Team 14: Facebook Data Breach

April Fitzpatrick, Akshay Goel, Leah Hamilton, Ramya Nandigam, Esther Robb

Final Presentation: 12/4/18

CS 4984/5984 Big Data Text Summarization - Taught by Dr. Edward Fox
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

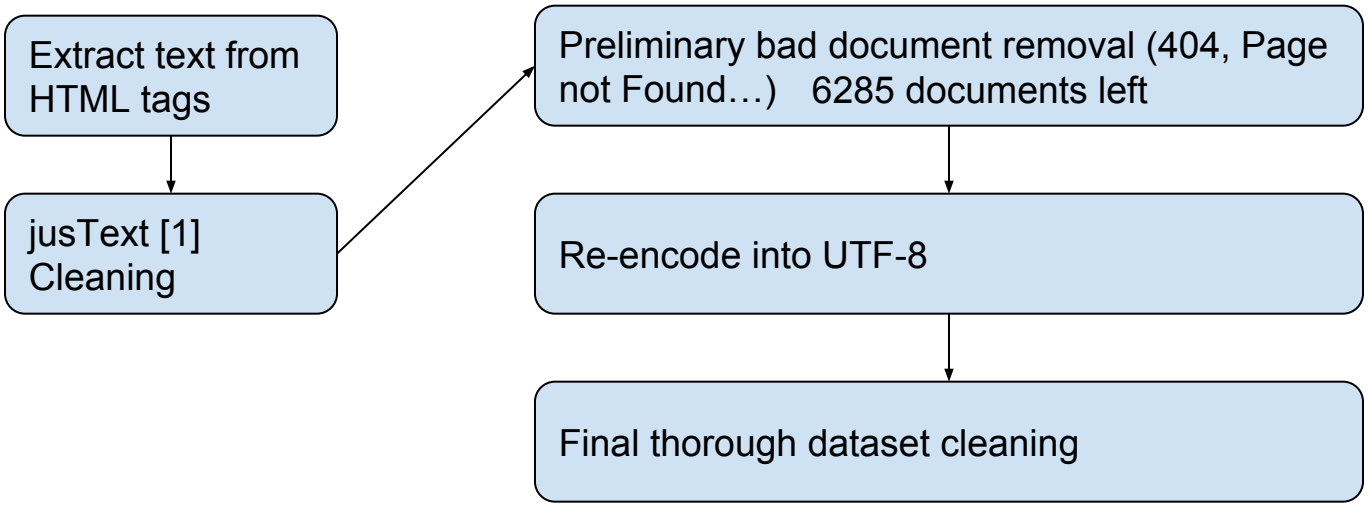
Presentation Overview

- Data and Tools
- Pre-processing
  - Data Cleaning
  - Document Clustering
- Summaries
  - Template Summary
  - Extractive Methods
  - Abstractive Methods
- Evaluation
- Conclusions

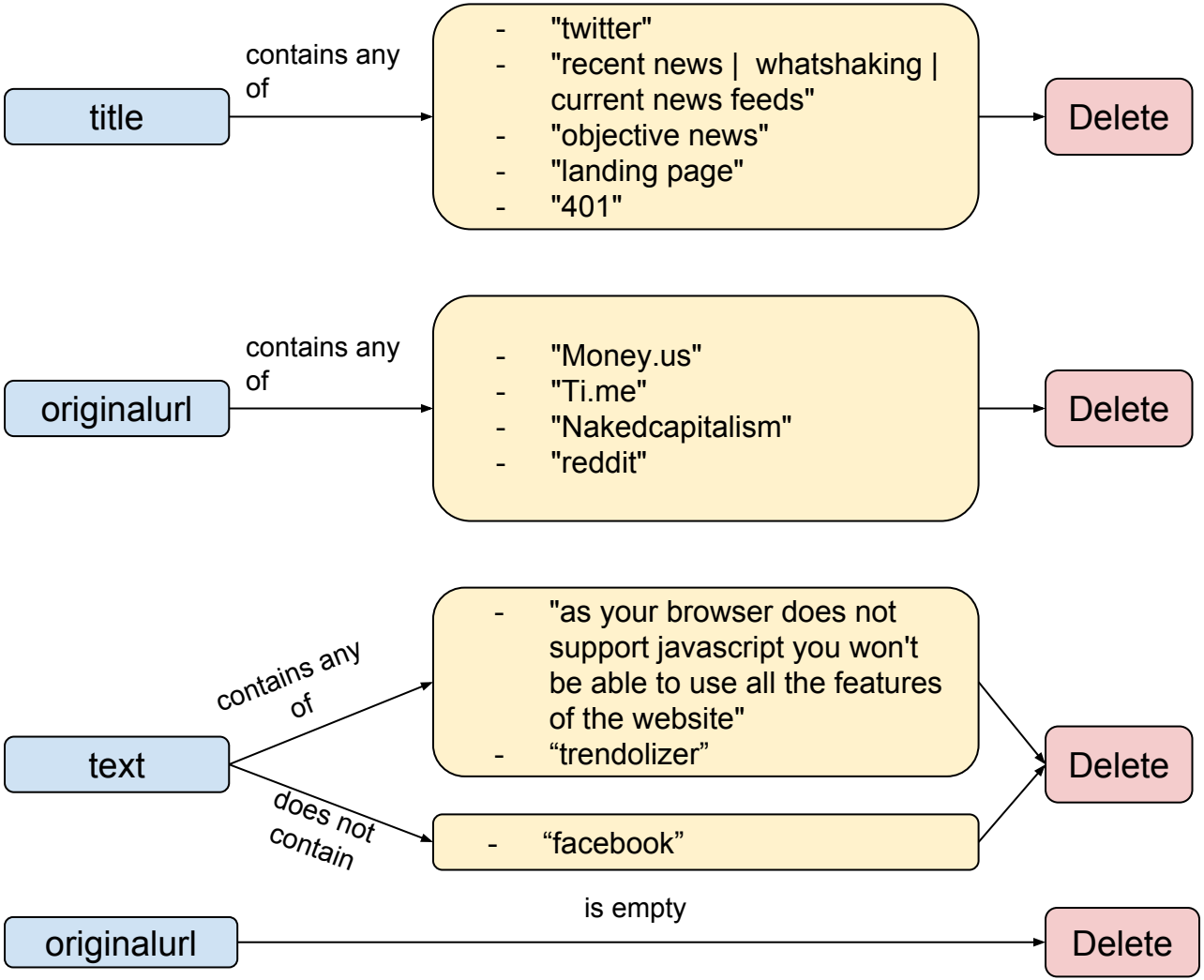
Project Overview

- Dataset: Facebook Data Breach
  - Number of docs: 10829
- Tools Used
  - Scala
  - ArchiveSpark
  - PySpark
  - Python
  - NLTK
  - Gensim
  - SpaCy
  - Pointer-Generator
  - FastAbsRL

Data Cleaning Overview



Parameter Based Filtering



3887 documents left

Removing Similar Documents

2922 documents left

Many documents were exact copies or similar to other documents

$$\forall \textit{ document } D_i,\ D_j \textit{ where } len(D_j) < len(D_i):$$

$$\textit{Let } S_k \textit{ be the set of unique sentences for document } D_k$$

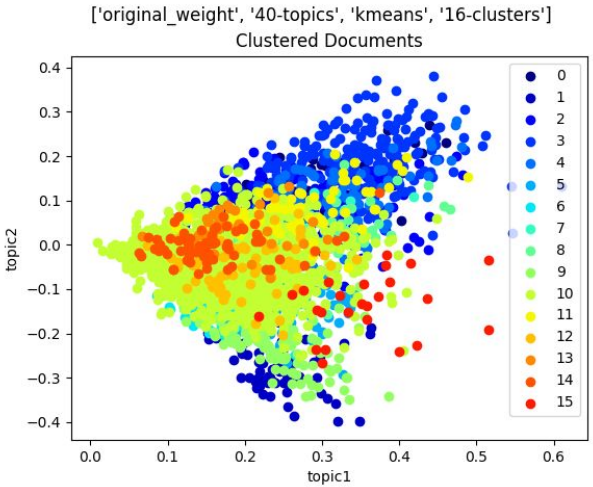
$$Similarity(i, j) = length(S_i \cap S_j) / length(S_j)$$

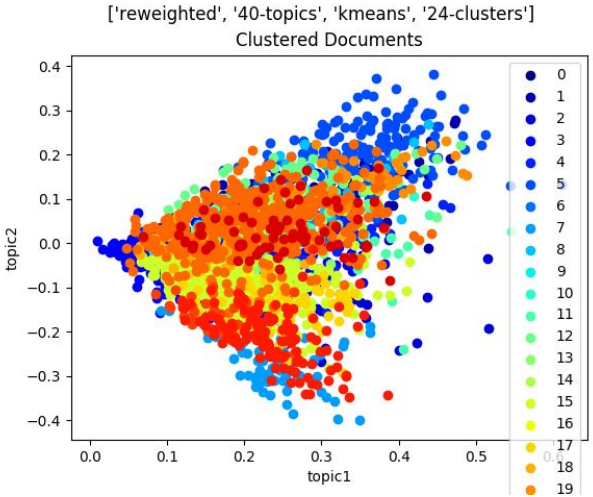
1. Sort all documents in descending order of size
2. Get similarity score for all document pairs (i,j) where j > i
3. If similarity is above threshold, delete document j

Threshold = 0.4

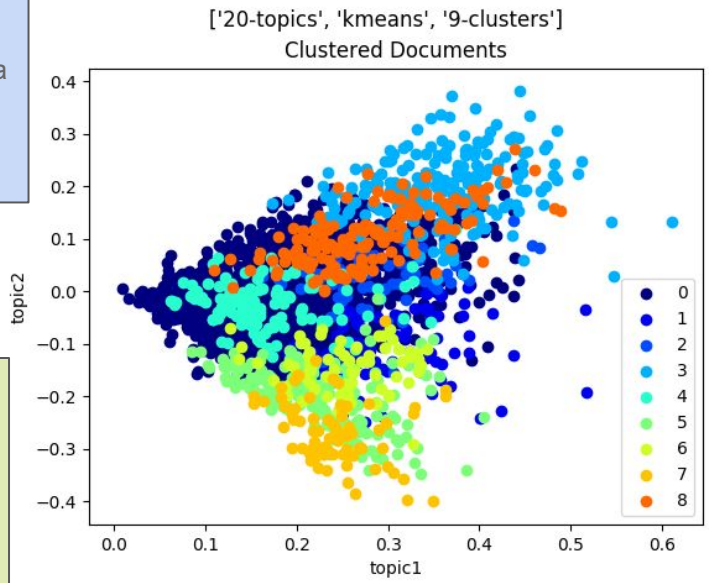
Clustering

- Use LSA to do topic modeling
  - Works better than LDA on our dataset
- Cluster based on LSA topic weights
- Observations:
  - Low number of topics is better
  - K-means works best





~~~~~~~~~~ Cluster 2 ~~~~~~~~~~~~
- **federal trade commission** investigation: is mark zuckerberg headed to facebook jail? #zuckerberg - tu.tv
- **ftc** launches probe into facebook privacy practices | mobile marketing magazine
- facebook **stock** slides after **ftc** launches data leak investigation
- **ftc**, eu, state attorneys general investigating facebook breach



~~~~~~~~~~ Cluster 8 ~~~~~~~~~~~~
- facebook ceo zuckerberg invited to **testify** by **senate** judiciary committee - the peninsula qatar
- facebooks discussions with **congress** signal mark zuckerberg will testify amid data-privacy scandal
- zuckerberg declines to **testify** in uk parliament - teletrader.com
- mark zuckerberg **refuses to give evidence** on facebook scandal | daily mail online

~~~~~~~~~~ Cluster 6 ~~~~~~~~~~~~
- facebook **changing privacy controls** as criticism escalates : the two-way : npr
- facebook **changes layout** to highlight privacy settings | wben 930am
- facebook **announces overhaul** of security and privacy settings - red team news
- facebook data scandal **prompts redesign** of settings, privacy pages | fox news
- facebook **will make it easier** for you to control your personal data | wired

~~~~~~~~~~ Cluster 3 ~~~~~~~~~~~~
- **trump-linked firm** collected data from 50 million facebook profiles -axios
- facebook suspends **trump-linked data firm cambridge analytica** (update: response)
- how **trump** consultants exploited the facebook data of millions | the seattle times
- facebook bans **trump** campaign's data analytics firm, **cambridge analytica** | breitbart
- **trump-linked firm** obtained data of 50m facebook users - cnet

Named Entity Recognition (NER)

- Used SpaCy's [2] Named Entity Recognizer
- Capitalization apparently necessary for tagging of names

Entity Type	Named Entity
'DATE'	'2016'
'QUANTITY'	'as many as 50 million'
'MONEY'	'billions of dollars'

Template Summary

- Used spaCy NER for dates and names.
- Used spaCy word relations and parse trees for phrases.
- Used num2words [3], word2number [4], and regular expressions for numbers:

```
\d+(?=[ ,-.]{1,2}([^ .,:;]+? )?(?:user|profile|customer)s?(?:(?: of )([^ .,:;]+?)[ .,:;])?)
```

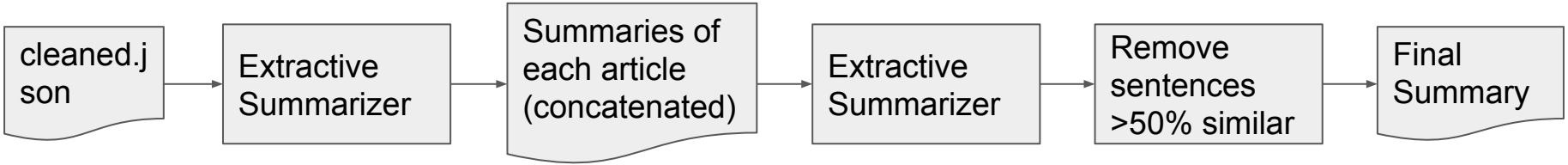
In 2014, the data of 50000000 users of Facebook was compromised. Information from the accounts friends profiles as well as updates, likes, and in some cases private messages was illegally obtained by <ATTACKER>. The incident was made public by <WHISTLEBLOWER> on <DATE OF ANNOUNCEMENT>. Facebook has said <COMPANY STATEMENT> and will <CHANGES EFFECTED>.

Extractive Summaries

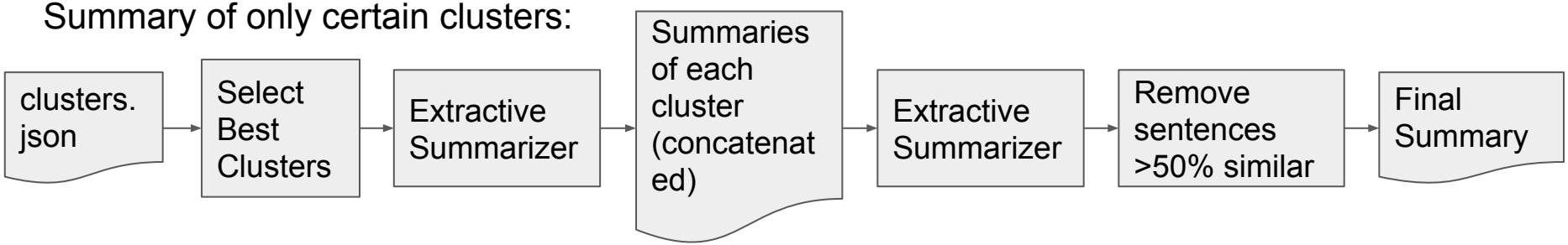
Two methods:

- Both use
  - Gensim's [5] TextRank for summarizing
  - difflib SequenceMatcher [6] for finding similar sentences

Summary on all articles in data set:



Summary of only certain clusters:



Extractive Summaries

- Clustered approach was better
  - Easier to customize important information
- Also broke it up into two paragraphs
  - Each paragraph was a different topic of clusters
  - Got us an even broader range of topics covered
- Clusters chosen:

Paragraph 1: Hack/Security Details	Paragraph 2: Effects of Hack
5 - actions that facebook has taken to improve privacy	11 - zuck testifies before us congress (no uk information)
8 - ftc investigates fb	14 - facebook stock falls
27 - zuckerberg's response to the whole fb crisis	30 - zuck testifies before us congress
28 - christopher wylie, the whistleblower	35 - FB under pressure, value falls
33 - actions that facebook has taken to improve privacy	

Extractive Summary (Paragraph 1)

Tom Pahl, the acting director of the Federal Trade Commission's Consumer Protection Division, wrote in a statement Monday morning that the agency is investigating Facebook's privacy practices a week after news broke that the Trump campaign's political-data firm, Cambridge Analytica, inappropriately obtained data on more than 50 million Facebook users and then allegedly lied about deleting it. Facebook is attempting to do a face saving act following severe criticisms against it so that it is able to maintain its user base and therefore the flow of advertisements and advertisers and investor. The largest social media platform in the world is facing close scrutiny of its privacy policies and actions both in the U.S. and the U.K. Last week there were allegations against Facebook that it did nothing to prevent the use of personal data of approximately 50 million Americans by British consultancy Cambridge Analytica which allegedly had misused the data during the 2016 Presidential elections in the U.S. The firm was appointed to assist President Donald Trump during the campaign. This weekend, a man named Christopher Wylie spoke with the New York Times about a consulting company he founded called Cambridge Analytica that, according to him, developed Facebook ads for the Trump campaign with the help of Steve Bannon and data stolen from the pages of 50 million Facebook users (including personal details, rather than passwords or private information).

Extractive Summary (Paragraph 2)

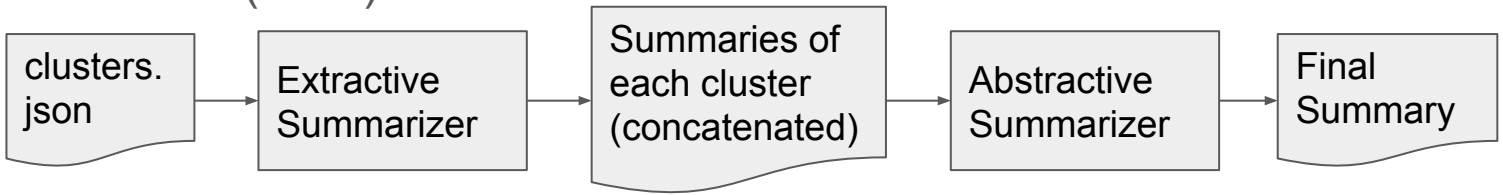
Facebook ended the day down nearly 7 percent, to US$172.56 making it the worst performing stock in the S&P 500, as the company sought to stem the damage from media reports that Cambridge Analytica, the U.S. data-mining arm of a Britain-based research firm, had improperly accessed personal details from nearly 50 million Facebook users to help Trump campaign advisers target political ads during the 2016 election. Calls for probe of misappropriation of the private information of tens of millions of Americans. Former Cambridge Analytica employee Chris Wylie said the company used information to build psychological profiles so voters could be targeted with ads. Wylie criticized Facebook for facilitating the process, saying it should have made more inquiries when they started seeing the records pulled a collection of powerful U.S. senators are demanding that Facebook explain how a third-party firm with ties to the Trump campaign was able to gain access to data on 50 million of its users. Washington revelations that a political data firm may have gained access to the personal information of as many as 50 million Facebook users drew new bipartisan calls on Capitol Hill Monday for Facebook CEO Mark Zuckerberg and the heads of other social media companies to answer questions from Congress.

Abstractive Summaries

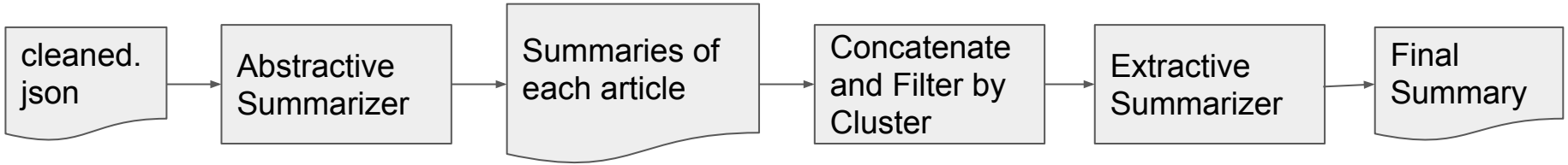
Three Methods:

- Pointer-Generator Network (PGN) or FastAbsRL as abstractive summarizer
- Different sources of input (extractive summary/individual articles)

Method A (PGN)



Method B (PGN and FastAbsRL)



Abstractive Summaries

- PGN (Method B) [7]:
  - Many sentences identical to extractive summary
  - Slightly less repetitive and shorter than pure extractive summary
  - Some issues with pronouns/clauses (e.g., “The company” instead of “Facebook”)
  - Some additional names and information not present in extractive summary
  - Coherency of second paragraph is low
- FastAbsRL [8]:
  - Highly abstractive
  - Some issues with noun repetition (“Facebook, Facebook and Facebook…”)
  - Some issues creating coherent abstractive sentences

PGN Abstractive Summary (Paragraph 1)

The company announced a suite of new, more intuitive privacy controls Wednesday morning, including a way to download and delete data, a redesigned settings menu, and additional shortcuts for controlling private information. Tom Pahl, the acting director of the Federal Trade Commission's Consumer Protection Division, wrote in a statement Monday morning that the agency is investigating Facebook's privacy practices a week after news broke that the Trump campaign's political-data firm, Cambridge Analytica, inappropriately obtained data on more than 50 million Facebook users and then allegedly lied about deleting it. Facebook CEO Mark Zuckerberg apologized on Wednesday for the social media website's role in what he previously called the "Cambridge Analytica Situation" wherein the research firm allegedly accessed 50 million Facebook user profiles improperly. Christopher Wylie, who previously revealed that consultancy Cambridge Analytica had accessed the data of 50 million Facebook users to build voter profiles on behalf of Donald Trump's campaign, said AggregateIQ (AIQ) had built software called Ripon to profile voters. Facebook has announced new controls, privacy shortcuts, and tools to delete facebook data but said these were in the works before the cambridge analytica scandal exploded.

PGN Abstractive Summary (Paragraph 2)

The invitation asking Zuckerberg to answer questions at an April 10 hearing comes as the Federal Trade Commission confirmed it's investigating Facebook's privacy practices after reports the company allowed political consulting firm Cambridge Analytica to harvest 50 million users' data. Analyst: "the risk here is that Facebook is paramount to the future of this company." Facebook shares 6 percent and were on track for their worst day in more than three years on reports that a political consultancy worked on president Donald Trump's campaign gained inappropriate access to data on more than 50 million users. Lawmakers in the United States, Britain, and Europe have called for investigations into media reports that political analytics firm Cambridge Analytica had harvested the private data on more than 50 million Facebook users to support Trump's 2016 presidential election campaign.

ROUGE Evaluation Scores

Summary Type	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
Cluster-based extractive	0.06557	0	0.06557	0.01198
Extractive	0.10714	0	0.10714	0.01316
PGN Method A	0.10169	0	0.0678	0.01863
PGN Method B	0.1	0	0.1	0.01829
FastAbsRL	0.09091	0	0.09091	0.01099

Conclusion & Lessons Learned

- Cluster-based extractive summary was the most useful summary we produced
- Importance of data cleaning
- Should’ve started on clustering and big summaries earlier

Future Work

- Refine clustering results
- Finish information extraction for template
- Try other abstractive algorithms or workflows
- Use summarization techniques on the Solar Eclipse dataset

Acknowledgements

This project would not have been possible without the NSF-funded (NSF: IIS-1619028) Global Event and Trend Archive Research (GETAR) project used to create our collections.

We would like to thank first and foremost, Dr. Edward Fox (fox@vt.edu).

We would also like to thank the Teaching Assistant for this course, Liuqing Li (liuqing@vt.edu).

We would like to thank our fellow classmates for the sharing of their ideas, workflows, and in some cases code.

And, finally, all of the presenters and consultants who took time to help us as well:

Xuan Zhang (xuancs@vt.edu)

Srijith Rajamohan (srijithr@vt.edu)

Michael Horning (mhorning@vt.edu)

Matthew Ritzinger (mritzing@vt.edu)

Ziqian Song (ziqian@vt.edu)

References

[1] M. Belica. “jusText heuristic based boilerplate removal tool,” *GitHub*, 5-Mar-2017. [Online]. Available: https://github.com/miso-belica/jusText. Commit ad05130df2ca883f291693353f9d86e20fe94a4e. [Accessed: 28-Nov-2018].

[2] Explosion AI. “spaCy v2.0,” *GitHub*, 2018. [Online]. Available: https://spacy.io/. [Accessed: 28-Nov-2018].

[3] V. Dupras, “num2words,” *GitHub*, 17-Nov-2018. [Online]. Available: https://github.com/savoirfairelinux/num2words. Commit 58613e0a18ad51e5372b22b59e2d304e958a3ec3. [Accessed: 28-Nov-2018].

[4] A. Nagpal, “Word2Number,” *GitHub*, 27-Jun-2017. [Online]. Available: https://github.com/akshaynagpal/w2n. Commit 33aac8a1d71ef1dffd4435fe6e9f998154bcb051. [Accessed: 28-Nov-2018].

[5] R. Řehůřek , “Gensim: topic modelling for humans,” *RaRe Consulting*, 20-Sep-2018. [Online]. Available: https://radimrehurek.com/gensim/summarization/summariser.html. [Accessed: 28-Nov-2018].

References

[6] Python Software Foundation. “7.4. difflib - Helpers for computing deltas,” *Python 2.7.15 documentation*, 08-Nov-2018. [Online]. Available: https://docs.python.org/2/library/difflib.html. [Accessed: 28-Nov-2018].

[7] A. See, “abisee/pointer-generator,” *GitHub*, 09-Jul-2018. [Online]. Available: https://github.com/abisee/pointer-generator. Commit a7317f573d01b944c31a76bde7218bcfc890ef6a. [Accessed: 28-Nov-2018].

[8] Y. C. Chen, “ChenRocks/fast_abs_rl,” *GitHub*, 06-Aug-2018. [Online]. Available: https://github.com/ChenRocks/fast_abs_rl. Commit aebf539107caba5be35720f5d1f9f98989a069e8. [Accessed: 28-Nov-2018].