Big Data Text Summarization - 2017 Westminster Attack

CS4984/CS5984 Final Presentation

Team 4: Aaron Becker, Colm Gallagher, Jamie Dyer, Jeanine Liebold, Limin Yang

Instructor: Dr. Edward A. Fox

Virginia Tech, Blacksburg, VA

Objective

To generate summaries for a large dataset of web archives containing information about the 2017 Westminster Attack

Westminster Attack Dataset

	Number of Documents	Number of Cleaned Documents	Number of Cleaned Documents Filtering Duplicates
Small Dataset	422	299	N/A
Big Dataset	11298	6900	3996

Summaries Generated

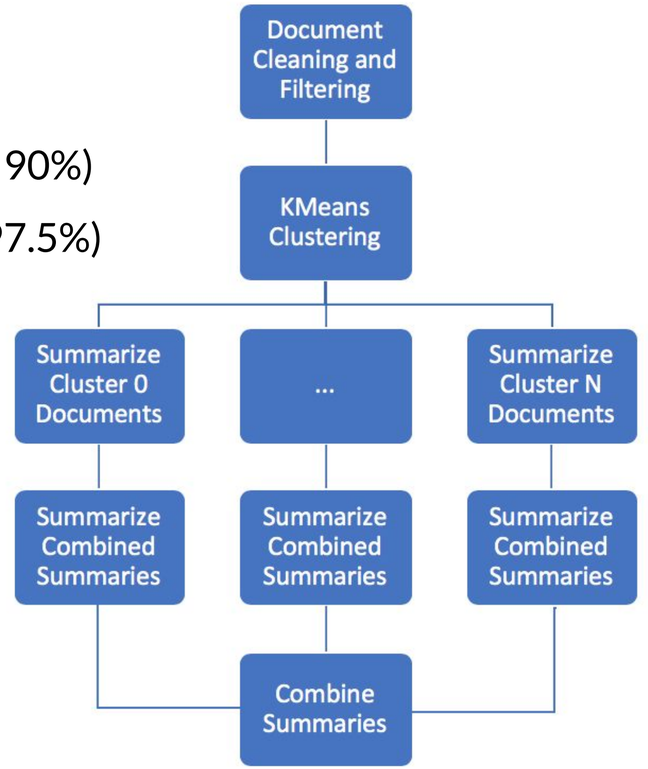
- A Set of Frequent Words
- A Set of Frequent Word Synsets
- A Set of Frequent Nouns, Verbs, and other Parts of Speech
- A Set of Frequent Named Entities for Multiple Categories
- A Set of Document Topics
- An Extractive Summary Formed of Important Sentences
- A Set of Values Extracted and Formed Into a Template
- An Abstractive Summary Formed With Deep Learning

General Methodology

- Preprocessing: jusText (remove boiler template), sklearn’s TFIDFVectorizer (remove duplicates)
- Unit 1, 2, 3: NLTK built-in functions or corpus like word_tokenize, stopwords, Wordnet, Brown, Regular
- Unit 5: spaCy’s statistical entity recognition system
- Unit 6: Gensim's Latent Semantic Analysis
- Unit 7: TextRank (gensim implementation) with K-Means clustering
- Unit 8, 9: regular expression and spaCy’s rule-based matching
- Unit 10: Pointer-generator Network

Methodology for Extractive Summary

- Narrow dataset by ranking documents by use of frequent words from Unit 1
- Filter out sentences longer than 50 words/tokens
- Perform K-Means to cluster documents into N clusters
- Generate summaries for each document in cluster using TextRank (reduced size by 90%)
- Concatenate summaries for each cluster and run TextRank again (reduced size by 97.5%)
- Concatenate cluster summaries as paragraphs in final result



Other Strategies Attempted:

- Summarizing entire concatenated cluster
- Clustering after summarizing documents
- Clustering individual sentences, as well as groups of two or three sentences

First paragraph of Extractive Summary

Five killed and 40 injured in Westminster terror attack. ***An assailant stabbed a policeman and was shot by police just outside Britain's parliament building in London on Wednesday (March 22) in what police described as a "terrorist incident".*** *The ministry said minister Marco Minniti convened the Committee of Strategic Anti-terrorism Analyses following "the tragic facts in London," in which* ***a vehicle mowed down pedestrians on a bridge and the attacker then stabbed a police officer outside the British Parliament***. *At least four people were killed and at least 20 injured in London on Wednesday after a car plowed into pedestrians and an attacker stabbed a policeman close to the British parliament, in what police called a terrorist incident.*

Second paragraph of Extractive Summary

While police believe ***Khalid Masood, a 52-year-old British man with a history of violent crimes, carried out the attack*** *on Westminster Bridge and outside parliament alone, they are investigating what help he may have received and whether he had any accomplices. The attacker who killed three people near parliament in London before being shot dead was named on Thursday as British-born Khalid Masood, who was once investigated by MI5 intelligence officers over concerns about violent extremism. An attacker - now named as 52-year-old Khalid Masood - killed three pedestrians and injured around 40 other people as he mowed down members of the public with a car on Westminster Bridge at about 2:40pm, before crashing into the railings in front of Parliament.* *An attacker - now named as 52-year-old Khalid Masood - killed three pedestrians and injured around 40 other people as he mowed down members of the public with a car on Westminster Bridge at about 2:40pm, before crashing into the railings in front of Parliament.*

Methodology for Slot Value Extraction

- Used spaCy's Named Entity Recognition and Pattern Based Matching to extract values
- Extracted: Date, Location, Nearby Features, Attacker Name, Type of Attack, Number Killed, and Number Injured
- Challenges:
  - Motivation of the attacker was unclear in articles
  - Attack method vs. weapon too difficult to extract
  - People killed and number injured changed when injured victim died weeks later

Query Name	Extracted Value	Count
Type of Attack	Terrorist Attack	119
Attacker	Khalid Masood	50
Location	London	358
People Killed	4	81
Date	Mar. 22, 2017	72
Nearby Feature	Westminster Bridge	242
People Injured	50	35

Slot Value Extraction Summary

Template:

On DATE a TYPEOFATTACK occurred in LOCATION near NEAR. The police investigated that the attacker was ATTACKER. During the TYPEOFATTACK the attacker killed KILLED people and injured INJURED.

Filled in Template:

On Mar. 22, 2017 a terrorist attack occurred in London near Westminster Bridge. The police investigated that the attacker was Khalid Masood. During the terrorist attack the attacker killed 4 people and injured 50.

Methodology for Abstractive Summary

- Model: Pointer-generator Network (PGN, pre-trained on CNN/Daily Mail dataset)
- Test Dataset:
  - 999 articles (ranking every document by its inclusion of frequent words obtained in Unit 1)
  - 3 most relevant topics and corresponding concatenated articles (about 1500 articles)
- Methodology:
  - Concatenated 999 articles together and ran PGN on it
  - Ran PGN on the 3 topics to generate 3 summaries and concatenate together

Results

- Extractive summary
  - Quite relevant
  - Includes important info like what, when, where, perpetrators, why, who was affected, countermeasures.
- Abstractive Summary
  - More natural, but focus on government’s attitude and reactions instead of the event itself
  - Includes info like what, when, who was affected, countermeasures

Evaluation - Mode 1

- ROUGE-1, ROUGE-2: overlap of unigrams and bigrams between our summary and the golden standard
- ROUGE-L: overlap of longest common sub-sequences (LCS)
- ROUGE-SU4: overlap of skip-grams with a max length of 4 as well as unigrams

	ROUGE			
	1	2	L	SU4
Extractive Summary	0.21429	0.03704	0.10714	0.05263

Evaluation - Mode 2 (Extractive Summary)

- Max ROUGE-1 score: 0.71429
  - [Predicted Sentence]
    - An attacker - now named as 52-year-old Khalid Masood - killed three pedestrians and injured around 40 other people as he mowed down members of the public with a car on Westminster Bridge at about 2:40pm, before crashing into the railings in front of Parliament.
  - [Golden Sentence]
    - The attacker was later identified as 52 year old Briton Khalid Masood.
- Max ROUGE-2 score: 0.33333
  - [Predicted Sentence]
    - While police believe Khalid Masood, a 52-year-old British man with a history of violent crimes, carried out the attack on Westminster Bridge and outside parliament alone, they are investigating what help he may have received and whether he had any accomplices.
  - [Golden Sentence]
    - The attacker was later identified as 52 year old Briton Khalid Masood.

Evaluation - Mode 3 (Extractive Summary)

- Entity Coverage: 13.68%
- Examples of matched named entities:
  - Khalid Masood
  - Keith Palmer
  - Adrian Russell
  - Westminster
  - Westminster Bridge
  - London
  - The Houses of Parliament
  - Islamic
  - 52-year-old
  - Five

Lessons Learned

- Data cleaning is difficult but important (and needs to be done early)
- A classifier is an important tool which could have helped select documents for the extractive and abstractive summaries
- Splitting the big dataset into smaller clusters saved time
- Various issues may arise when the dataset gets bigger
- Figure out time complexity before running the code