ANCSSC

“A cloud solution for analysing patterns in NGO projects”

Team Number: Team 36

Names: Rachel Mattoo, Mark Anson, Yansong Liu

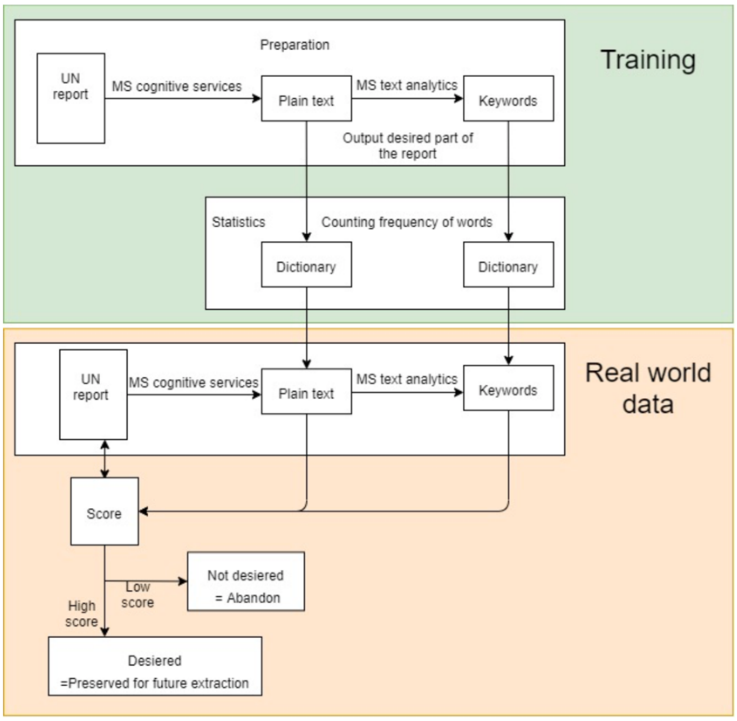
Date: 24/01/2020

Report of progress to ANCSSC

A large portion of the team's time was spent on designing and implementing a pdf extraction tool in the last two weeks. The programming section involved a multi-step design process.

Since a pdf file consists of unstructured data, the biggest challenge in the process so far has been designing an algorithm to extract this information, so that it can be stored in a structured way. The extraction tool is a two-step process. The first part has been developed, which involved determining which part of each annual report contains relevant information (e.g. financial information). This involved having to create a separate training data set on a desired input and using a scoring algorithm to compare this to any input to determine the relevancy of the information. The second challenge is to extract the most important information in a given text and recognising this information in context. For example, if we extract a number, we want the information around it to determine what it represents (e.g. profit/revenue). This challenge will be overcome by consulting PHD students who specialise in NLP.

Diagram + Algorithm explanation overview



We input a desired particular part of the UN report as the training dataset. In this example, the training dataset consists purely of financial information. This is because financial information is fit for our purpose because the information is quantifiable, and so we can easily spot trends and patterns.

The Microsoft Cognitive Services Ink recognition API converts this pdf text into plain text. The Microsoft Cognitive Services Text Analytics API is then run on this plain text to extract the key words from the file. In order to build the dictionaries, two empty dictionaries are created initially. Each word in the plain text is checked to see if it exists in the dictionary. If it doesn't, then it is added to the dictionary. If not, a counter is incremented to indicate a higher frequency for the corresponding word. The same process is applied to create a keyword dictionary.

The resulting dictionaries are exported as a search facility for another data set, which could be another UN report, for example. The same process is applied to this section, where Ink Recognition API is used to convert the pdf into plain text and the MS Text Analytics API is used to extract the keywords from the file. Afterwards, each word in the plain text and keywords are checked against the corresponding dictionaries (generated by the training dataset) and allocated a score. The score is higher for keyword match than a plain text word match.

At the end, both the plain text and keyword scores are accumulated. If the score is high enough, the input will be preserved for future extraction. If it's too low, the input is not desired so this part will not be considered for future extraction. Essentially, the aim of this process is to check the relevancy of the information in each section of the report so that we only preserve the relevant sections for future analysis. The information in the relevant sections will be used to build the database.

Video creation

We created a video explanation and demonstration of the code.

Website creation

We created a website to document our progress and prototype design.

Meetings we have had so far

Dean

- Consult an NLP specialist
- ANCSSC have money problems
    - Money problems with the ANCSSC so costing is something we should account for

Yun

- Scores should be expressed as a percentage of the total length
    - Prevents bias in length of text

List of tasks completed and whether project is running on time

We have successfully managed to submit our Prototype on time, and all the requirements which came with it. Here is a list of everything we have done so far:

- Early pdf extraction tool
    - Scoring algorithm to find region of text

        - Related to financial information
- Meeting with Dean
- Feedback from Yun on early pdf extraction tool

Plan for the coming two weeks

- Continue to work on the pdf extraction tool
    - Get pickling working
    - Create percentage function
    - Integrate everything
    - Figure out a way to package the pdf extraction tool so it can be easily deployed
    - Creating a database plan
- Figure out a more cost-efficient solution for data extraction
- Consult an NLP specialist on the next step of our extraction tool