School of Computing
Year 4 Project Functional Specification

Project Title: **Q&A Extraction from Factual Text**

Student Name: **Traian Svinti**
Student ID: **1443218**
Stream: **CASE4**

Project Supervisor: **Yvette Graham**

Completion Date: 18/11/2018

Table of Contents

1. Introduction

1.1 Overview

For my final year project, I aim to create an efficient and accurate questions and and extractor from factual text. This will be done using Natural Language Processing techniques (NLP). This webapp will take text as input and process this text to extract questions and answers based on the main ideas of the text and rank these questions based on how accurate and advanced they are. The user could modify, rank, download or answer these questions on the webapp. This is aimed for educational purposes where a teacher could create automatically generate questions or a student could generate questions to assess their retained knowledge of a specific piece of text.

1.2 Business Context

The area that this product will be used is in the educational field. As previously described, students and teachers alike can use this product to, more efficiently, create questions based on a text and quiz themselves/their students on the knowledge of this text. This is aimed to be fast and easy to use in contexts where time is vital, such as before an exam.

1.3 Glossary

- **NLTK -** Natural Language Toolkit for Python.
- **Django -** Backend framework that connects the UI to the database.
- **SQLite -** Database for storing information and submission/questions.
- **AWS -** Amazon Web Server is used as the deployment server of the product.

2. General Description

2.1 Product/System Functions

The process of using this product is described below:

Once the web-app is accessed, the user will have the option to login or simply extract questions from a text and download this text. If the user signs in, they will be able to view previously uploaded documents and the questions generated. There will be an upload button for uploading a document or scan of a text that will be extracted. The user will be given the option to modify this text before the question extraction begins. Once questions are returned; they user has the option to rank questions based on their accuracy and relevance based on the main ideas of the text, this ranking will also be saved on the users profile.

The user will have the option to also answer these questions and an accuracy percentage will be returned as to how accurate the answer is to the text.

2.2 User Characteristics and Objectives

This product is intended to be as simple as possible, with no more knowledge needed by the user than to use a simple, one page website. The user would need to click the upload button, search for the file or simply copy and paste the text. A login feature will also be implemented so knowledge of signing up using an email would be needed.

This is intended to be accessible by users of all ages, whether more senior teachers to primary level children.

2.3 Operational Scenarios

The operation of the system does not change on the type of user. The use cases are described below:

Use Case 1 - User accessing website without logging in

The user will access the website, be given the option to login or directly begin the extraction process. If the user does not log in, they will still be able to use the product normally but it cannot be saved until they sign in.

The user would input the text, and the output of question and answers and then downloadable/printable.

Use Case 2 - User accessing website with logging in

The user will access the website and login. Once they are logged in, they can upload their text, and questions and answers will be outputted. These questions and answers will be available to download/print and they will also automatically save to the user's profile.

Use Case 3 - Logged in as teacher

Once the user is logged in, they do not have to have a new piece of text to use the system. They can go over previous submission, edit them and provide these questions for their students to see. Once these questions are answered by the student, the answers will be returned to the teacher. The teacher can make the text visible or hidden to the student.

Use Case 4 - Logged in as student

Once a student logs in, they can also view their previously extracted text, questions and answers. They will also have a section where they are able to answer questions provided by their teacher based on a piece of text.

Use Case 5 - Incorrect file input

The user cannot input files that are not accepted text files. These files will be rejected with a clear error. The accepted text files are .txt and .pdf.

2.4 Constraints

Below is a list of constraints that will be faced with executing this project:

Time - No matter what project is being done, as long as there is a point when the project needs to be submitted, time will be a constraint. This will be an issue due to other college work which needs just as much attention and also, I work a part-time job and do as many hours as possible to sustain myself. Balance is key.

Efficiency - The efficiency of the processing text will need to be substantial as large amounts of texts can be submitted and there needs to be no longer than several seconds of wait time before questions are outputted.

Technique - There are so many different techniques in NLP that can be utilized for text manipulation and selecting the best for question extraction will be difficult.

Ease of use - The aim of the UI is to be as simple to use as possible. This is vital; hard thought would need to go into the design.

3. Functional Requirements

3.1 Logging In/Account creation

Description - Once the website is accessed, the user will be given the **option** to log into their account. This is done with an email and password. This email will need to be verified.

If the user continues without logging in, they will be prompted to create an account after the extraction process.

Criticality - Low (*optional)*

Technical Issues - Creating the login process should be simple through UI. The actual users will be saved and accessed by Django.

Dependencies - This is an independant step. The login process is optional to the user and is not dependant for the system to be utilized.

3.2 Text input/modification

Description - The user will be able to generate questions based on a number of inputs.

1. Pasting/Typing text into textbox
2. TXT/PDF file
3. Scanned file with text (PNG)

Once the user inputs this text, depending on the inputted, it will be extracted (OCR from scan) and the text will be provided in a text box for modification (removal of chapters, mistakes, irrelevant information). This will then be submitted to the next step.

Criticality - High (*essential)*

Technical Issues - There will be a simple function that will be able to take any of the above as input, an OCR will also be implemented for the scanned image of text.

Dependencies - This is an independant step, each subsequent step depending on text being inputted.

3.3 Natural Language Processing

Description - Once the text is inputted for processing, NLP techniques will need to be applied to extract questions and their associated answers from the text. This is the main and most difficult part of this project.

Criticality - High (*essential)*

Technical Issues - Creating efficient and accurate processing will be a challenging task. NLP has many techniques for text manipulation, knowing which ones to use and how to use them will be an important step.

Dependencies - This is dependant on the previous step (3.2). Text will need to be inputted to be manipulated and modified to extract questions and answers.

3.4 Post-processing

Description - Once the questions and answers have been extracted. The user will have the option to print the questions, print the answers or answer the question in the website. Also, if the user is logged in or signs up, the submitted text, questions and answers will be saved on their profile for future use.

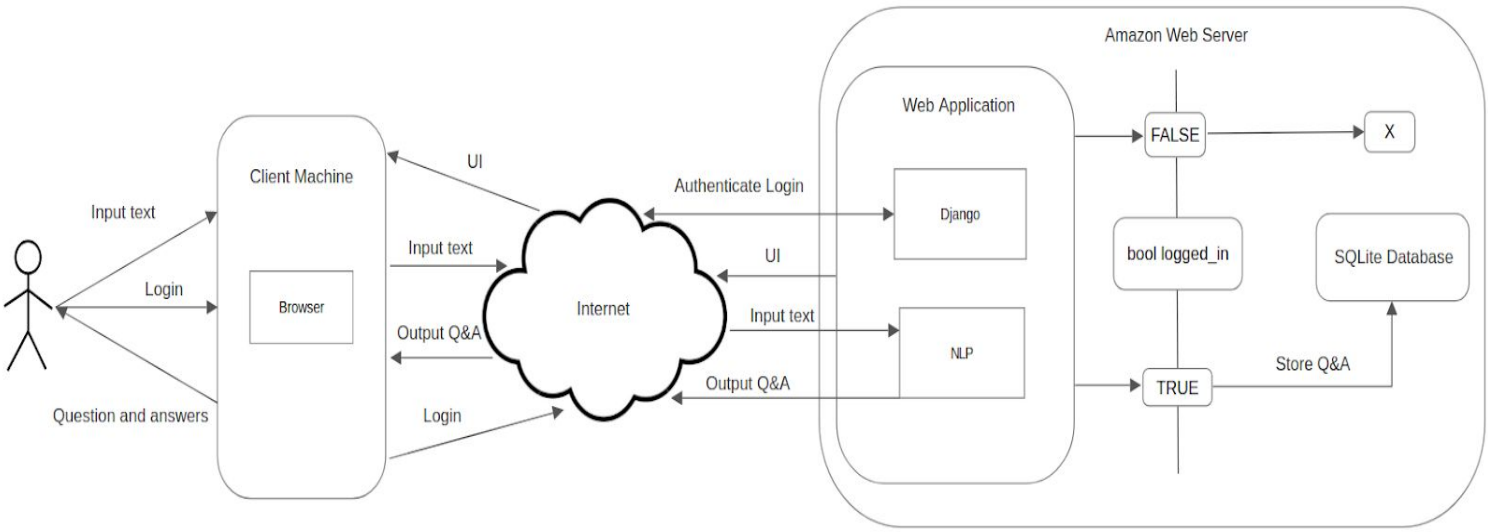
Criticality - Medium (*User may not want to use immediately)*

Technical Issues - Storing this information on each profile may prove difficult. Also making sure the UI is in an acceptable for to display this information and be as easy to use as possible will not be easy.

Dependencies - This is dependant on the previous step (3.3). There are no options in this step if processing is not done.

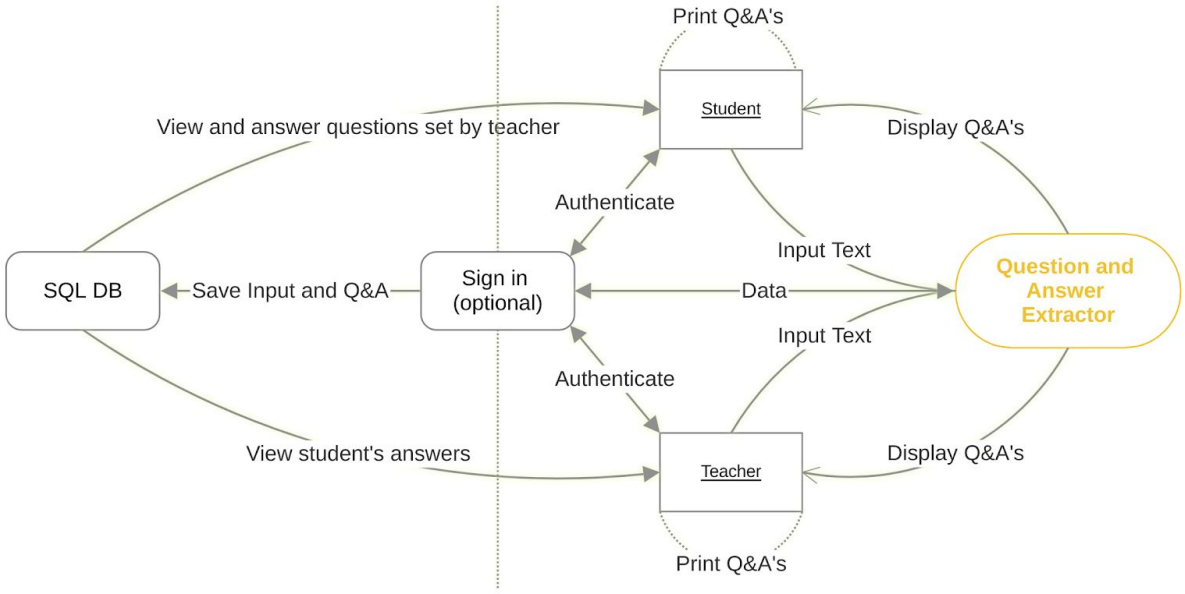
4. System Architecture

4.1 Figure A - System Architecture Diagram

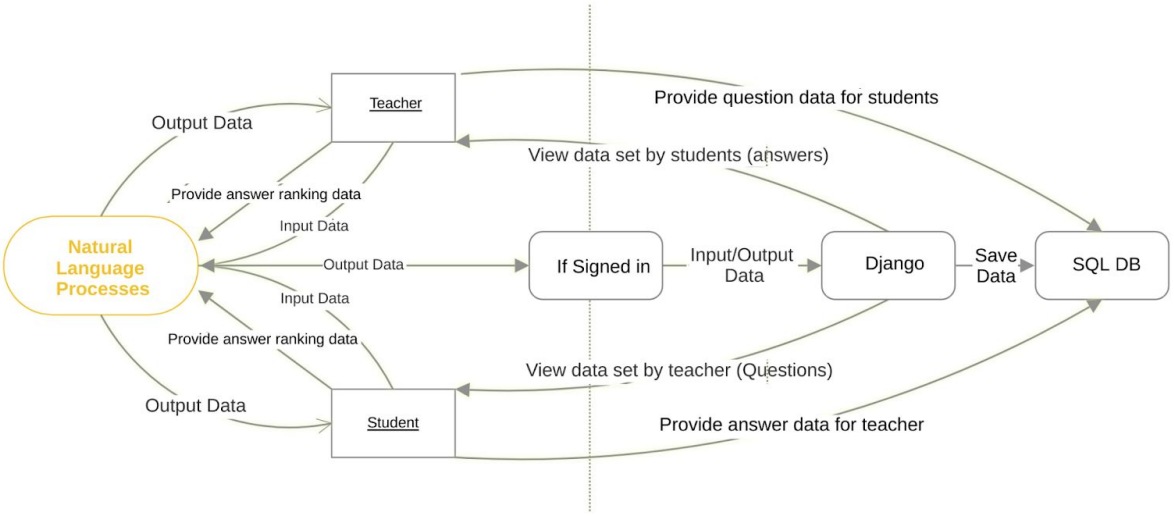


5. High-Level Design

5.1 Figure B - Context Diagram



5.2 Figure C - Data Flow Diagram



5.3 High-Level Design Description

- ➔ User uploads text as input the the processor.
- ➔ Question and answers are outputted to the user. These can be ranked.
- ➔ User is prompted to sign in to save input text and question/answer to the DB, if not, user can print or sign up.
- ➔ If the user signs in as a teacher, they can submit questions for a specified group of people to answer.
- ➔ If the user signs in as a student, they are able to answer the questions provided by the teacher. These answers are then returned to the teacher.

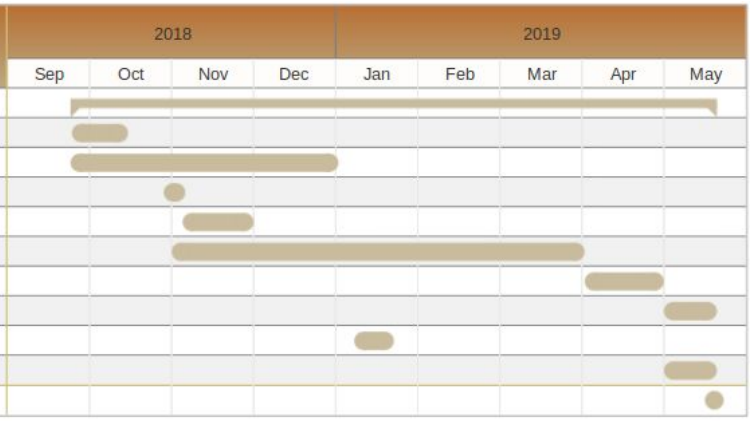
6. Preliminary Schedule

6.1 Overview

The time schedule in figure D below was created using SmartDraw. It shows each of our activities (task and events) that we have accomplished and future undertakings. On the left of the Gantt chart we are shown a list of activities and along the top are a suitable time scale. Each activity is represented by a bar. The position and length of this bar reflects the start date, duration and the end date of each activity

6.2 Figure D - Gantt Chart

	Task		Start	End	Dur
	Design Project ⊖		9/24/18	5/20/19	171
1	Identify a Project		9/24/18	10/15/18	15
2	Research/Brainstorming		9/24/18	1/1/19	72
3	Project Proposal		10/29/18	11/5/18	6
4	Functional Specification		11/5/18	11/30/18	20
5	NLPs		11/1/18	4/1/19	108
6	Django		4/2/19	4/30/19	21
7	Frontend		5/1/19	5/20/19	14
8	Semester 1 Exams		1/8/19	1/22/19	11
9	Semester 2 Exams		5/1/19	5/20/19	14
10	Project Submission		5/19/19	5/20/19	1



7. Appendix

https://www.nltk.org/
https://www.djangoproject.com/
https://aws.amazon.com/
https://www.sqlite.org/index.html
https://www.smartdraw.com/
https://www.nltk.org/book/ch07.html