“A cloud solution for analysing patterns in NGO projects”

Team Number: Team 36

Names: Rachel Mattoo, Mark Anson, Yansong Liu

Date: 07/02/2020

Report of progress to the ANCSSC

So far, we have managed to fix any remaining bugs and completed the first part of the data extraction tool. We have also had a meeting with a master's team and Team 37 and added a database backend (ANCSSC database – Database 1) for these teams to our user requirements. This change to our user requirements was strictly necessary because it was only in Term 2 and during this meeting that we realised how our client's requirements related to our project brief, and so our user requirements had to be updated to reflect these.

Meeting with master's team

After our meeting with the master's team, we realised that the main aspects to focus on (for the ANCSSC database – Database 1) would be geolocation and a description of the list of projects that the organisation is currently collecting.

The key aspects for each project are:

- dates
- how much money they have?
- from whom
- how many people were working on the project?
- keywords - technology called x

After some discussion with Dean, the masters team realised that the following deliverables must be created:

Team 36 database – need to create a table to search by region, grants, searching for people such as staff and volunteers, filtering in a table when projects are expiring (when is the grant ending), red list - grant ending but project not finished, orange list - a disaster / emergency happened so these charities need help immediately. Ines and their partners need to authorize any changes

They also provided an early data schema for their front-end web app for the ANCSSC:

NGO:

NGO Name
Comms officer name
email
phone number
username
password

Project:

dates - start and end (when project ends - switch status to finished??)
how much money they got?
from whom
how many people were working on the project?
keywords (technology names?)

Type: orange/ red/ normal
description field - abstract
description filed - detailed overview
images - links / BLOBs
[contact card displayed]
discussion forum - maybe like a comment section
Status: pending / confirmed by admin
Status 2: Active / finished
Some sort of a measure for how successful the project is

Admin
create projects
create NGOs
confirm any changes made by the NGO
monitor the system (filter the comments section, delete / add info etc)

Meeting with NLP Specialist, Dr. Pontus Stenetorp – 07/02/2020
Our main question to answer for this meeting was "How to extract data in context?". Here are some of the points we covered:

- "Shallow parsing"
- Chunking – infer given phrases and is a brute force alternative to machine learning
- Constituency and dependency are factors to consider
- No need for black box, which is a concern with Deep Learning techniques
- BIO – begin inside outside
    - Part of named entity recognition
    - Each label is associated with a given token
- NLTK/Spacey – Example of chunking software
- RoBERTa – Question and answering system
    - Based on BERT and allows the user to ask questions to any text
    - Trained on the SQuAD dataset
- ALBERT
    - Point at text and play around with the questions
    - These are computation heavy
    - Few gigs of memory
- Precision, recall, F1 – interpolation
    - Methods for determining the accuracy of the machine learning models listed above

- Removing bias
    - Overfitting problem
        - A problem with the model itself
    - Programming
        - Context switching, run it on one document and then next one so the RAM isn't being used

List of tasks completed and whether project is running on time

Project is on track on meeting the intended list of requirements.

Here is a list of things we have done in the past two weeks:

- NLP meeting with Dr. Pontus Stenetorp
    - Introduced us to both brute-force (chunking) and machine learning methods (BERT)
    - Also introduced us to hugging face python wrapper for BERT
- Remaining bugs for early pdf extraction tool were fixed
- Feedback on early pdf extraction tool taken on board and implemented

Plan for coming two weeks

- Remove bias in our program
- Test different NLP techniques
    - Chunking
    - BERT
- Test different BERT models to determine most accurate model
    - BERT
    - RoBERTa
    - ALBERT
- Train different models on SQuAD a different number of times
- Create a training data set to fine-tune models on our own data