Shraddha Kulkarni (6117955457)
Bharath Chandra Thota (7232807134)

PROJECT PROPOSAL

Engrapho – Search App

What does your app do?

- Engrapho is basically a Web Application, that lets the user search on any topic, and as the search Results all kinds of documents (word, ppt, excel, pdf) which are related to that particular search term will be presented. Additionally, the user interface will consist of a Filter section (Faceted Search) for the most appropriate results.

Where do you plan to obtain the documents?

- Google Scholar, SlideShare, Springer, LinkedIn, Elsevier.

What metadata do you plan to extract?

- Standard Properties: Title, Author, Subject, Keywords, Comments, File size, References
- Dynamically updated Properties: No. of pages, paragraphs, Lines, words
- For Word Docs only: Table of Content.

https://stackoverflow.com/questions/14209214/reading-the-pdf-properties-metadata-in-python

Where (i.e., which cloud database) do you plan to store the metadata?

- AWS S3

How do you plan to implement programs for extraction and uploading of extracted metadata to the cloud database?

- To extract the metadata, we plan on using the third-party modules pypdf2, eyeD3 for pdfs and mp3 respectively and use **requests** module to upload and query the metadata.
- The backend system, implemented in Flask, is responsible for uploading and querying the server for metadata.

What kind of user interface do you plan to implement (web browser, Android, iOS)?

- Web Application

Which programming languages and software libraries will you use?

- Languages, Frameworks, and Tools: Python, Flask Framework, AWS S3
- Web Dev Technologies: HTML5, CSS3, Javascript, jQuery, Bootstrap
- Database and APIs: MongoDB, Youtube API, AWS API
- IDE: Visual Studio Code, Sublime Text
- Metadata Extraction Tool: http://meta-extractor.sourceforge.net/
- Other Tools: https://www.forensicswiki.org/wiki/Document_Metadata_Extraction

Shraddha Kulkarni (6117955457)
Bharath Chandra Thota (7232807134)

Group formation: who are in your group? What is each person's responsibility? Is your group equipped to implement the application by the end of the semester?

- Bharath Chandra Thota (7232-8071-34) & Shraddha Kulkarni (6117-9554-57) (1st Year Data Informatics Graduate Students)
- Both will work on creating html templates for web pages. Will add Bootstrap, Javascript & jQuery snippets for better look and feel.
- Bharath will be responsible for plugging in these web templates into the Flask framework. Besides that, he'll be working on code to extract metadata from the Documents.
- Shraddha will be handling the Database section, i.e. connecting the AWS S3 Storage and MongoDB, querying for fetching data and loading it into the web templates.
- The project idea seems to be feasible. Once we are done collecting all the document files, we will write code for fetching it from the NoSQL Database and displaying onto the webpage. Most of the tools required like MongoDB, AWS S3, Flask, is already set up.
- Both of us have well equipped laptops with 8GB of RAM & about 1TB HDD/256GB SDD. Plus, we could use the AWS Cloud services in case we need more computation power.

Types of Metadata:

File Type	Metadata
MP3 Files	Song Name, Artist, Album, Year, Comment, Track, Genre, Band, Supports both ID3TagV1 and ID3TagV2
Word Files	File Name, Author, Subject, Keywords, Comments, File size, References, No. of Pages, Paragraphs, Lines, Words, Table of Contents
PPT Files	File Name, Author, Subject, Keywords, Comments, File size, References, No. of Slides
PDF Files	File Name, Author, Creator, Creation Date, Modified Date, Producer, File size

Milestones: a project timeline with milestones.

Time	Activity	Description
September 9, 2018	Submit Proposal	Submission of project proposal.
September 9, 2018 – October 5, 2018	Datasets development	Collecting the data (Pdfs, Word, Excel, ppts, Images, Mp3) from which metadata is to be extracted.
October 6, 2018	Submit the Status	Report the work done till date.
October 7, 2018 – November 17, 2018	Development	Developing web pages, backend and metadata extraction.
November 18, 2018 – November 27, 2018	Testing	Testing the application.
November 30, 2018	Submit Final Status	Submission of final report.

Sample Code for PDF and MP3 Meta Data Extraction:

```
1  from PyPDF2 import PdfFileReader
2  from mp3_tagger import MP3File
3
4  print("PDF Extraction")
5  pdf_to_get = PdfFileReader(open('testdocs.pdf','rb'))
6  pdf_info = pdf_to_get.getDocumentInfo()
7  print(str(pdf_info))
8  print("MP3 Extraction")
9  mp3 = MP3File('DearJhon.mp3')
10 print(mp3.get_tags())
```

Sample Execution:

```
F:\Python Files>python meta_data_extractor.py
PDF Extraction
{'/Author': 'Bharath Chandra Thota', '/Creator': 'Microsoft® Word 2016', '/CreationDate': "D:20180903122207-07'00'", '/ModDate': "D:20180903122207-07'00'", '/Producer':
 'Microsoft® Word 2016'}
MP3 Extraction
{'ID3TagV1': {'song': 'Dear John(Taylor Swift)', 'artist': 'Taylor Swift', 'album': 'Taylor Swift', 'year': '2010', 'comment': 'Downloaded From www.krazywap', 'track':
99, 'genre': None}, 'ID3TagV2': {'artist': 'Taylor Swift', 'band': 'Taylor Swift', 'album': 'Taylor Swift', 'song': 'Dear John(Taylor Swift)', 'comment': 'XXX\x00Downlo
aded From www.krazywap.com visit now for latest music, games and softwares', 'year': '2010'}}

F:\Python Files>
```