

Information Storage and Retrieval

Team: CME

Kulendra Kumar Kaushal, Rutwik Kulkarni, Aarohi Sumant,
Chaoran Wang, Liling Yuan, Chenhan Yuan

CS 5604: Information Storage and Retrieval, Fall 2019

Instructor: Dr. Edward A. Fox

Department of Computer Science, Virginia Tech

Blacksburg, VA 24061

12/05/2019

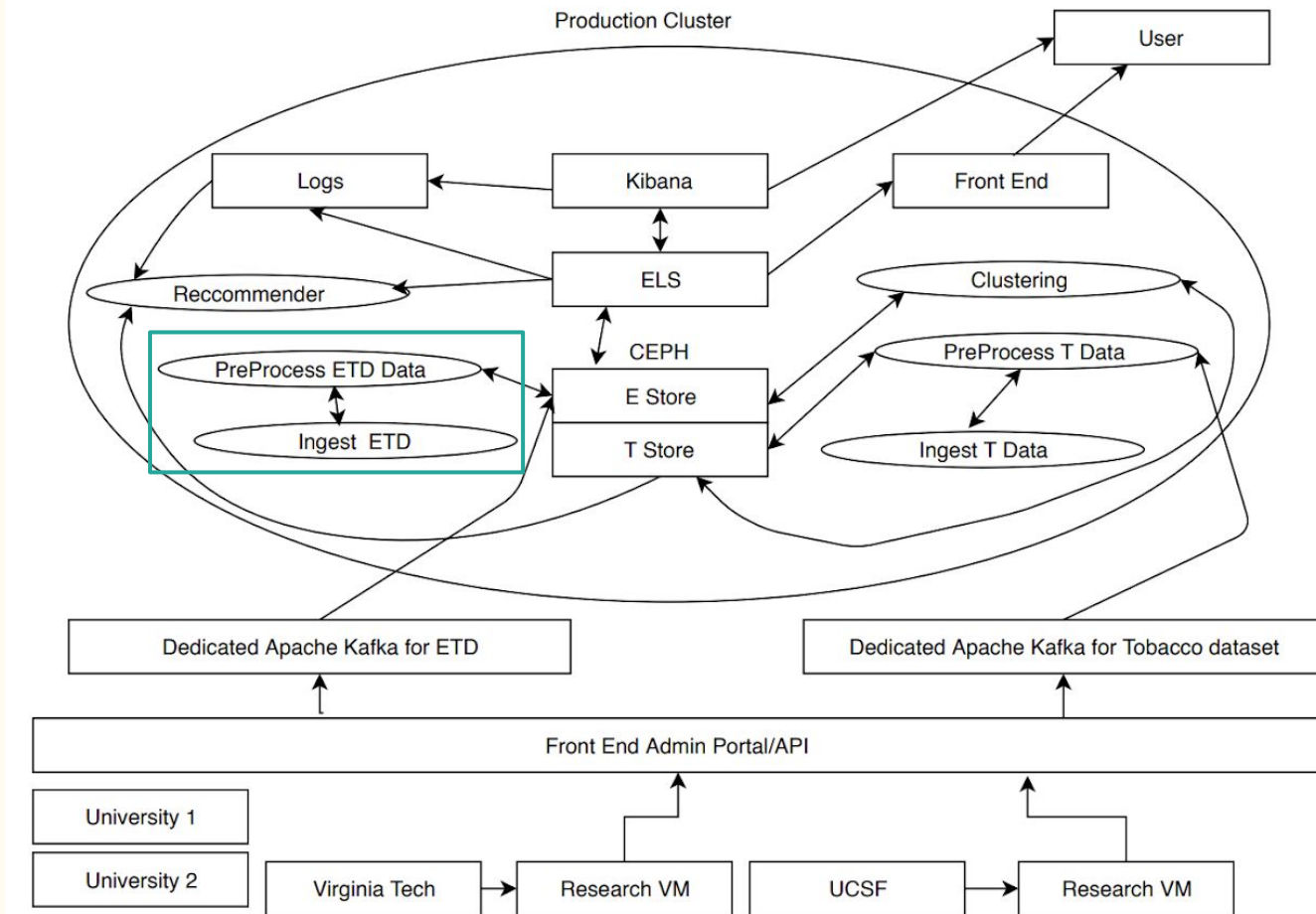
Outline

1. Introduction
 2. Metadata Extraction
 3. Text Extraction and Preprocessing
 4. Chapter Level Text Extraction
 5. Development of an Automated Suite
 6. Contribution
 7. Future Scope
 8. Acknowledgements
-

Introduction



Our position in the system



Research Questions

Our team addressed the problems related to the extraction of metadata and preprocessing from the ETDs by addressing the following research questions:

- **RQ 1:** Can we extract metadata from an ETD document, and transform it into a format that can be ingested into Elasticsearch?
- **RQ 2:** Can we produce text files from PDF files as well as from extracted elements, thus having content suitable for subsequent indexing and searching?
- **RQ 3:** Can we expand the extracted data by including a file for each chapter?
- **RQ 4:** Can we develop an automated system that can extract the metadata from new documents, format it and ingest it into Elasticsearch?

Research different parsers and pick the best ones for this project.

1. Organize metadata.
2. Clean text.
3. Submit visualization requirements.
4. Set up GROBID.

Submit presentation and final report.



09.19.2019

10.10.2019

10.31.2019

11.21.2019

12.05.2019

1. Parse and organize ETD data and put in ceph.
2. Chapterwise extraction.
3. Extract TF-IDF tags.

1. Create an Automated Suite.
2. Upload all scripts on Gitlab.

Metadata Extraction

—

Comparison of Parsers

1. Different parsers can extract metadata and text from an ETD document. Each parser has its pros and cons. We performed a detailed review of the existing parsers for deciding the most accurate parser for our problem statement.
2. The parsers which we studied include:
 - a. GROBID
 - b. Apache TIKA
 - c. Science Parse
 - d. PyPDF2
 - e. PDFMiner

Parser comparison

Parser	Extraction Format	Advantages	Disadvantages
GROBID	XML	<ol style="list-style-type: none">1. Easy to set up2. Structured format3. Can be tuned	<ol style="list-style-type: none">1. Slow parsing2. Not able to extract chapterwise content
Apache TIKA	Text	<ol style="list-style-type: none">1. Can be used for different file formats2. Able to process tables3. Can extract content as well as metadata	<ol style="list-style-type: none">1. It is hard to extract combination of two different types of information
Science Parse	JSON	<ol style="list-style-type: none">1. Structured format2. Detects abstract correctly	<ol style="list-style-type: none">1. Difficult to set up2. Skips or merges chapters3. Skips some references
PyPDF2	Text	<ol style="list-style-type: none">1. Easy to set up2. Extracts text and document information	<ol style="list-style-type: none">1. No output in JSON or XML format
PDFMiner	Text, XML, HTML	<ol style="list-style-type: none">1. Easy to set up2. Compatible with both Py2.x & 3.x	<ol style="list-style-type: none">1. It cannot process table and image

GROBID-Generation Of Bibliographic Data



```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- ... -->
  </teiHeader>
  <text>
    <front>
      <!-- front matter of copy text, if any, goes here -->
    </front>
    <body>
      <!-- body of copy text goes here -->
    </body>
    <back>
      <!-- back matter of copy text, if any, goes here -->
    </back>
  </text>
</TEI>
```

GROBID takes the PDF format of each scientific document as the input and makes use of machine learning models (cascading of linear-chain CRF) for extracting the metadata from the document in XML format. As GROBID could extract **maximum metadata** with **minimum noise** it was selected as the parser for extracting the metadata.

Conversion of metadata into ingestible format

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- ... -->
  </teiHeader>
  <text>
    <front>
      <!-- front matter of copy text, if any, goes here -->
    </front>
    <body>
      <!-- body of copy text goes here -->
    </body>
    <back>
      <!-- back matter of copy text, if any, goes here -->
    </back>
  </text>
</TEI>
```



```
{
  "contributor-author": "Aatique, Muhammad ",
  "date-accessioned": "2011-08-04T21:27:39Z ",
  "date-available": "2011-08-04T21:27:39Z ",
  "date-issued": "1997-08-06 ",
  "identifier-other": "etd-82597-03345 ",
  "identifier-uri": "http://hdl.handle.net/10919/9558 ",
  "description-abstract": "This is abstract ",
  "description-provenance": [
    { "key0": "ETDs_20110726_kdweeks " },
    { "key1": "unrestricted " },
    { "key2": "Made available..." },
    { "Author Email": ["aatique@qualcomm.com " ] },
    { "Advisor Email": ["woerner@vt.edu " ] }
  ],
  "format-medium": "ETD ",
  "publisher-none": "Virginia Tech ",
  "relation-haspart": "aatique.pdf ",
  "rights-none": "I hereby grant to Virginia Tech...",
  "subject-none": ["GPS ", "geolocation ", "DOA ", "AOA "],
  "title-none": "...",
  "type-none": "Thesis ",
  "contributor-department": "Electrical Engineering ",
  "description-degree": "Master of Science ",
  "contributor-committeechair": "Woerner, Brian D. ",
  "contributor-committeemember": ["Tranter, William H. "],
  "identifier-sourceurl": "http:...",
  "date-sdate": "1997-08-06 ",
  "date-rdate": "1997-10-01 ",
  "date-adata": "1997-10-01 ",
  "degree-name": "Master of Science ",
  "degree-level": "masters ",
  "degree-grantor": "Virginia Polytechnic Institute and State University",
  "handle": "9558",
  "searchAuthorStr": "aatique muhammad ",
  "searchTitle": "evaluation ...",
  "text": "This is text"
}
```

- **RQ 1:** Can we extract metadata from an ETD document, and transform it into a format that can be ingested into Elasticsearch?
 1. Comparison of parsers.
 2. Use GROBID to extract metadata.
 3. Convert the extracted metadata into ingestible format.
 4. The data of small subset (2017 ETDs) and larger dataset (all 30K ETDs) is ingested.
 5. Evaluated whether the obtained data is in correct format (to be discussed in the next sections).

Text Extraction and Preprocessing

—

Overview and Need

1. In order to facilitate full text search for the ETDs, extraction of full texts and adding this as a field in the metadata is necessary.
2. Additionally, the Text Analysis and Machine Learning Team needed the preprocessed data for implementing machine learning algorithms.

Text Extraction

Toolkits used to convert PDF to text:

- PyPDF2

APIs are reachable;

- PDFminer.six

Only scripts are offered to extract PDF, no APIs documentation is available.

Still under development

PyPDF2

```
ERROR : TypeError: ord() expected string of length 1, but int found
```

This happens when decoding LZW algorithm

Lempel–Ziv–Welch (LZW) is a compression algorithm based on reference dictionary

```
Line 205      nextbits=ord(self.data[self.bytepos])
```

The decoding processing will probably generate **integer** instead of string.

To address this issue, a new **ord()** function is re-defined (override the built-in)

Garbage values in the extracted text

The full text documents contained a lot of garbage data and redundant numerical values which needed to be removed.

For example:

- b' at the start of the extracted text.
- [23]: Reference numbers are garbage values for clustering
- Figure 4.1: Figure numbers area also irrelevant for clustering
- PDF parsers are not able to extract contents of the image and hence produce garbage value, so such garbage strings are removed,
51sin()kNzzkFmgmx12223242(.)NNNNfFFFF1234(.)
- PDF parsers are also not able to extract the formulae from the document and these formulae appear as garbage and are removed

Preprocessing

1. Used NLTK library in Python to remove stop words.
`stopwords.words('english')`
2. Used regular expression to remove the garbage values and irrelevant numerical data. For example:
 - Replace "..." by ""
 - "[\d{1,20}]"
 - "[\d{1,20}\.\d{1,20}]"
 - Replace "[\(\([\].*?[\)\]\]]" by ""
3. Extracting TF-IDF tags and abstracts separately for clustering



Natural Language Analysis
with Python NLTK

1the four design parameters length, breadth, height distance tip axis rotation. These four parameters run optimization algorithm MATLAB find optimal values satisfy output conditions. The governing equations relating torque applied angle deflection follows Where, (4.7) (4.8) (4.9) Fr normal force applied elastic beam element, θ_b bending angle beam tip, θ_r bending angle shaft, θ_0 initial angle shaft, mr mass object used testing, lm distance object axis shaft, E modulus elasticity elastic beam, I moment inertia beam, R distance axis shaft center pin applying force beam, l length beam δ horizontal distance displaced pin. 33 $\theta \cos() \cos() r b r r r m r F R T m g l \cos() 2 r b b r F l F d E I \cos() 2 \cos() (\cos) \sin 32 r b r r r F l F d R R l R E I$ Figure 4.7: Dynamics Rotary SEA upon applying force/torque To experimentally test RSEA mechanism, test rig built known load applied known location generate torque shaft axis. The deflection measured using bourns 3382G rotary potentiometer. The data read teensy 3.6 microcontroller processed MATLAB. The experimental plot upon observation close linear relationship. Therefore, linear equation curve fitted experimental data given Equation (4.9). But analytical data closely matching experimental result could attribute d several factors like error young's modulus, manufacturing errors friction loss. 34 $\theta_r \theta_b R l \delta d F r$ After applying torque Weight(mrg)lm Figure 4.8: Plot comparison experimental data, fitted data theoretical solution RSEA deflection vs. applied torque

1 four design parameters length breadth height distance tip axis rotation four parameters run optimization algorithm find optimal values satisfy output conditions governing equations relating torque applied angle deflection follows Where normal force applied elastic beam element b bending angle beam tip r bending angle shaft initial angle shaft mr mass object used testing lm distance object axis shaft E modulus elasticity elastic beam moment inertia beam R distance axis shaft center pin applying force beam l length beam horizontal distance displaced pin cos sin : Dynamics Rotary upon applying force/torque experimentally test RSEA mechanism test rig built known load applied known location generate torque shaft axis deflection measured using bourns G rotary potentiometer data read teensy microcontroller processed experimental plot upon observation close linear relationship linear equation curve fitted experimental data given Equation analytical data closely matching experimental result could attributed several factors like error youngs modulus manufacturing errors friction loss r b applying torque Weight lm : Plot comparison experimental data fitted data theoretical solution RSEA deflection vs applied torque

- **RQ 2:** Can we produce text files from PDF files as well as from extracted elements, thus having content suitable for subsequent indexing and searching?
 1. We extracted full text from all 30K ETD documents.
 2. We ingested this text in the metadata to facilitate full text searching.
 3. We preprocessed the data, so that it could be used for training machine learning algorithms.
 4. We extracted the abstracts and TF-IDF tags of all the ETDs to facilitate clustering.

Chapter Level Text Extraction

—

Xpath based Chapter Level Text Extraction

- Various projects have successfully used GROBID for capturing the structure of ETD documents. Therefore, due to previous successful usage and ease of installation, we decided to use GROBID for chapter level text extraction.
- The TEI output format does not explicitly define a chapter tag `<chapter>`, neither does it provide `@type=chapter` attribute for the `<div>` element.
- Therefore, due to the lack of explicit tags for the indication of the start or end of a chapter, chapter level extraction from ETD documents is a difficult task.

Xpath based Chapter Level Text Extraction

“Chapter Name” is generally present in the `<head>` tag which is wrapped inside the `<div>` tag.

XPath expression used:

`/tei:TEI/tei:text/tei:body/tei:div[tei:head]`

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- ... -->
  </teiHeader>
  <text>
    <div xmlns="http://www.tei-c.org/ns/1.0">

      <head><!--Chapter Name --> </head>
      <p> <!-- Chapter Content--></p>
    </div>
    <front>
      <!-- front matter of copy text, if any, goes here -->
    </front>
    <body>
      <!-- body of copy text goes here -->
    </body>
    <back>
      <!-- back matter of copy text, if any, goes here -->
    </back>
  </text>
</TEI>
```

Results

1. Method extracted more chapters
2. Treated subsections as chapters
3. The ETD having 5 chapters was segmented into 15 chapters!!

Name	Date modified	Type	Size
73987.txt	04-09-2019 11:33	Text Document	4 KB
73987.xml	05-10-2019 17:40	XML Document	232 KB
Bailey_JM_D_2017.pdf	04-09-2019 11:46	Adobe Acrobat D...	8,915 KB
Bailey_JM_D_2017.pdf.jpg	04-09-2019 11:33	JPG File	4 KB
Bailey_JM_D_2017.pdf.txt	04-09-2019 11:33	Text Document	200 KB
chapter0.txt	05-10-2019 18:26	Text Document	2 KB
chapter1.txt	05-10-2019 18:26	Text Document	8 KB
chapter2.txt	05-10-2019 18:26	Text Document	16 KB
chapter3.txt	05-10-2019 18:26	Text Document	6 KB
chapter4.txt	05-10-2019 18:26	Text Document	15 KB
chapter5.txt	05-10-2019 18:26	Text Document	6 KB
chapter6.txt	05-10-2019 18:26	Text Document	34 KB
chapter7.txt	05-10-2019 18:26	Text Document	2 KB
chapter8.txt	05-10-2019 18:26	Text Document	20 KB
chapter9.txt	05-10-2019 18:26	Text Document	1 KB
chapter10.txt	05-10-2019 18:26	Text Document	1 KB
chapter11.txt	05-10-2019 18:26	Text Document	4 KB
chapter12.txt	05-10-2019 18:26	Text Document	1 KB
chapter13.txt	05-10-2019 18:26	Text Document	3 KB
chapter14.txt	05-10-2019 18:26	Text Document	1 KB
contents	04-09-2019 11:33	File	1 KB
dublin_core.xml	04-09-2019 11:33	XML Document	8 KB
handle	04-09-2019 11:33	File	1 KB
metadata_thesis.xml	04-09-2019 11:33	XML Document	1 KB

Chapter Level Text extraction based on Table of Contents

Table of Contents provides the page numbers on which a user can find these sections and subsections.

Contents

Abstract	vi
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Overview	1
1.2 VTechWorks ETD Dataset	2
1.3 Problem Definition	3
2 Literature Review	5
2.1 PDF Processing	5
2.1.1 Overview	5
2.1.2 Evaluation of Open-Source Bibliographic Reference and Citation Parsers	5
2.1.3 Big Data Text Summarization	6
2.1.4 GROBID	6
2.1.5 Science Parse	8
2.1.6 Apache Tika	9
2.1.7 PDFMiner	9
2.1.8 PyPDF2	10
3 Requirements	11
3.1 Extract Metadata and Text for ETD corpus	11
3.2 Pre-processing the ETD corpus	12
3.3 User support	12

Chapter Level Text extraction based on Table of Contents- Issues

Chapter 1 Introduction

In order to meet stringent requirements for next generation commercial aviation, aircraft will need to employ more aerodynamic designs to 1) reduce noise, 2) improve landing, takeoff, and cruise emissions, and 3) improve fuel consumption [1,2]. These designs are expected to take the aviation industry away from traditional 'tube-and-wing' architectures into a more hybrid or blended wing body (HWB, BWB) design that will satisfy aircraft operation improvement goals. A comparison between modern and proposed aircraft architectures is shown in Figure 1.1.



Figure 1.1. Modern aircraft 'tube-and-wing' design (Left) and future blended wing design (Right).

A common feature of these HWB airframes is the integration of the engines with the aircraft fuselage. This relocation changes the inlet flow environment as the engines are no longer mounted below the wings where uniform inlet flow is ingested for the majority of the flight envelope. Mounting the engines in close proximity to the airframe, or embedding them within the airframe, results in non-uniform (distorted) inlet flow caused by the airframe body or inlet ducts. In the interest of performance and efficiency, modern commercial transport engines are designed to tolerate only small levels of distortion, such as during takeoff or strong crosswind conditions. Integrated propulsion systems will experience distorted inflow over the entire flight envelope and their interaction with such conditions will need to be studied to determine how engine performance is affected, and to support the design of distortion-tolerant engines.

To determine how well engines can withstand such an environment, three criteria will need to be evaluated in great detail; aeromechanics, stability, and performance. Although studies have shown that engine operation penalties are found in each one of these, the overall benefit of highly integrated airframe and propulsion systems can justify the operational challenges introduced to the engines [4].



Chapter 1 - Introduction 1 Chapter 1 Introduction In order to meet stringent requirements for next generation commercial aviation, aircraft will need to employ more aerodynamic designs to 1) reduce noise, 2) improve landing, takeoff, and cruise emissions, and 3) improve fuel consumption [1,2]. These designs are expected to take the a(cid:89)iation industry(cid:82) a(cid:90)a(cid:92) from traditional (cid:181)tube-and-(cid:90)ing(cid:182) architectures into a more hybrid or blended wing body (HWB, BWB) design that will satisfy aircraft operation improvement goals. A comparison between modern and proposed aircraft architectures is shown in Figure 1.1. Figure 1.1. Modern aircraft (cid:181)tube-and-(cid:90)ing(cid:182) design (Left) and future blended wing design (Right). A common feature of these HWB airframes is the integration of the engines with the aircraft fuselage. This relocation changes the inlet flow environment as the engines are no longer mounted below the wings where uniform inlet flow is ingested for the majority of the flight envelope. Mounting the engines in close proximity to the airframe, or embedding them within the airframe, results in non-uniform (distorted) inlet flow caused by the airframe body or inlet ducts. In the interest of performance and efficiency, modern commercial transport engines are designed to tolerate only small levels of distortion, such as during takeoff or strong crosswind conditions. Integrated propulsion systems will experience distorted inflow over the entire flight envelope and their interaction with such conditions will need to be studied to determine how engine performance is affected, and to support the design of distortion-tolerant engines. To determine how well engines can withstand such an environment, three criteria will need to be evaluated in great detail; aeromechanics, stability, and performance. Although studies have shown that engine operation penalties are found in each one of these, the overall benefit of highly integrated airframe and propulsion systems can justify the operational challenges introduced to the engines [4].

Courtesy Boeing®

Courtesy NASA

Chapter 1 - Introduction 1 Chapter 1 Introduction In order to meet stringent requirements for next generation commercial aviation, aircraft will need to employ more aerodynamic designs to 1) reduce noise, 2) improve landing, takeoff, and cruise emissions, and 3) improve fuel consumption [1,2]. These designs are expected to take the a(cid:89)iation industry(cid:82) a(cid:90)a(cid:92) from traditional (cid:181)tube-and-(cid:90)ing(cid:182) architectures into a more hybrid or blended wing body (HWB, BWB) design that will satisfy aircraft operation improvement goals. A comparison between modern and proposed aircraft architectures is shown in Figure 1.1. Figure 1.1. Modern aircraft (cid:181)tube-and-(cid:90)ing(cid:182) design (Left) and future blended wing design (Right). A common feature of these HWB airframes is the integration of the engines with the aircraft fuselage. This relocation changes the inlet flow environment as the engines are no longer mounted below the wings where uniform inlet flow is ingested for the majority of the flight envelope. Mounting the engines in close proximity to the airframe, or embedding them within the airframe, results in non-uniform (distorted) inlet flow caused by the airframe body or inlet ducts. In the interest of performance and efficiency, modern commercial transport engines are designed to tolerate only small levels of distortion, such as during takeoff or strong crosswind conditions. Integrated propulsion systems will experience distorted inflow over the entire flight envelope and their interaction with such conditions will need to be studied to determine how engine performance is affected, and to support the design of distortion-tolerant engines. To determine how well engines can withstand such an environment, three criteria will need to be evaluated in great detail; aeromechanics, stability, and performance. Although studies have shown that engine operation penalties are found in each one of these, the overall benefit of highly integrated airframe and propulsion systems can justify the operational challenges introduced to the engines [4].

1.1 Literature Review Studies have been performed on the evaluation of engines subject to non-uniform inlet flows produced by highly integrated engine-airframe systems. The three metrics that must be most considered are; aeromechanics, stability, and performance. Researchers can find background into aeromechanical fatigue caused by distorted inflow through earlier publications [5, 6]. As this work focuses on the impact of stability and performance, the literature review will investigate previous work on these subjects in detail. An important component of the integrated engine-airframe problem is the issue of fan-distortion interaction, in which a coupling effect between the distortion or distortion producing device and the fan exists. As will be shown later, these components act together as a single system that must be considered when designing distortion-producing devices within the context of distortion tolerant fan research. 1.1.1 Stability Stability is a qualitative measure of how well an engine can handle an off-design

Manual Chapter Level Extraction

We finally did a manual chapter level extraction from 20 ETD documents.

This gives the gold standard result.

Differences of number of chapters

Document	XPath	Manual	Match
73987	15	5	No
73988	9	7	Close
74003	52	5	No
74047	3	1	
74048	36	5	No
74049	46	5	No
74050	75	5	No
74233	5	5	Yes
74234	40	7	No
74235	12	5	No
74236	31	6	No
74237	23	5	No
74238	2	5	No
74239	154	7	No
74275	13	ETD in slides format	
74302	50	7	No
74383	85	5	No
74395	21	5	No
74396	3	1	
74398	0	1	
74423	31	6	No

Evaluation

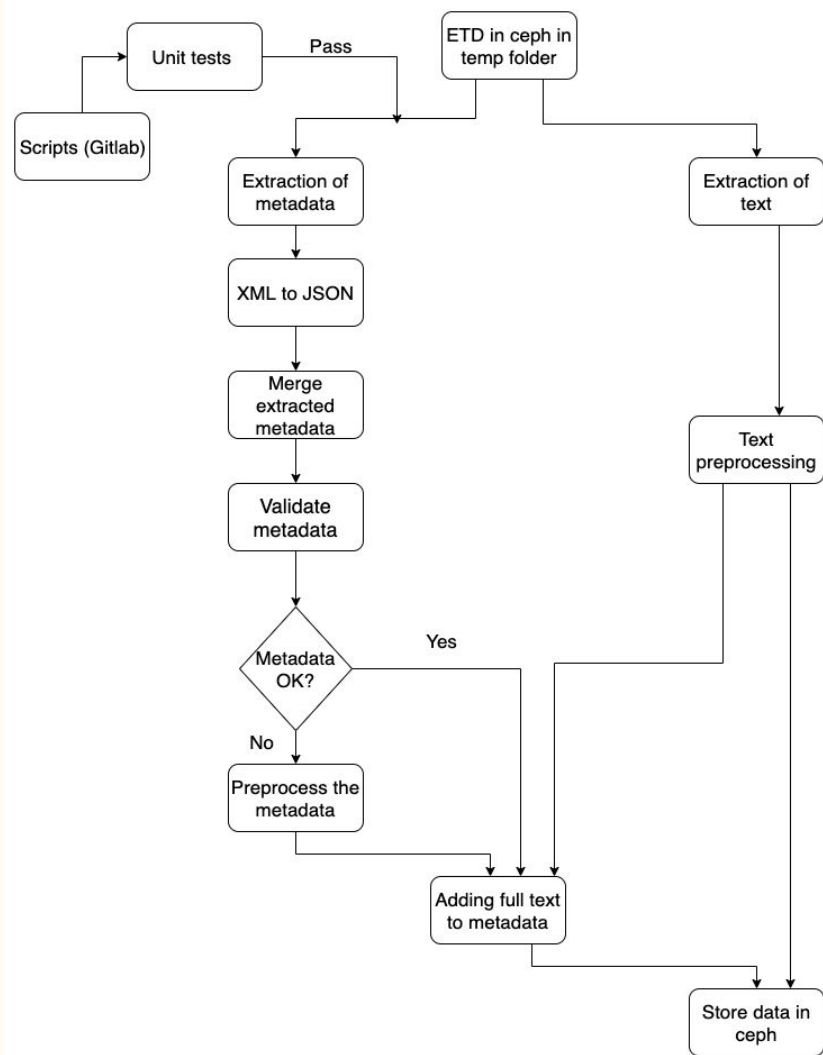
	XPath	Manual
Chapter completeness on average	43.90%	90.88%
Formulas	No	Yes but lots of illegal characters
Illegal characters	No	Some letters are converted to {cid:}
References in-text	No	Yes
References	No	Yes
Texts in figures	No	Yes but many illegal characters

- **RQ 3:** Can we expand the extracted data by including a file for each chapter?











1. Tried Xpath based and table of contents based method for chapter level text extraction
2. Evaluated these methods against gold standards (manual chapter level text extraction)
3. This result can be used for big data summarization problem
4. There is a scope for improvement

Development of an Automated System

Flowchart



Script Description

Name	Last commit	Last update
 AddTextToMetadata.py	Initial commit	1 week ago
 DataPreProcessing.py	Initial commit	1 week ago
 DriverScript.py	Initial commit	1 week ago
 MergeMetaData.py	Initial commit	1 week ago
 MetadataExtractor.py	Initial commit	1 week ago
 TextExtractor.py	Initial commit	1 week ago
 XML2JSONConverter.py	Initial commit	1 week ago
 __init__.PY	Initial commit	1 week ago
 config.py	Initial commit	1 week ago
 util.py	Initial commit	1 week ago

Features and Advantages

1. Unit tests to test the scripts (e.g., check whether GROBID is running or not)
2. Automatic extraction of the new metadata and merging it to the existing file
3. Unit tests to validate the metadata
4. Extraction of text that can be used by Text Analysis and Machine Learning Team

```
grobid running
Path Correct
Path Correct
Path Correct
Path Correct
Path Correct
Path Correct
Grobid Running and XML output path Correct
Grobid Running and XML output path Correct
Grobid Running and XML output path Correct
Grobid Running and XML output path Correct
Grobid Running and XML output path Correct
C:/Users/hp/Downloads/ETDSummarizationSourceCode.tar/ETDSummarizationSourceCode/pipeline/Etd Data/test_sub/73987\Bai
ETD found
C:/Users/hp/Downloads/ETDSummarizationSourceCode.tar/ETDSummarizationSourceCode/pipeline/Etd Data/test_sub/74003\Bate
ETD found
C:/Users/hp/Downloads/ETDSummarizationSourceCode.tar/ETDSummarizationSourceCode/pipeline/Etd Data/test_sub/74048\Lee_
ETD found
C:/Users/hp/Downloads/ETDSummarizationSourceCode.tar/ETDSummarizationSourceCode/pipeline/Etd Data/test_sub/74049\Pain
ETD found
C:/Users/hp/Downloads/ETDSummarizationSourceCode.tar/ETDSummarizationSourceCode/pipeline/Etd Data/test_sub/74050\Wyga
ETD found
-----
Ran 4 tests in 254.392s
```

Limitations and Assumptions

1. Cannot scrape the data from VTechWorks or ir.cs.vt.edu (the new data should be added in a folder called “temp:” on ceph).
2. The folder structure of ETD document should be the same as we are currently using.

- **RQ 4:** Can we develop an automated system that can extract the metadata from new documents, format it and ingest it to Elasticsearch?
 1. Such an automated system which will automate the entire project done by CS5604 class can be developed.

Our contribution

1. Full text extraction and preprocessing
2. Metadata extraction and conversion into ingestible format
3. Automation suite for adding new documents

Future Scope

—

1. Improve chapter level text extraction.
2. Batch metadata processing of ETD documents.

Acknowledgements

—

This project was completed over the course of CS5604: Information Storage and Retrieval at Virginia Tech. ETD data analyzed was provided by Virginia Tech University Libraries, and the students who have submitted theses or dissertations in connection with their Virginia Tech studies.

The authors would like to thank Dr. Edward A. Fox, Ziqian Song, and our classmates for providing insight and expertise that greatly helped us accomplish this research project.

We thank Bipasha Banerjee for her help and insights on ETD processing.

We thank the other teams -- TML, CMT, FEK, INT, and ELS -- for their valuable help and coordination.

Last but not least, we are immensely grateful to the creators of all the open source software we used to create this project. Thank you very much!

Thank you!

Any Questions?

Team: CME

Kulendra Kumar Kaushal, Rutwik Kulkarni, Aaroahi Sumant,
Chaoran Wang, Liling Yuan, Chenhan Yuan