Outline

1. Project Overview
2. Data Extraction
3. Summarization
4. Frontend
5. Acknowledgements/References
Team Members

Joe Grilli
Front End Main
Communications Main
Communications Backup

Manasi Peta
Communications Main
Back End Backup

Nandha Chokkan
R&D Main
Front End Backup

Aidan Simms
Back End Main
R&D Backup
**Project Overview**

Client - Bipasha Banerjee: PhD student, Natural Language Processing (NLP) researcher

**Problems**

- Extremely long and not searchable
- ETDs are not very accessible
- Minimal metadata
- Hard to discover interdisciplinary work
Backend

50+ STEM ETD PDFs from around the US

Manually segment into separate chapter PDFs

Extract semantically useful text; remove stopwords

Apply relevant tokenizer if needed

Ingest into abstractive summarization model from HuggingFace.co

Output summary text; should be around 1/5 length of input

50+ Non-STEM ETD PDFs from around the US
How? Object Detection (Intro)

Code and Research from ETD “Parsing Electronic Theses and Dissertations Using Object Detection” (Aman Ahuja, Alan Devera, Edward A. Fox, and others)
Finding Summarization Models

Used pretrained text transformation models hosted at https://huggingface.co/
Implementing Summarization Models

❖ Deliverable: 4 summarization models implemented

➢ BART
➢ T5
➢ BigBirdPegasus
  ■ Block Sparse Attention
➢ Longformer-Encoder-Decoder (LED) - Book Summarization
  ■ Longer input ability than most summarization models (~16000 tokens)

https://www.projectpro.io/article/transformers-bart-model-explained/553
Team members wrote the above article taking information from various sources like Dr. Fox’s Virginia Tech and personal website.

Summary:

Dr. Fox has been at Virginia Tech since 1983. He works in the areas of digital libraries, information storage and retrieval, machine learning/AI, and NLP. His works have been cited 20,300 times.
Presented by: Fall 2022 CS5604 teams.

- Ran summarization models and text extraction
- Backend server no longer connected
- Basic design with typos, unintuitive labels, and some bugs
- Different scope

Presentation from Fall 2022 CS5604 teams.
Frontend Changes

- Removed components outside of scope
- Altered page navigation and fixed typos

Implemented page limits

Disabled buttons until file uploaded

Custom page labels
Summary Database

- Used to view summaries produced by pipeline
- Searchable by chapter name or ID
- Display useful metrics like algo used
Future Plans

- Implement summarize functionality for website
- Explore different text extraction methods for cleaner and more complete output
- Eventually implement with VTechWorks
- UI alterations
Acknowledgements

Thanks to Aman Ahuja, Alan Devera, Edward Alan Fox, and others for the research in Object Detection in ETDs.

Thanks to the CS5604 Fall 2022 team guided by Dhanush Dinesh involved in creating the base frontend for our project.
Acknowledgements

Special thanks to Bipasha Banerjee for providing us with the opportunity to work with a great client on such a meaningful project. Also, to Dr. Fox for guiding and informing the process thus far.
References

Parsing electronic theses and dissertations using Object Detection (Aman Ahuja, Alan Devera, Edward Alan Fox, and others)

PDF2Image Documentation (Edouard Belval)

PDFPlumber Documentation (Jsvine)

PyMuPDF Documentation (Jorj X. McKie, Ruikai Liu, Artifex)

Wordninja Documentation (Derek Anderson)

Yolov7 Documentation (Kin-Yiu Wong)
References

**HuggingFace**

**Hugging Face - BART** (Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer)

**Hugging Face - BigBirdPegasus** (Zaheer, Manzil and Guruganesh, Guru and Dubey, Kumar Avinava and Ainslie, Joshua and Alberti, Chris and Ontanon, Santiago and Pham, Philip and Ravula, Anirudh and Wang, Qifan and Yang, Li and others)

**Hugging Face - T5** (Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu)

**Hugging Face - Longformer Encoder Decoder** (Iz Beltagy, Matthew E. Peters, Arman Cohan)