EE/CprE/SE 492 - sddec22-19

Text extractions from documents into Elasticsearch

Report 13

10/26/22 - 11/08/22
Client: Kayla Gilleland, Buildertrend
Faculty Advisor: Goce Trajcevski

Team Members:

Bruce Bitwayiki
Jared Hayashi
Rushal Sohal
Tiffany Mayberry

Weekly Summary

The focus of the past two weeks has been further extending the communication between individual components. We were able to get past a big hurdle of not being able to get our text extractor component to communicate with Elasticsearch. We spent time researching and expanding our search implementation to search by keywords and filter based on file type. In addition our upload feature is now able to upload files onto our server. We also spent time putting together and presenting our second update of the semester in class.

Past Week Accomplishments

- Tika is now able to send JSONs to Elasticsearch
- Tika now detects and extracts metadata based on the file type and constructs a JSON from it
- Users can now upload a file or multiple files. The uploaded files are stored on the server for now.
- Users are able to search by keywords (instead of a full exact content match) and add an additional filter to limit the search even more based on the file type

Individuals Contributions

Name	**Individual Contributions**	**BiWeekly Hours**	**Hours Cumulative**
Bruce Bitwayiki	ElasticSearch analyzer research & implementation	7	50
Jared Hayashi	File specific JSON construction and communication between Tika and Elasticsearch now working	7	52

Rushal Sohal	Upload functionality done with styling and server setup	9	54
Tiffany Mayberry	Extended search querying to be able to filter by file type and search by non exact keywords	8	71

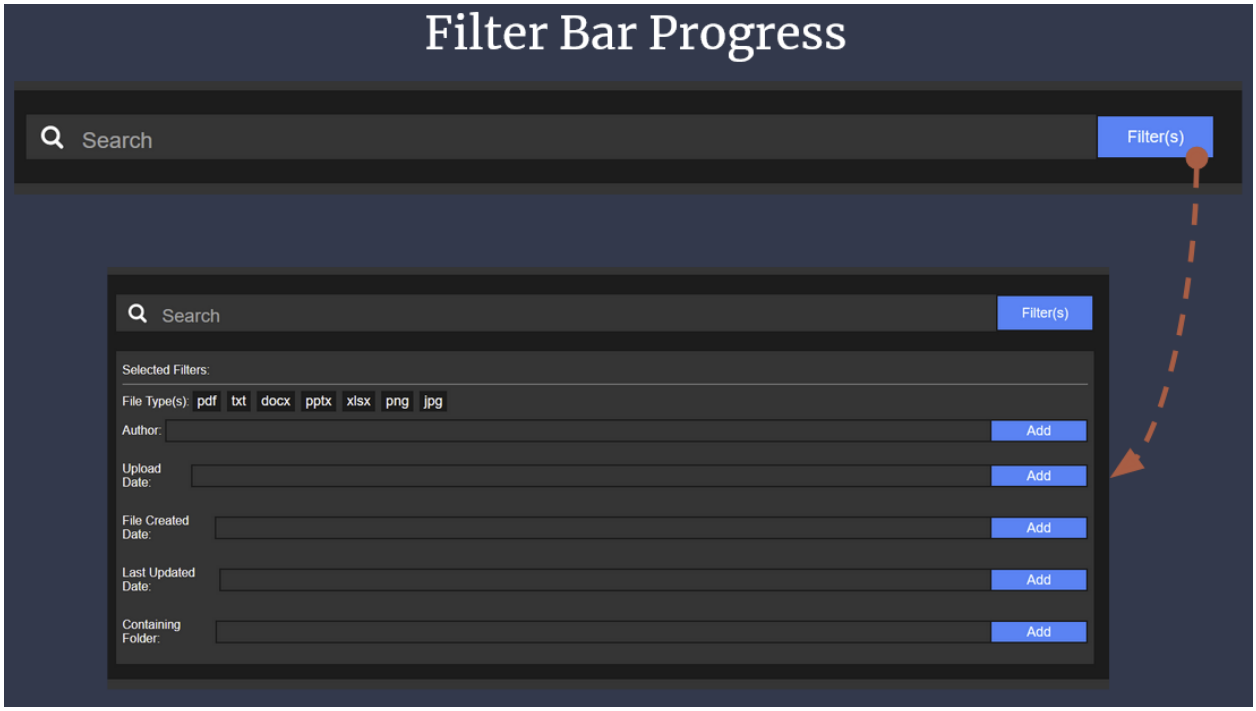
Plans for the Upcoming Weeks

- Tika OCR for text extraction from images
- Spring server for Tika to receive files from the frontend
- Getting the UI together and testing
- Deploy UI onto server
- Create a video demoing of our biweekly progress to replace our client meeting

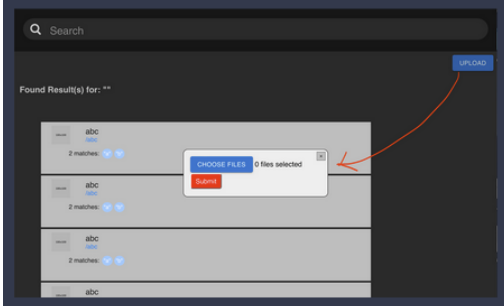
Summary of Weekly Advisor Meeting

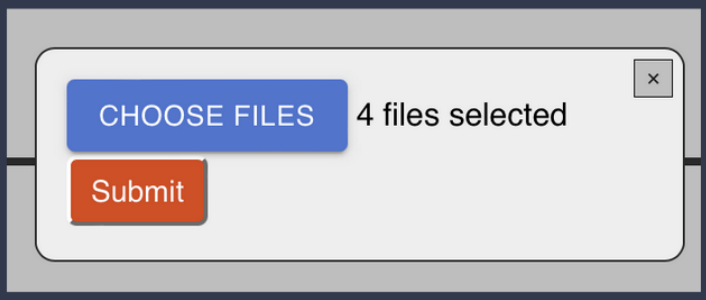
- Canceled due to faculty sickness and no major topics to discuss

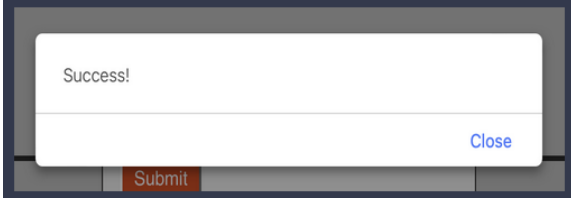
Highlights from InClass Presentation

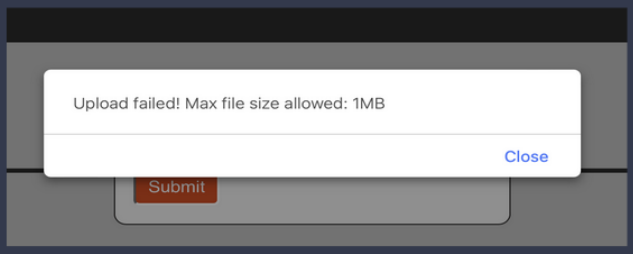


Upload Progress









ElasticSearch keyword + file filter query