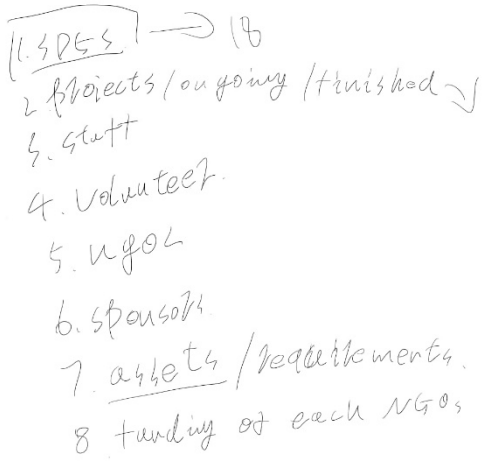
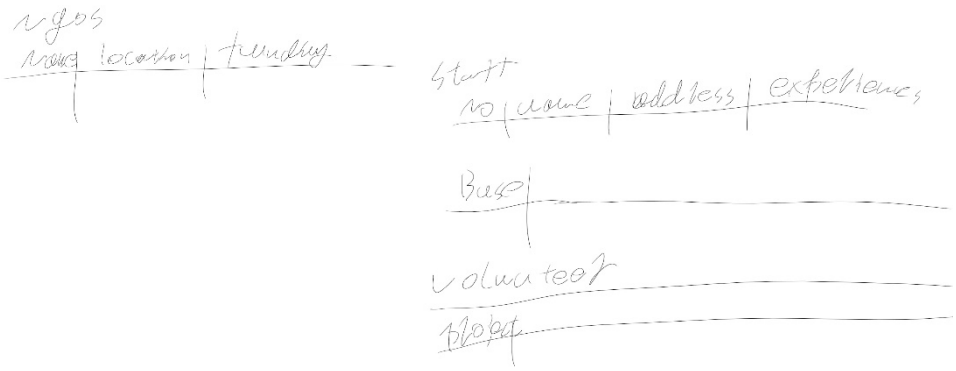
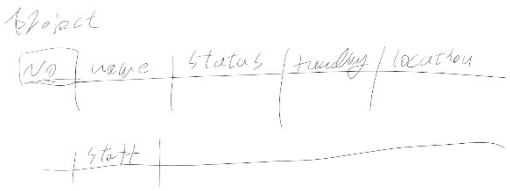
Fortnightly Report – 21/11/2019

Report of progress to ANCSSC:

So far, we have had 3 meetings in the last two weeks. In our first meeting, we created a database schema. The schema is essentially a design of the database. It outlines how to store the data extracted from the pdf. The sketches of the design are shown below:

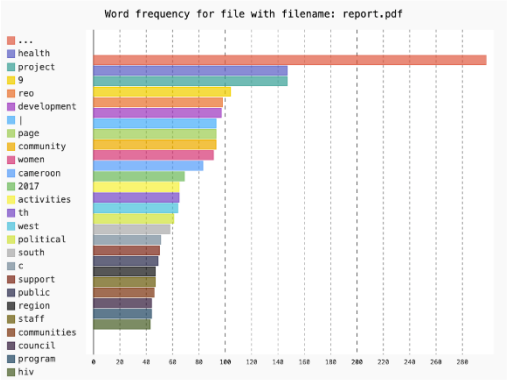






In our second meeting, we were trying to determine and research ways to extract data from pdf files. We decided to seek more assistance on our third meeting, from our technical expert, Joseph Connor. We were given a dissertation and code from a previous MSc student, who used NLP (natural language processing) techniques to do frequency analysis on pdf files in order to determine which words appear most frequently in each document. Although this is not directly applicable to our project, we can still use this feature to classify the documents based on which words appear most frequently as it gives us a general overview of the document.

Essentially, the MSc project was a visualisation tool to present the words which appear most frequently in a document in a graphical format. In the left image, the size of the word corresponds to the frequency of appearance of that word in the document. In the right image, the size of the bar corresponds to the frequency of appearance of that word in the document.



As part of our meeting today, we looked through the PowerPoint of requirements to create a timeline for an upcoming deadline in January 2020. This includes having to perform a demo of our prototype, submit our Portfolio 1, which includes the source code, the website and the individual report. The website should include the home page, the requirements page, our research, the HCI, our prototype, our achievements, our plans for term 2 and the appendices. In order to complete the large number of tasks within a smaller deadline, we allocated each person in the team a task - two people are working on the portfolio due in January and one person is focussing on the analysing existing data for the main project.

We also had a discussion with both Dean and our TA, to gain a further understanding on how to extract the data. Our TA advised us not to use existing code from the MSc project, unless it is directly applicable to our project. This is because we should use pre-existing code only if it can be justified, due to the limitations of the code and risks associated with using unknown source code. If we were to use the code, we would have to include strong reasons as to why it is more efficient that our own solution.

We have set up azure accounts and experimented with the different features available on the service. We set up admin-permissions for each member of our team.

Self-Evaluation of progress

I think we have made some progress as a team in the past two weeks. We have had multiple meetings and mainly focussed on data extraction from pdf files. This is because it is important to extract relevant and accurate data, so that the data can be synthesised at a later stage to produce a good model.

Plan for the coming two weeks

In the next two weeks, we aim to complete a large portion of our Portfolio 1, including the website, the individual reports, and source code. Our aim for the project is to:

- Make significant progress on the prototype.
- Decide whether to use a rules engine instead of NLP techniques to extract data.
- Decide on exactly what data we need to extract – facts + figures
- Start creating a database outline in Azure