'press', 'NN'), (u'jarrod', 'NN'), (u'taken', 'VBN'), (u'many', 'JJ'), (u'sun', 'JJ'), (u'female', 'JJ'), (u'multiple', 'NN'), (u'officials', 'NNS'), (u'home', 'VBP'), (u'took', 'VBD'), (u'like', 'IN'), (u'back', 'RP'), (u'opened', 'VBD')]

POS Tagging

Nouns:

[u'school', u'police', u'capital', u'gazette', u'news', u'people', u'mills', u'students', u'newspaper', u'rollins', u'officer', u'ramos', u'county', u'student', u'gunman', u'gun', u'baltimore', u'authorities', u'cameron', u'resource', u'media', u'rights', u'parents', u'office', u'victims', u'building', u'thursday', u'june', u'reporter', u'annapolis', u'law', u'man', u'officers', u'tuesday', u'court', u'fire', u'video', u'journalists', u'editor', u'enforcement', u'please', u'press', u'jarrod', u'multiple', u'officials']

Verbs:

[u'said', u'killed', u'told', u'according', u'fired', u'identified', u'heard', u'injured', u'scene', u'anne', u'died', u'trump', u'get', u'taken', u'took', u'opened']

LDA

- Split each document into sentences using NLTK tokenizer,
- Run LDA topic model on our big dataset

Topic 1 : 0.011*"capital" + 0.010*"shooting" + 0.008*"gazette" + 0.006*"news" + 0.006*"newspaper" + 0.006*"said" + 0.005*"one" + 0.005*"ramos" + 0.004*"rights" + 0.004*"people"

Topic 2 : 0.011*"said" + 0.008*"capital" + 0.007*"shooting" + 0.006*"gazette" + 0.006*"police" + 0.004*"news" + 0.004*"people" + 0.004*"maryland" + 0.004*"victims" + 0.004*"newspaper"

Topic 3 : 0.019*"school" + 0.011*"high" + 0.010*"police" + 0.008*"great" + 0.008*"said" + 0.008*"mills" + 0.006*"students" + 0.006*"said." + 0.006*"shooting" + 0.005*"officer"

Kmeans

Newsroom shooting with k=10

centroid: ['bill:7.108', ' ramos:5.781', ' annapolis:4.272', ' gazette:3.786', ' june:5.137', ' inside:5.143', ' court:4.921', ' suspect:4.735', " here's:8.002", ' suspected:6.040', " suspect's:7.203", ' district:6.393', ' 29:5.792', ' hearing:8.337', ' arundel:5.143', ' anne:5.093', ' sketch:12.463', ' maryland:3.735', ' captured:8.120', ' scene:4.801', ' shooting:3.274', ' shooter:4.385', ' jarrod:7.529', ' 2018:4.601', ' during:5.202', ' bail:9.279', ' capital:3.718', ' et:6.733', ' courtroom:12.870', ' from:3.416']

Kmeans

School shooting with K=10

centroid: ['attacked:7.381', ' state:5.461', ' gunman:4.481', ' shot:5.620', ' unsupported:8.423', ' school:3.202', ' new:4.725', ' panel:8.605', ' your:4.661', ' police:3.472', ' external:7.649', ' links:7.170', ' window:7.912', ' media:6.831', ' injured:4.978', ' gunfight:8.135', ' maryland:3.724', ' our:4.541', ' dead:5.252', ' officer:4.280', " caption'greatest:9.116", ' device:7.219', ' two:4.272', ' us:5.252', ' playback:8.423', ' teenage:7.270', ' after:3.666', ' high:3.953', ' open:6.996', ' children:5.444', ' who:3.707', ' close:6.631', ' students:4.211', ' nightmare:8.828', ' share:5.784']

Named Entity

Using NLTK's ne_chunk()

said	VBD
student	NN
shooting	VBG
Rollins, Mills, Great	PERSON - NNP
officer	NN
School, High, Maryland	PERSON - NNP
shooter	NN
Cameron	PERSON - NNP
Tuesday	NNP
gun	NN
parent	NN
fired	VBD
Gaskill	PERSON- NNP

Named Entity

Using SpaCy's Named Entity Recognition

Tuesday	DATE
Mills	PERSON
Leonardtown High	PERSON
Maryland	GPE
Blaine Gaskill	PERSON
Wyatt Rollins	PERSON
Great Mills	ORG
20 2018	DATE
Tim Cameron	PERSON
Austin Wyatt	PERSON
14-year-old	DATE
Jaelynn Willey	PERSON
St Marys	PERSON

Regex extraction

What is the age of the shooter?
Regex used: ([0-9][1-9]-year-old|[0-9][1-9],|[0-9][1-9]-yo)
‘16-year-old’
‘17-year-old’
‘17-year-old’
‘16-year-old’
‘17-year-old’
‘17,’
‘17-year-old’
‘17,’
‘17,’
‘17,’
‘17,’
‘16-year-old’
‘17,’
‘16-year-old’
‘17,’
‘17,’
‘17,’
‘17,’

Did place of shooting change their security measures after the shooting?
Regex used: ([school]*[control]*
.{0,100}measures[a-zA-Z0-9 ,-]*.)
‘school systems have pressed for other measures, such as increasing security at school facilities.’
‘gun control measures.’

Punishment of shooter(s), did the shooter plead insanity?
Regex used: (((face |charges
).*(penalty)[\w\s\d,-]*\.)|(Ramos|He).{0,30}(plead[\w\s\d,-]*.)
)
‘he had pleaded guilty to criminal harassment.’
‘Ramos ended up pleading guilty to a misdemeanour harassment charge.’
‘He was charged and pleaded guilty.’
‘He pleaded guilty in July 2011 to harassment and was sentenced to 18 months of supervised probation and ordered to attend counseling.’

Extractive Summary - Newsroom shooting

On June 28, 2017, Jarrod Ramos, 38 opened fire at Capital Gazette, Maryland at 2:34 p.m. The shooter killed 5 and several were left wounded before getting caught. The victims were aged between 34 and 65. County Police arrived in 1-minute. The shooter fired some rounds using 12-gauge pump-action shotgun. The motive behind this attack was he had a previous dispute with the newspaper. Ramos faced a grand jury indictment within the next 30 days. In the wake of that attack, and the general rise in threats, many newsrooms have stepped up emergency drills and taken other measures to increase security.

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
0.33333	0.08571	0.25	0.13

Extractive Summary - School shooting

On March 20, 2018, Austin Wyatt Rollins, 17 opened fire at Great Mills High School, in St. Mary's County, MD at 7.55 am in the morning. The shooter killed one and was left wounded. The victims were aged between 14 and 17. The shooter fired one round using a semi-automatic glock. The shooter had ended his relationship with the girl he shot and that is believed to be the motive. Moments after the gunfire, a school officer was alerted, and he fired a round in order to engage the shooter. Austin was killed by the officer's bullet. The students asked for further school safety measures and called for an end to gun violence.

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
0.26471	0.15152	0.23529	0.09574

Deep Learning: Pointer Generator

- Broke down articles into sentences and clustered sentences by K-means with Cosine similarity distance metric.
- Used the Pointer Generator method as described in “Get To The Point: Summarization with Pointer-Generator Networks” to summarize each of the clustered sentences.
- Concatenated the summaries from each of the clusters to get a final summary.
- The repository used for the process was https://github.com/chmille3/process_data_for_pointer_summrizer.
- We used the pretrained model with the CNN/Daily News as the vocab files for the training set.

Abstractive Summary

Newsroom Shooting Summary Snippet:

5 dead in ’targeted attack’ at Capital Gazette newspaper on June 28, 2018, in Annapolis, Maryland, with many others wounded. The suspect was identified in reports as Jarrod Ramos. There were about 170 people in the paper’s Annapolis office when a gunman shot through a glass door, opening fire on multiple employees. He is accused of entering the capital gazette office, firing through a glass door with a 12-gauge shotgun, hunting for victims.

Evaluation - Rouge Paragraph

Event	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
Newsroom Shooting	0.36111	0.11429	0.19444	0.1
School Shooting	0.20588	0.12121	0.17647	0.07979

Evaluation - Rouge Sentence & Entity Coverage

Newsroom shooting:
Max ROUGE-1 score among sentences: **0.73333**

Rob Hiaasen, Wendi Winters, Rebecca Smith, Gerald Fischman and John McNamara were shot and killed thursday afternoon. [Predicted Sentence]

Of the 11 employees there that day, 5 were killed: John McNamara, Gerald Fischman, Rob Hiaasen, Wendi Winters, and Rebecca Smith. [Golden Sentence]

Entity Coverage: **38.10%**

School shooting:
Max ROUGE-1 score among sentences: **0.88889**

He shot Jaelynn Willey, 16, in the head, and the bullet also struck a 14-year-old boy, Desmond Barnes, in the leg. [Predicted Sentence]

The bullet that hit Willey also struck 14-year-old Desmond Barnes in the leg. [Golden Sentence]
Entity Coverage: **27.27%**

Challenges

- Team members had no prior knowledge of summarization or deep learning.
- The dataset contained a lot of noise and irrelevant content which proved to be a major challenge.
- The biggest challenge we faced was having two shooting instances in our dataset. So distinguishing between the two events proved to be a challenge.

Takeaways

- Understanding and cleaning the dataset is important.
- Working with big collection is difficult.
- Communication and patience are essential for teamwork.
- Don’t be afraid to reach out and ask for help.

Acknowledgement

We would like to sincerely express our gratitude to Dr. E. Fox for his constant motivation and encouragement. We would also like to thank our TA , Mr. Liuqing Li for addressing to our queries and helping us out in every aspect.

We would also like to thank Chreston Miller for his help with the pointer generator technique.

We would also like to acknowledge our sponsor Global Event and Trend Archive Research (GETAR) project, supported by the National Science Foundation under Grant No. IIS-1619028.

Resources

[1] E. K. Steven Bird and E. Loper, Natural Language Processing with Python. [Online]. Available: https://www.nltk.org/book/

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “BleiNgJordan2003.pdf,” vol. 3, pp. 993–1022, 2003.

[3] “K-means clustering.” [Online]. Available: https://mahout.apache.org/users/clustering/k-means-clustering.html

[4] A. See, P. J. Liu, and C. D. Manning, “Get To The Point: Summarization with Pointer-Generator Networks,” 2017. [Online]. Available: http://arxiv.org/abs/1704.04368

[5] “Get to the point: Summarization with pointer-generator networks.” [Online]. Available: https://github.com/chmille3/pointer-generator

[6] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Icml03-Nb.Pdf,” People.Csail.Mit.Edu, no. 1973, 2003. [Online]. Available: http://people.csail.mit.edu/jrennie/papers/icml03-nb.pdf{\%}5Cnpapers2://publication/uuid/B3C25F39-7B42-4ADB-A47C-42595E1D526B

Questions?