Team 7 → Final Presentation

CS4984/CS5984 Big Data Text Summarization
November 29, 2018
Blacksburg, Va 24061

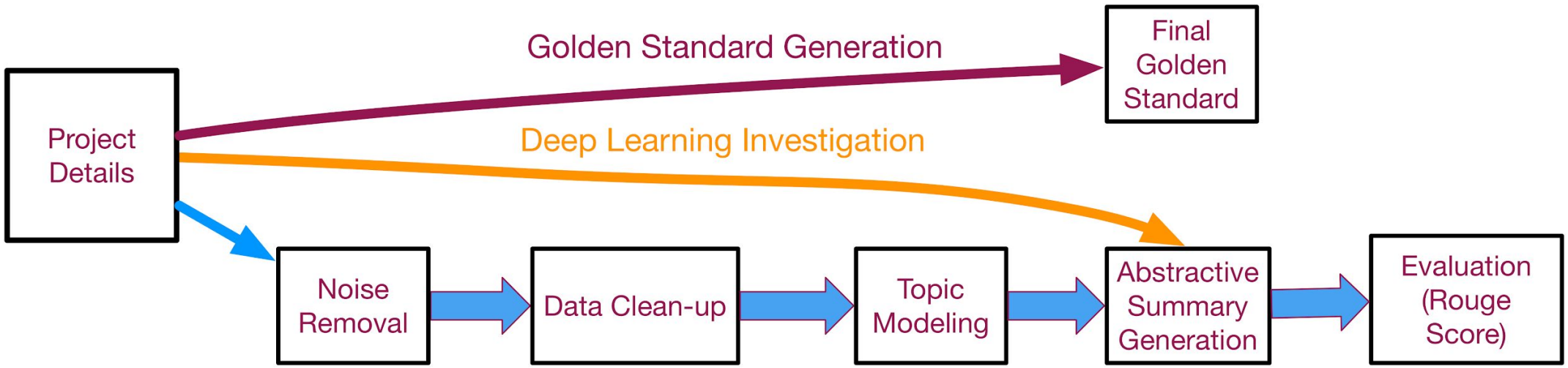
Anuj Arora
Jixiang Fan
Yi Han
Shuai Liu
Chreston Miller

VIRGINIA TECH

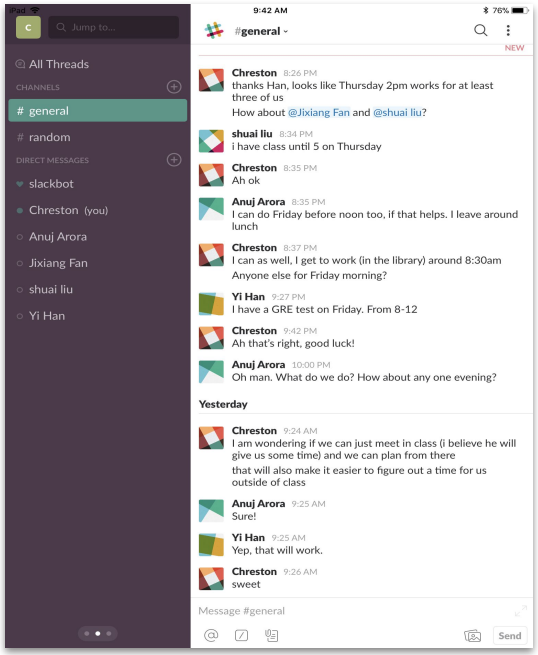
Roadmap

- Introduction
- Project Management
- Motivation
- Golden Standard
- Article Collection
- Pre-Processing Data
- Deep Learning for Summarization
- Post-Processing
- Lessons Learned
- Conclusion

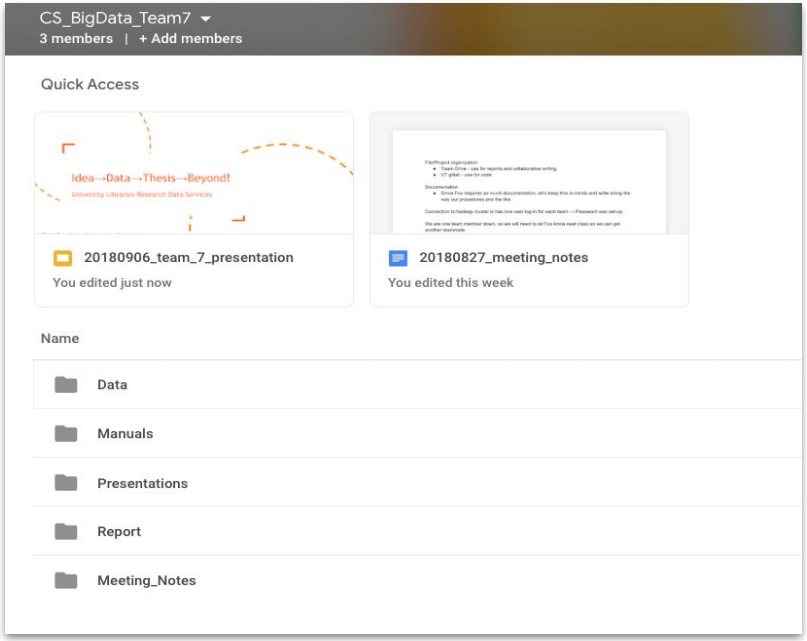
Introduction



Project Management



Slack



Team Drive

GitHub

Code

…and Weekly Team Meetings

VIRGINIA TECH

Motivation

How do we summarize multiple articles, when manually reading through each of them becomes impractical?

Golden Standard

- Create outline
- Use Solr and Google search related material
- Update outline
- Built-in summary tools in Mac OS
- Finish first draft
- Read feedback
- Update summary

Article Collection - NeverAgain

- Collection of articles concerned with and focused around school shootings and gun policy
- The following table shows the article counts before and after the cleanup

Dataset	Before Noise Removal	After Duplicate Removal	After noise filtering	Duplicate Percentage	Total Reduction Percentage
NeverAgain	12,111	7,631	3,669	37%	69%

Pre-Processing

- **Raw Data Conversion and Indexing** (WARC/CDX to JSON)
- **Duplicated Entries Removal**
- **Noise Filtering**
  - Iteration 1: Word Filtering
  - Iteration 2: jusText and Word Filtering
  - Iteration 3: jusText and more restrictive Word Filtering
  - Stopwords, Punctuation Removal and Part of Speech Tagging
  - Lemmatization and Tokenization

Exploratory Results

- After tokenization, we came up with a list of the most frequently occuring words in the cleaned corpus. The top words are shown below

Word	Frequency
gun	25,009
school	21,726
people	12,853
student	12,234
shoot	10,838

Exploratory Results

- NLTK's Named Entity Chunker was used to generate named entities, the most frequent of which are shown below

Named Entities	Frequency
('ORGANIZATION', 'NRA')	4,510
('GPE', 'Parkland')	4,265
('GPE', 'Florida')	2,470
('PERSON', 'Trump')	2,234
('GPE', 'U.S.')	2,233

Topic Modelling

- The tokenized sentences were vectorized using the Bag of Words (BOW) approach
- Gensim was used to implement the LDA (Latent Dirichlet Allocation) algorithm

Topic	Keywords
Topic 0 (Santa Fe)	0.024*"school" + 0.021*"police" + 0.019*"shoot" + 0.012*"people" + 0.012*"kill" + 0.011*"2018" + 0.011*"high" + 0.010*"santa" + 0.010*"fe" + 0.010*"shot"
Topic 1 (General Shootings)	0.030*"school" + 0.021*"student" + 0.014*"gun" + 0.011*"people" + 0.010*"shoot" + 0.010*"high" + 0.009*"parkland" + 0.008*"life" + 0.008*"like" + 0.008*"violence"
Topic 2 (Policy)	0.025*"gun" + 0.012*"state" + 0.008*"nra" + 0.008*"law" + 0.007*"trump" + 0.006*"people" + 0.006*"year" + 0.005*"use" + 0.005*"right" + 0.005*"firearm"

Deep Learning

- Method Exploration
- Separate summaries based on LDA identified topics
- Pointer-Generator Network
  - Hyperparameter tuning
- Training data - CNN/Daily Mail
  - Used a pre-trained model - saved much time
- Processing
  - Local Ubuntu server with a high power graphics card (GPU)

Deep Learning Visualization

- Generated Summary
  - Highlighted = High Generation Probability

suspect dimitrios pagourtzis , 17 , has cooperated with police . henry denied bail for the student , who is accused of capital murder of multiple people and aggravated assault on a public servant . students at santa fe high school , not far from houston in southeastern texas , scrambled for safety after they heard shots just after class began friday morning . nine students and one teacher were killed , a law enforcement official says . `` we need to do more than just pray for the victims and their families , '' gov. greg abbott tells cnn the alleged shooter was `` really quiet and he wore like a trench coat almost every day '' cnn affiliate ktrk where a bullet went in the back of his head and came out near his left ear , substitute teacher

Post-Processing

- Removing redundancy
- Fixing flow issues
- Rearranging sentences
- Given the separate topics, arranging them in a logical order and creating transitions between them

Lessons Learned

- GPU acceleration - surprisingly did not need a cluster
- Software versions are very important
- Ubuntu Virtual Machine would not work with GitHub
  - This was using Parallels virtualization software
- How to use Docker
- How to use Solr
- Pre-processing large datasets to extract meaningful information
- Deep Learning architectures

Conclusion

- Abstractive summarization is challenging
- Careful pre-processing is necessary for filtering articles
- Deep learning solutions/approaches are available via open source repositories
  - Provides a working foundation
- Current methods for abstractive summarization are not a complete solution, however, techniques continue to improve

Questions!

Anuj Arora: aarora23@vt.edu

Jixiang Fan: jfan12@vt.edu

Yi Han: hy@vt.edu

Shuai Liu: shuail8@vt.edu

Chreston Miller: chmille3@vt.edu