Chapter Summarization

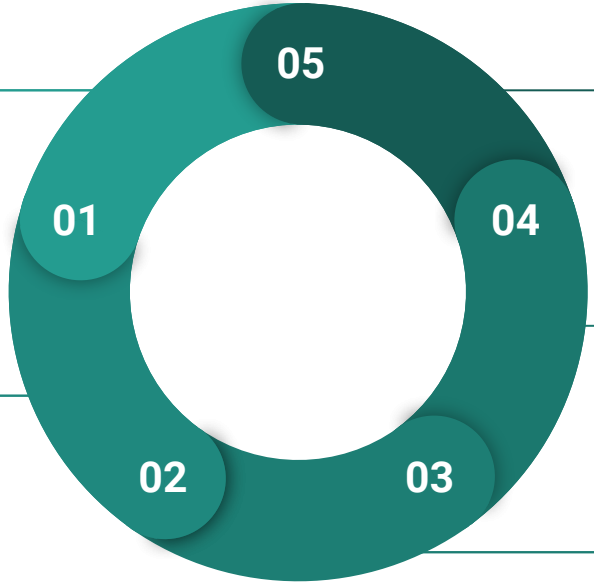
Manasi Peta, Joe Grilli, Aidan Simms, Nandha Chokkan

CS 4624: Multimedia/Hypertext, and Information Access
taught by Dr. Edward A. Fox
Virginia Tech, Blacksburg VA 24061
4/20/2023

Outline

Project Overview

Data Extraction



Acknowledgements/ References

Frontend

Summarization

Team Members

Joe Grilli

Front End Main

Communications Backup

Manasi Peta

Communications Main

Back End Backup

Nandha Chokkan

R&D Main

Front End Backup

Aidan Simms

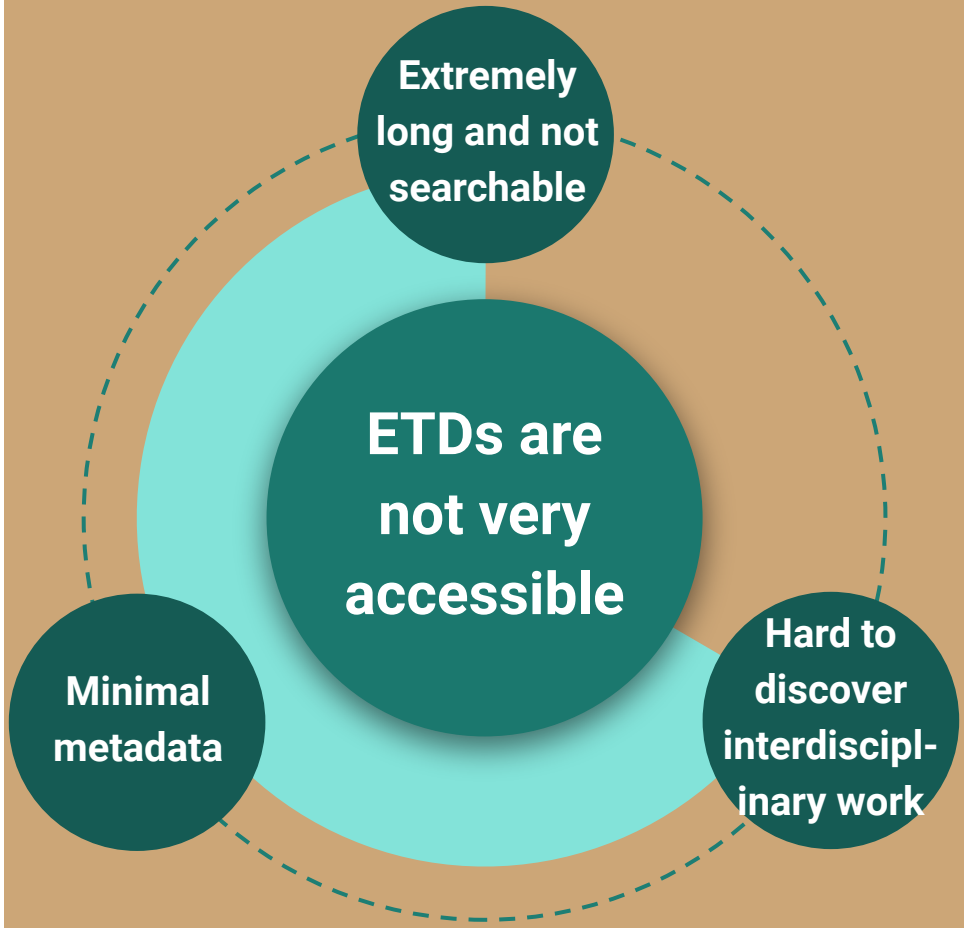
Back End Main

R&D Backup

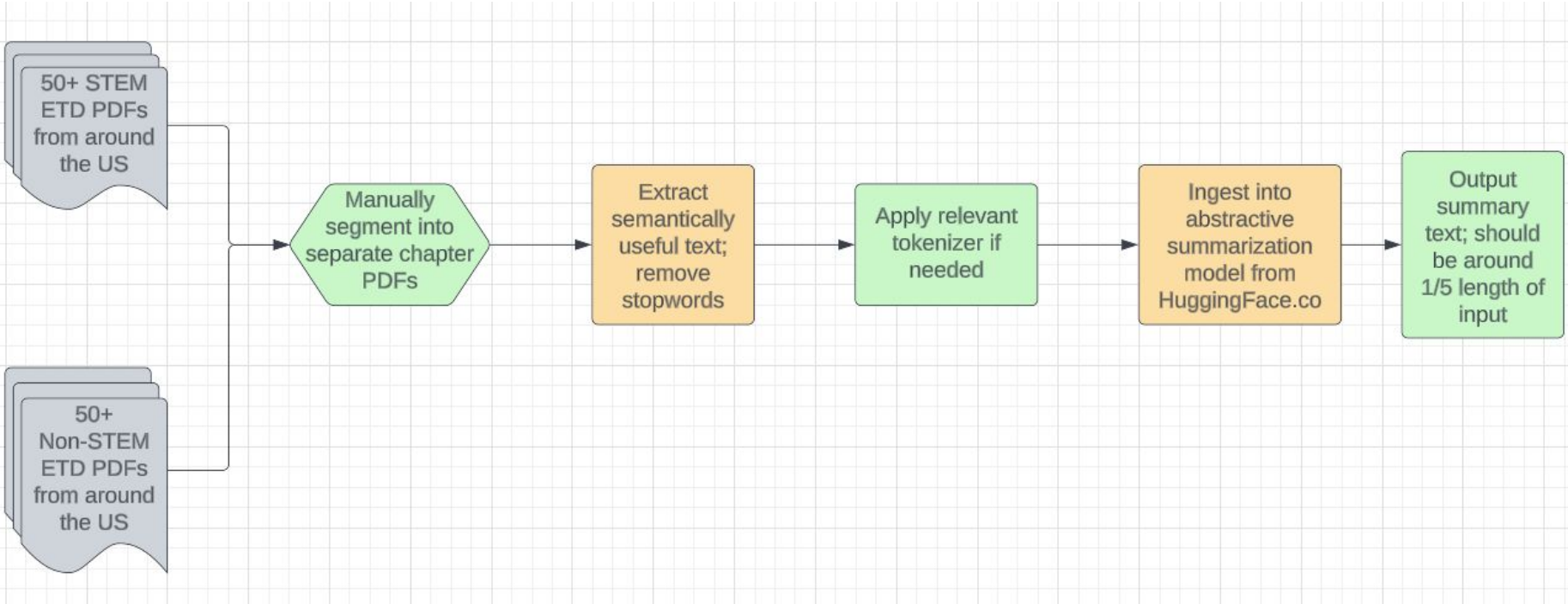
Project Overview

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

Problems

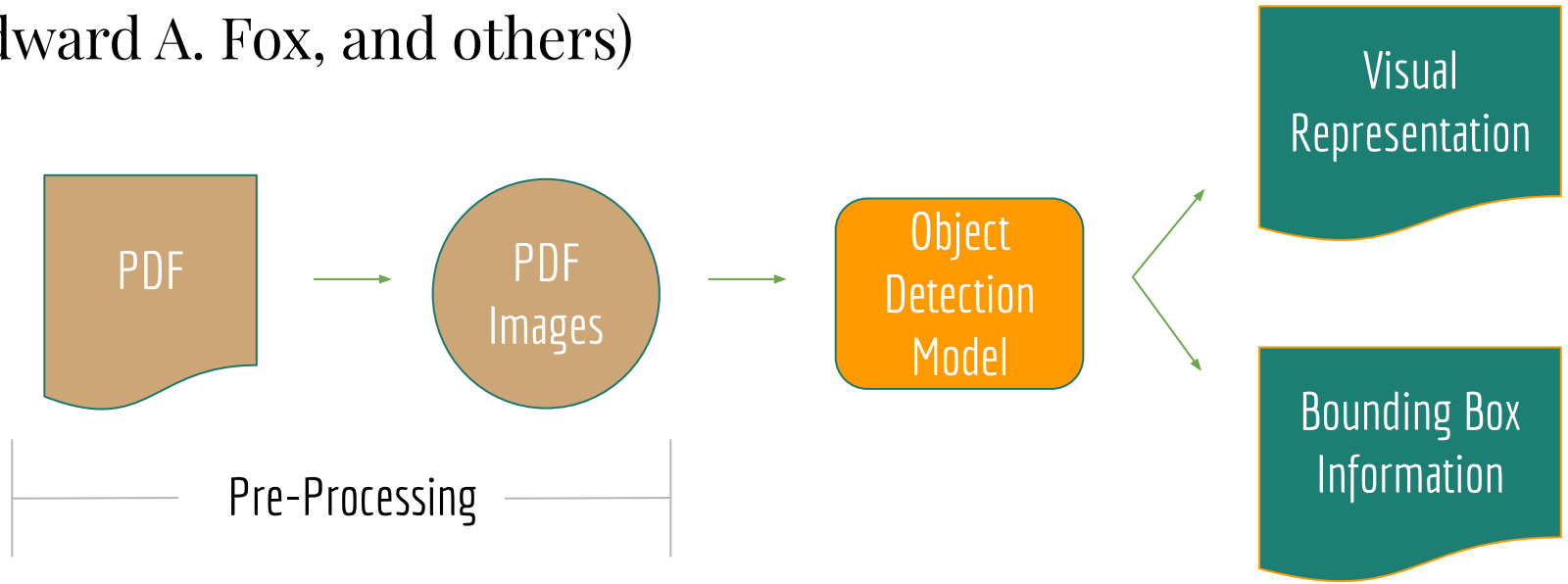


Backend

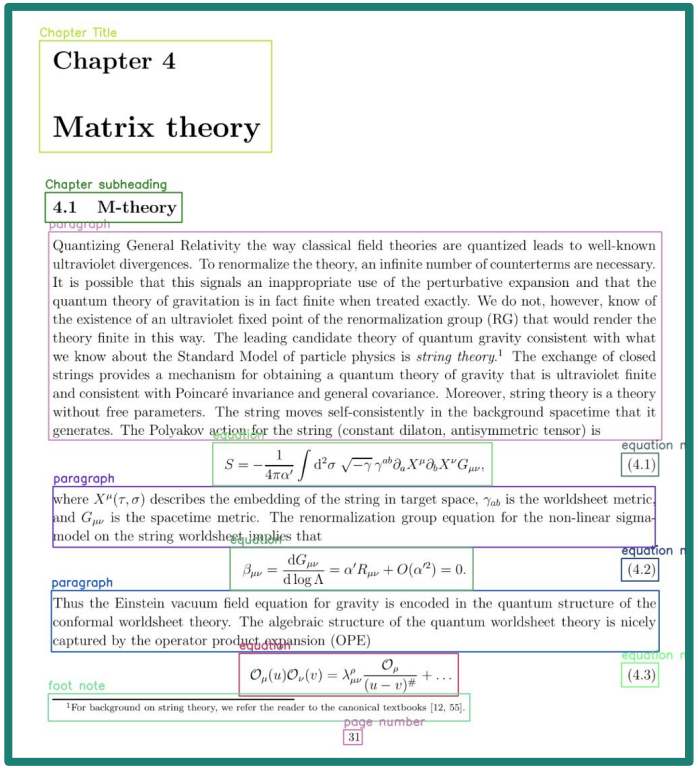

50+ STEM ETD PDFs from around the US
50+ Non-STEM ETD PDFs from around the US
Manually segment into separate chapter PDFs
Extract semantically useful text; remove stopwords
Apply relevant tokenizer if needed
Ingest into abstractive summarization model from HuggingFace.co
Output summary text; should be around 1/5 length of input

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)



Object Detection Output (Partial)



chapter4_marked_output.pdf

ETD: Foundational Investigations & Astronomical Implications of Quantum Gravity (Michael James Kavic)
http://hdl.handle.net/10919/29331

```
xmin,ymin,xmax,ymax,label,page_no
188.3175811767578,1698.3909912109375,1592.609130859375,1839.9718017578125,paragraph,0
182.08834838867188,871.515380859375,1598.4638671875,1351.841064453125,paragraph,0
192.91961669921875,1458.1927490234375,1583.4493408203125,1598.0052490234375,paragraph,0
561.0928344726562,1357.4459228515625,1206.6009521484375,1456.724853515625,equation,0
601.601806640625,1599.8541259765625,1162.9056396484375,1697.5162353515625,equation,0
622.2537231445312,1843.5489501953125,1128.97119140625,1941.720703125,equation,0
1505.6903076171875,1381.119873046875,1591.758544921875,1434.232421875,equation number,0
1505.728759765625,1624.0128173828125,1591.7479248046875,1676.9234619140625,equation number,0
161.81370544433594,430.79217529296875,697.962158203125,687.8905029296875,Chapter Title,0
174.02259826660156,780.2891235351562,490.81622314453125,849.1450805664062,Chapter subheading,0
1505.9725341796875,1865.1124267578125,1591.942626953125,1920.61181640625,equation number,0
181.55751037597656,1936.95556640625,1155.1824951171875,1999.767578125,foot note,0
863.8097534179688,2019.783447265625,906.0684814453125,2053.639892578125,page number,0
85.73263549804688,1220.8145751953125,1523.7064208984375,1995.0465087890625,paragraph,1
95.57185363769531,221.09280395507812,1498.1405029296875,331.8503112792969,paragraph,1
84.30024719238281,570.5612182617188,1523.229248046875,1161.4912109375,paragraph,1
```

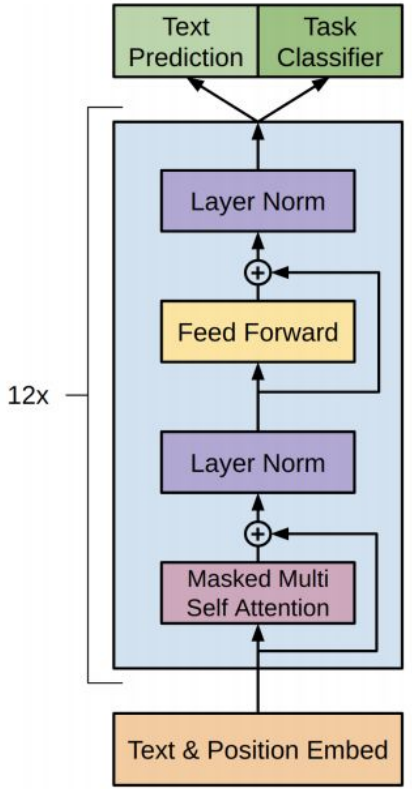
chapter4_bounding_box.pdf

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

Example Summarization Scripts

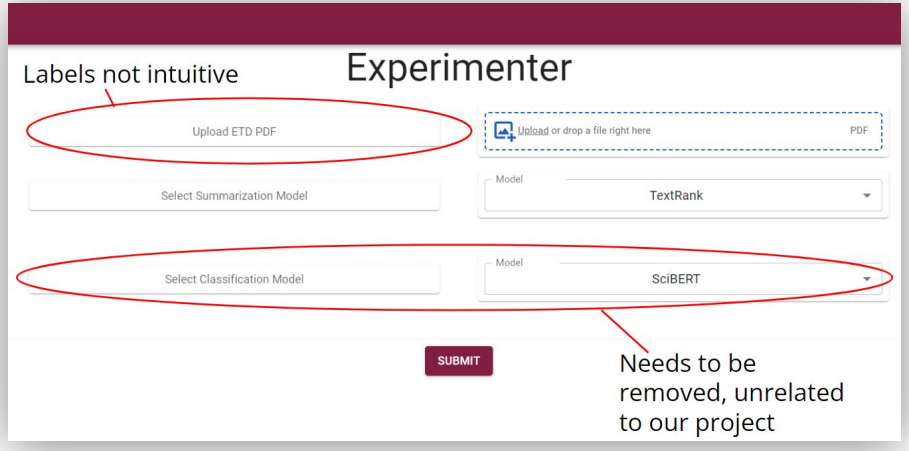
```
from transformers import pipeline

summarizer = pipeline("summarization", model="facebook/bart-large-cnn", device=-1)

ARTICLE = """Dr. Fox is an excellent professor. He has been at Virginia Tech since 1983.
He works in the areas of digital libraries, information storage and retrieval, machine learning/AI,
computational linguistics (NLP), hypertext/hypermedia/multimedia, computing education, electronic publishing,
and expert systems. In these areas, he has participated in, organized, presented, and/or reviewed for
hundreds of conferences and workshops. Over 139 grants have funded his research. He has (co)supervised
through completion over 94 doctoral or masters students. Papers, publications, and presentations total
over 1275: over 88 tutorials in over 28 countries; 19 books; over 144 journal or magazine articles;
over 665 related keynotes, papers, book chapters, posters, demonstrations, and reports; and over 359 other
presentations. Google Scholar reports his works have been cited 20,300 times, with an h-index of 64 and i10-index of 279."""
print(summarizer(ARTICLE, max_length=50, min_length=20, do_sample=False)[0]['summary_text'])
```

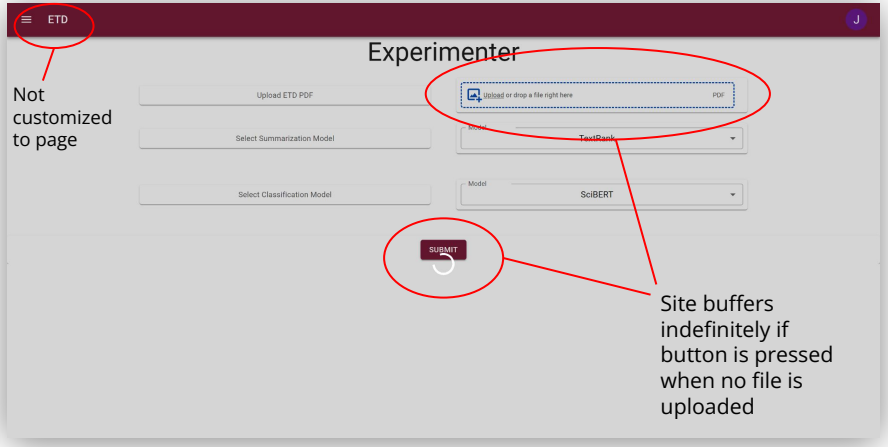
Team members wrote the above article taking information from various sources like Dr. Fox's Virginia Tech and personal website.

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.
```

Original Frontend Work



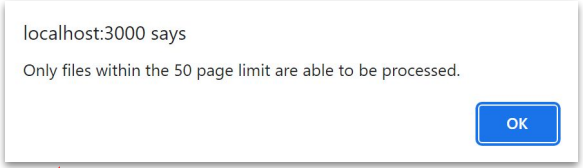


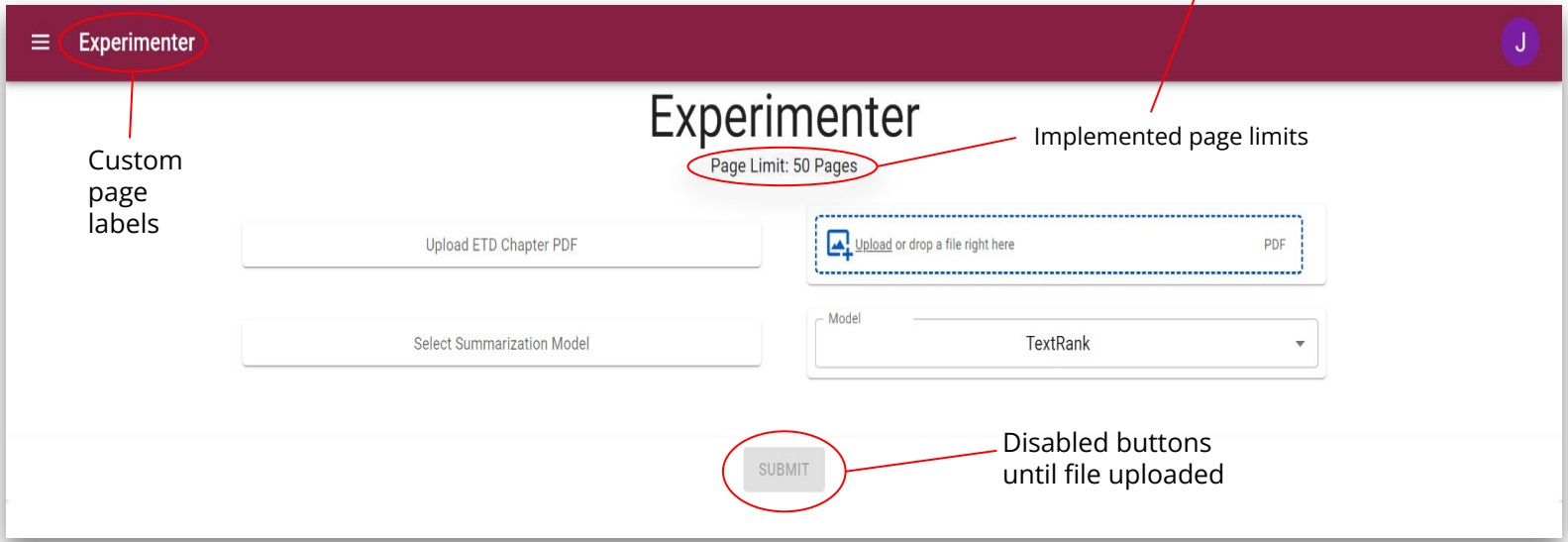
Presentation from Fall 2022 CS5604 teams.

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

Frontend Changes

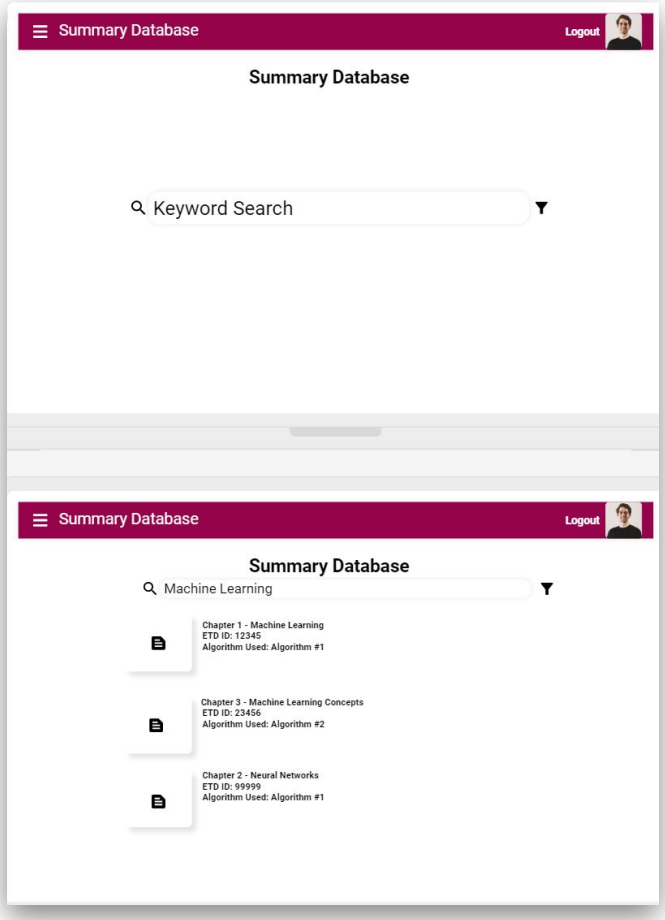
- Removed components outside of scope
- Altered page navigation and fixed typos





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)