Automated Document Summarization using Natural Language Processing

Functional Specification

Student Name: Niall Egan

Student Number: 19378906

Student Name: John Mc Cormack

Student Number: 19517396

Date Complete: 10/1/22

Table of Contents

Introduction:

- ## Overview:

  Our system will automatically produce summaries of pdf documents using Natural Language Processing techniques. This system could be useful in a number of scenarios. For example, in an online repository with large amounts of pdf documents and new ones being added regularly, it would be impractical for someone to provide a summary of each and every one of them. Our system would be able to handle this task.

- ## Business Context:

  N/A

- ## Glossary:

  - Natural Language Processing: This is a field of computer science, more specifically Artificial Intelligence, that's goal is to give computers the ability to understand language in the same way humans do.

General Description:

- ## Product / System Functions:

  Our system will generate a summary of a pdf document automatically. It will read in a pdf document and extract the text data. It will then conduct pre-processing on the data. Finally, the processed text will be passed to the Natural Language Processing algorithm to generate the summary.

- ## User Characteristics and Objectives:

  Users in our target community will likely be involved in a document repository platform. Users will be expected to have a high expertise with software systems. They will need to understand how to route documents in and out of the system.

  From the user's perspective, our system should be able to provide the summaries with a high level of accuracy and provide them in a reasonable amount of time. It should also be to conveniently interact with scripting, to allow automated summarization of PDF documents.

- ## Operational Scenarios:

1. **Use Case Name:** Summarize PDF
   Actor(s): User
   Flow of Events:
   1. User provides a PDF to the system.
   2. The system summarizes the PDF.
   3. The system return the summary to the User

- ## Constraints:

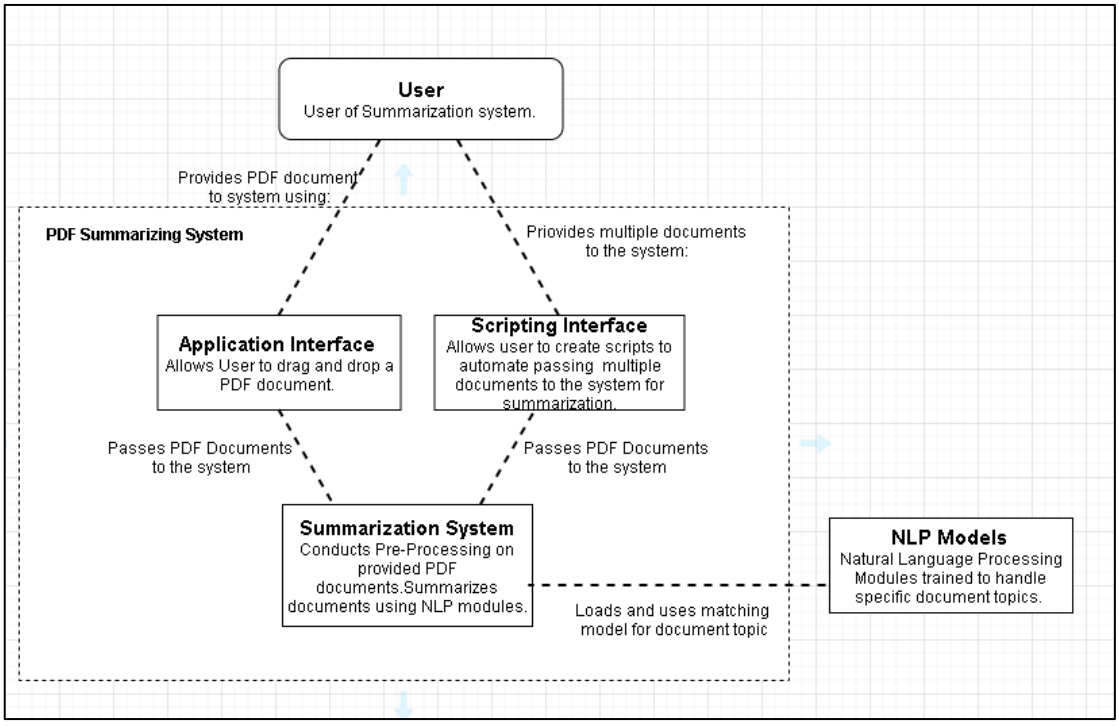
  The main constraint placed on the design team was regarding speed requirements. Our system must be able to provide summaries in a reasonable amount of time.

Functional Requirements:

- **Description:** System must be able to summarize a pdf document
- **Criticality:** This function is essential to the system. If the system cannot do this, it fails its main task.
- **Technical Issues:**
- **Dependencies with other requirements:** N/A

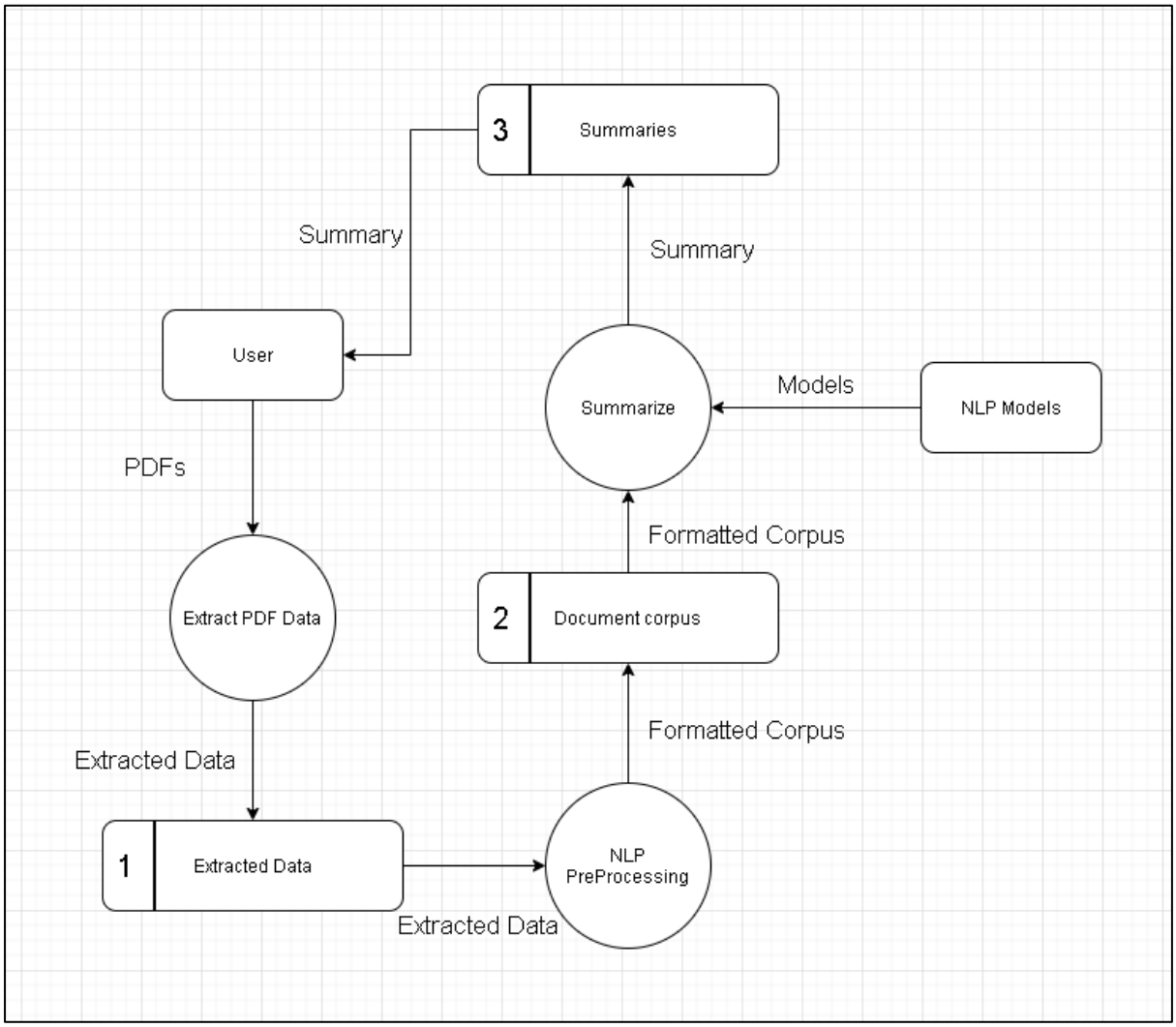
System Architecture:

Provisional System Architecture Diagram for the PDF Summarization System:



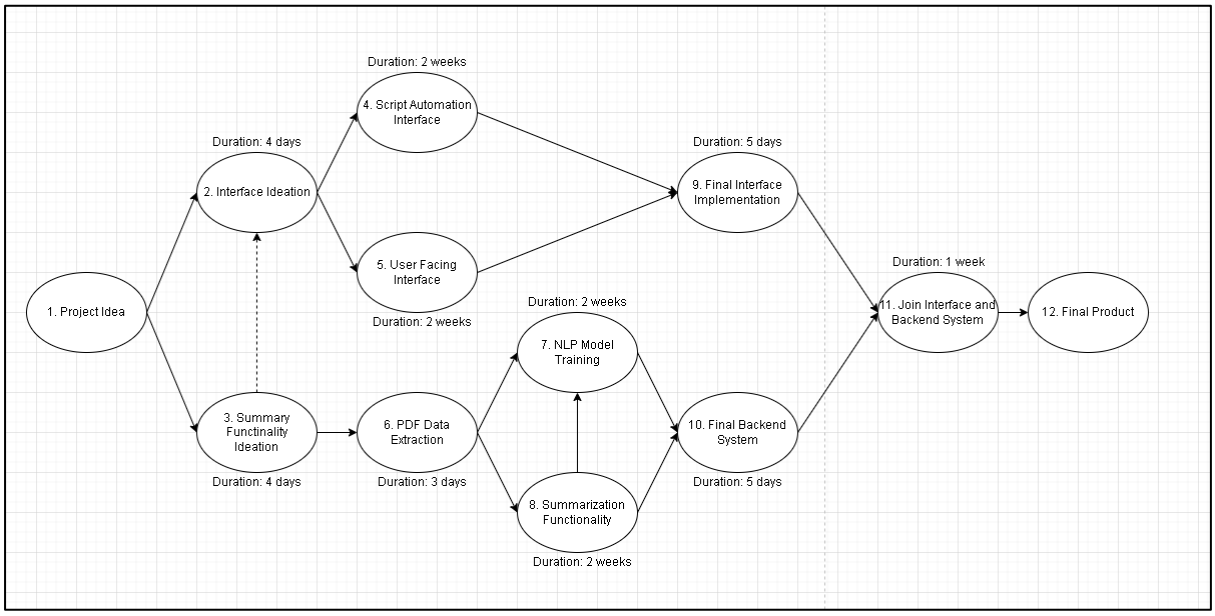
High-Level Design:

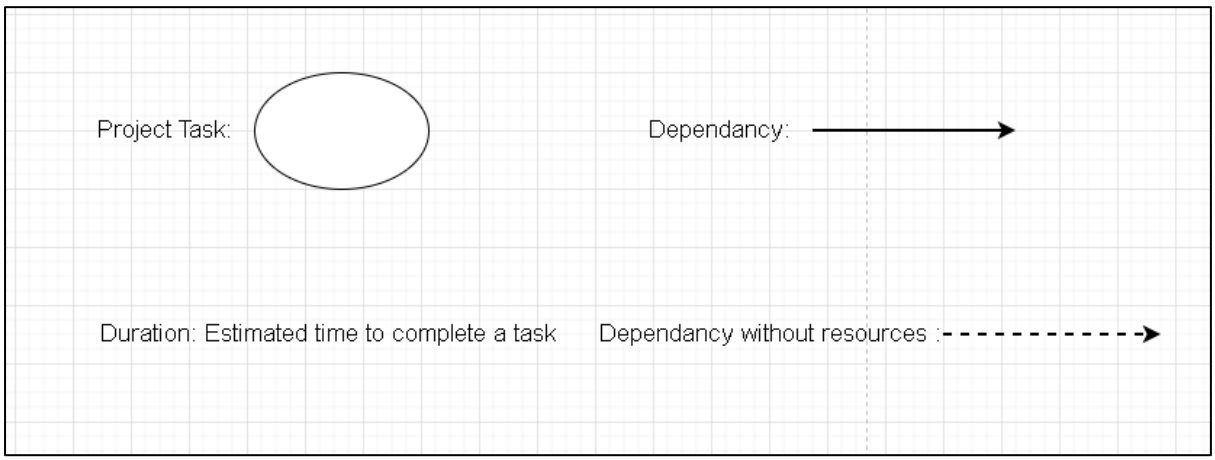
This is our Data Flow Diagram:



Preliminary Schedule:

This is our provisional PERT chart:





Appendices:

- **Documentation:**
  This is a list of important documentation we used when building the functional specification. This includes the documentation for relevant Python libraries:

    - Gensim Library Documentation: https://radimrehurek.com/gensim/auto_examples/index.html#documentation

    - Explanation of Natural Language Processing: https://www.ibm.com/cloud/learn/natural-language-processing

    - Natural Language Toolkit Documentation: https://www.nltk.org/book/

    - PyPDF 2 Documentation: https://pythonhosted.org/PyPDF2/PdfFileReader.html