

BuilderTrend Text Extraction

Jacob Kirshenbaum, Kyle Maynard, Pajie Bimpolo Maouene, Edward Niemann

Our project is a system that will make it possible for BuilderTrends' customers to search through documents they have uploaded. We will be researching and implementing a tool that will be able to extract text from many different file types. We will then take the text that was extracted from the documents and put it into a framework that is able to store the data. This framework will also act as a search engine for us to properly allow the text to be searched for. A simple user interface will also be developed in order to demonstrate the search feature.

This group is composed of four group members. Jacob Kirshenbaum is the project manager of our team. Kyle Maynard is the client liaison and the quality assurance/testing lead. Pajie Bimpolo is the requirements analyst and the lead researcher. Edward Niemann is the Architect/Technical lead and the scribe. The division of work for our group will be as equal as we are able to make it. Once we are able to meet with our sponsor to get more details we will discuss how the work will be divided.

There is not a huge deal of risk associated with the application we will be developing as it seems that it will primarily be used by employees of BuilderTrend. If that ends up not being the case then we will have risk associated with ensuring the security of the documents that we are scanning. We would need to ensure that no unauthorized individuals are given access to documents that were scanned through our application as there could potentially be some sensitive information within the documents that would be scanned.

As of right now, the only hardware requirement we have is that we will have to have computers that we are able to program on. Aside from that we will need any compatible software that will work as an IDE for the development of our system. Any tools or frameworks we will be

using will need to be compatible with the .NET framework as that is how we intend to implement our system.

Our work breakdown structure as of right now is mainly just based around research until we have more detail about the project. Each team member will be looking into programming in C# along with the Apache Tika and Elasticsearch frameworks. Once we have more details we will plan out how we want to develop the system and break it down from there.

Our work breakdown structure as of right now is split into five parts but is subject to change as we begin implementation of our application. The five parts will be implementing Apache Tika, implementing Elasticsearch, implementing a simple UI for the application, testing our application, and investigation of other tools similar to Apache Tika.

We intend to have our application fully functional within five milestones. For the first milestone we would like to have basic functionality of Apache Tika and Elasticsearch implemented in our application. For the second milestone we intend to have a simple UI to show the functionality of the application along with some further implementation of Apache Tika and Elasticsearch. By the third milestone we would like to have a fair amount of our testing of the application done along with most of the functionality for the application done. For the fourth milestone we will be putting finishing touches onto the application while also looking into the stretch goals that the sponsor has provided. We would also by this time like to look at implementing another tool like Apache Tika to show the benefits of a different tool. By the fifth milestone we would like to have at least two tools functionally scanning documents and a completely working application.

We will be using Trello as our project tracking system. This will act as a way for us to split up tasks and keep track of what everyone is working on. Github will be our version control

system as it will work the best for our needs. Also, we are using Google Drive to share project documents. We are using a discord server to coordinate with each other on what we are working on. We will begin having biweekly meetings with BuilderTrend on Tuesdays at 4pm on 2/22/2022. Our sponsors will also be creating a Teams group for us to communicate with them in. We do not currently have a set time that we meet with each other outside of class other than typically communicating after the class time on Mondays and Wednesdays.

Trello: <https://trello.com/invite/buildertrendtextextraction/c477b242004e70a51a05bcb7235c634d>

Github repository: <https://github.com/jacobkirshenbaum/BuilderTrend-Text-Extraction>

Google Drive has been shared with you.