Week 3

Report of progress to ANCSSC

Date: 21/11/19

So far, we have to figure out how to actually extract information from the pdf. Currently, we have been experimenting with OCR APIs such as Google's Cloud Vision API in order to extract information from documents which are entirely image based or in table format. We have also been experimenting with pdf to text converters and decided it would be easier to create a rules engine in Python to extract relevant information.

Date of meeting: 28/11/19
Who: Sheena + Dean

Sheena suggested that the reason the document is mostly in dollars is because it is a stable currency for conversion. We realised that all the financial data must be in the same format so it is comparable so we must use a tool which allows conversion between the different currencies.

After talking to Dean, we concluded that we must build a rules engine which searches for text and looks for finance information specifically. We clarified that the front-end must be a web app to find NGOs already mapped, progression and what it needs. We need a web app of data on different points which is a visual representation of sample code. We realised that we need to check how the report is different from the data you are collecting – if the data is the same, it is overcomplicated to mine data, all the data is open source, how does the data you're collecting from the NGOs differ from the data they are collecting.

For our next meeting, we are planning to figure out

- How to extract data from pdf
- Manually fill in the data – do we just use the one dataset?
- Do the reports match the NGO reports?

In order to clarify these technical requirements, we had a skype follow-up meeting with Dean.

Date of meeting: 03/12/19
Who: Dean

We had a meeting with Dean regarding a separate application to extract pdf files, so that any research group including us can extract pdf files. Dean spoke to a Microsoft team for us and asked me (Rachel) for a follow-up meeting. After speaking to Microsoft, Dean introduced us to the Cognitive Services API provided by Microsoft. The two APIs which are directly relevant for our purpose are:

1. **Ink Recognition:** Uses OCR recognition to convert images to text
2. **Text Analytics:** Uses semantic analysis to find the keywords in any given text

These APIs are fit for our purpose because they are: cheap & robust

We were also introduced to the following

- Microsoft store and download office lens
    - Flow automation
    - Power tools

    - Graph API

We decided not to go with the alternative option because this was an offline version so it would have a higher set-up cost, in terms of time and efficiency.

Dean was also kind enough to introduce us some new people at Microsoft who were willing to support us on this project:

- Lee - director of education
- Sam - lead technical expert of Microsoft Philanthropies
    - CTO of group

We were also introduced to new technologies at Microsoft:

- There is an app called Office Lens from Microsoft
    - Can be automated
    - Takes a picture of a page
- Feature found on excel is a component borrowed from office lens
    - No public API
    - Can use office lens and flow automation and power tools

At the end of the meeting, Dean also happened to mention that the final product will be branded by Microsoft philanthropy to support NGOs around the world!

Update – 09/12/19

- Ink Recognition API for data extraction is working
- Currently trying to work out the distance between two bounding boxes

Self-Evaluation of progress

We have made significant progress in terms of requirement gathering. We have a solid set of technical requirements and a good strategy for achieving the basic requirements on our MoSCoW list. Developing a strategy involved a lot of planning, mainly in the form of flow-diagrams and sketches which was then developed into an architecture diagram for the algorithm.

Plan for the coming two weeks

What we need to do

- Create a database for ANCSSC and a form so that they can manually input data
- Create a database for UN data
    - First figure out how to use cognitive services for pdf data extraction of annual reports

Main objectives for the coming weeks:

- Report website
- Form
- Cognitive Services API

Potential meeting with Husna

- Meeting next week – 12/12/2019
- Other UN people to give advice on what they want they want out of this
- Create a front-end to allow NGO to insert as many fields as possible
    - More important to have all the fields possible
    - Create a long schema and just record what they say as they approach charities
    - She wants to sit with us and go over how it will be used by the UN at a high level to govern NGOs