EE/CprE/SE 491 - sddec22-19
Text extractions from documents into Elasticsearch
Week 3 Report

2/14/22 - 2/20/22
Client: Kayla Gilleland, Buildertrend
Faculty Advisor: Goce Trajcevski

Team Members:

Bruce Bitwayiki
Jared Hayashi
Rushal Sohal
Tiffany Mayberry

Weekly Summary

This week's focus was to collect and organize our research in an easily accessible location for all team members. We brainstormed different tech stack options to implement the project. We created a list of clarifying questions that we believe will help better understand the goals. Gitlab issues were created to keep track of our project tasks such as our design decisions.

Past Week Accomplishments

- Setup shared docs in Google Drive for shared documentation of our research
  - Includes research into ElasticSearch and Tika
  - Also includes notes for past meetings
- Team Professionalism Assignment
  - Examined how the ACM code of ethics applies to our project
- Created Gitlab issues for document parsing, front end app, and design decisions. These include the following tasks
  - Experiment with Elasticsearch and possibly deploy a single-node instance
  - Experiment with Apache Tika parse Microsoft Office documents (docs, pdfs, powerpoint, excel) and examine the output.
- Brainstorm different tech stack options
  - We brainstormed different options for the frontend and backend of our application. We plan to look into the pros and cons of each options this upcoming week.
  - For the front end, we are considering different combinations of web frameworks and different Java and Python GUI libraries.
  - For the backend, we are considering java and python packages of Tika.

Individuals Contributions

Name	Individual Contributions	Hours This Week	Hours Cumulative
Bruce Bitwayiki	Initial gitlab issues (TIKA parsing and front end app)	1	5
Jared Hayashi	Documented Apache Tika notes to communal doc	1	3
Rushal Sohal	Some researching into lucene and documentation	1	3
Tiffany Mayberry	Brainstormed different tech stacks, uploaded past reports to website	5	10

Plans for the Upcoming Week

- Send follow up email to Kayla regarding Teams and questions about use cases/example datasets
    - Figure out what platform this project will be used on (e.g. website or desktop application)
    - What info from the files will be displayed by the application
    - What info will the users use to query and filter the files (e.g. file type, title, date, text body)
    - See the “Scope Questions” section at the end of this document
- Prototype system design and architecture
    - Use case diagram?
    - Block/class diagram?
    - Gantt chart?
- Experiment with Elasticsearch

Summary of Weekly Advisor Meeting

- Need to create document containing questions that we may have about use cases and the specifics of the project
- Should think about sketching design plans and system architecture

Project Scope Questions

1. Who will be doing the queries? Contractors? Customers? Builders? Customer Support?
2. What data is needed to be stored in elasticsearch?
    1. Metadata - what parts of metadata?
    2. File itself?
    3. View permissions? Read-Only? Editable?