Engrapho – Search App

**What does your app do?**

* Engrapho is basically a web Application, that lets the user search on any topic, and all kinds of documents (word, ppt, excel, pdf) which are related to that particular search term will be presented as the search results. Additionally, the user interface will consist of a Filter section 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?**

**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>

**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 & Shraddha Kulkarni (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 4-8GB of RAM & about 1TB HDD. Plus, we could use the AWS Cloud services in case we need more computation power.

**Milestones: a project timeline with milestones.**