Individual Report 2

Mark Anson (18004601)
Team 36 – with ANCSSC

1. Title

'Let's leave no one behind': A cloud solution for analysing patterns in NGO projects with a first stage synthetic data generator

2. Contributions

2.1 Coding

I noted in my first individual report that I wanted to become more engaged in the programming aspects of the project. I am happy to report that throughout the rest of the project, this has certainly been the case. I have made numerous contributions to different technical aspects of the project, including:

2.1.1 Management of the ANCSSC database (Database 1)

I spent a lot of time working on this database design. Following on from prototype 1, we learned a lot more about what database 1 would actually be used for, as well as an idea about what the schema would look like. This required the database to be completely re-written. I had to handle communications between our team, Team 37, and the Masters team all working with the ANCSSC in order to build one coherent database structure that would work for everyone.

2.1.2 Contributions to BERT and the PDF extractor

Of course, Yansong put the majority of his effort into the PDF extractor. I did still make contributions along the way. In the initial version of the PDF extractor, pre-BERT, I created a feature to allow users to save data they had created as "pickles" in python. My code would also allow the user to "revert" saves if they had saved accidentally. I also developed a test suite for the features we had created. For our reportQuery system, which replaced the previous one after we moved to using BERT, I worked on an algorithm to extract the names of NGOs from the documents. It would search the document name as well as email addresses and website domains within it to come up with the name of the NGO. It could then ask the user for confirmation. This algorithm was necessary as the name of the NGO was needed in order to build questions for the BERT model to answer, and we found it very difficult to use BERT to reliably get the name in the first place. The PDF name and corresponding NGO name would then be stored in a table within Database 2.

2.1.3 Integration of our reportQuery system with Database 2

I was in charge of linking our reportQuery system with database 2. We worked on a format of datatype that would be produced by the reportQuery system and passed to my insertion code. We decided on a dictionary of dictionaries, with keys matching the names and field names of tables, making integration as easy as possible. I constructed code to insert all of the data into the relevant part of the database. I also wrote code to fetch the name of the NGO using the data stored in the database from my name extraction algorithm.

2.2 Research

Research continued throughout our project as we explored new concepts. In particular, I put a lot of time into researching potential methods to extract data from tables within our NGO report PDFs. I looked into Microsoft's brand new "Form Recognition" tool, which seemed very promising. However, on further research, I discovered that it was not suitable for our dataset as it required training on a large number of similar tables, something we did not have. I also looked at a tool called "tablua" which seemed to be the premier form of table extraction from PDFs. I wrote an algorithm to utilise this on our PDFs, however, after some hard work this was eventually abandoned. The output we were getting from tablua contained too much garbage text and inconsistent formatting, making it impossible to automate with any kind of accuracy.

2.3 Client Liaison

My team and I continued to work with our clients throughout the project. Following on from my first individual report, I believe I have improved at contributing more to our meetings with them, asking pertinent questions and providing answers to theirs.

2.4 Report Website

Following on from Prototype 1, I was the main contributor to the updated website for our final submission. This took a lot of co-ordination across our team in order to update all relevant information, as well as provide detailed write ups of all new elements.

2.5 Video Editing

Due to the rise of COVID-19 and our team's abrupt return to our homes, we could not record our video as a group. Instead, the video production was split relatively equally among the team. We each recorded our own sections of the videos, and I provided editing help to Yansong, who did not have the facilities to do so on his own computer. I was then responsible for stitching the final video together.

3. Difficulties

3.1 Research and "finding our way" through our requirements

Looking back, I think our biggest hurdle in this project was figuring out our approach. We spent a very significant amount of time on research, mostly through the internet or talking to the various experts we could get in contact with. In particular, the transition from our initial version of the PDF extraction tool to the current one using BERT took a long time and plenty of research. Yansong had to go back and seek further advice from Dr Pontus Stenetorp after our initial conversation in order to get set up as the online documentation was proving insufficient.

3.2 Testing different BERT models

As part of the process of researching BERT, we wanted to compare different Hugging Face transformers models on two metrics, how increasing the number of “passes” would increase accuracy, and also compare different versions of the BERT model including ALBERT. This proved a very difficult process. The algorithm to run the test would take several hours, and we often found it would inexplicably crash or stop. We managed to get some useful information out after several days of constant trials. Eventually we had to stop testing and work only with the data we had managed to collect due to time constraints.

3.3 Management of the ANCSSC database (Database 1)

Team 37 and the Masters team had somewhat conflicting goals in regards to their use of database 1, and as such it took a lot of time to build a structure that worked for everyone. It was also a struggle to get the Masters team to pin down exactly what they wanted from the structure, as they themselves did not really know what they would need. Their deadline was also set after the completion of our project, so they were under less pressure to come up with a structure for the database right away. An example of this difficulty occurred a few weeks before the deadline of our whole project, where the Masters team requested us to look through their design document and figure out what changes to the schema would be necessary. I spent time implementing the changes as best I could following this, but we had to make clear to the team that our focus was very much on getting our own work complete due to the short amount of time we had left. I believe that the structure I have built will work well for their team and should be flexible enough to handle any changes in their requirements.

3.4 Integration of our reportQuery system with Database 2

This proved quite challenging as I had to make sure data was inserted correctly following Rachel’s database structure. There were also a few miscommunications that were made regarding the specifics of the dataset Yansong’s algorithm would produce, hence the debugging process took quite a while. I was unable to run the complete program from my computer as I was missing certain components required to run it, instead I had to push changes I made to UCL’s blaze computer in order to test it. After a lot of testing and bug fixing, we managed to produce a working system, which has now successfully processed all of our PDF reports.

4. Assessment of team members

Following on from my comments in the first individual report, I still feel very lucky to have been assigned my teammates. We have bonded well as a team and have managed to keep on top of our tasks well. My team is very communicative and open to working through problems. Towards the end of the project we would spend hours almost every day working on the project together.

4.1 Rachel Mattoo

Over the course of this module, Rachel has made herself out to be an excellent team leader. She is definitely well suited to this role. She has been great at coordination and communication within our team, with clients, and with fellow teams also working with the ANCSSC. She has involved herself in all aspects of this project, from programming to bi-weekly reports, and has shown great competence in all of them. Rachel has said herself that she still wants to be more pro-active with documentation writing in the future, following on from prototype 1, although I think she has certainly gotten better at it.

Strengths: Organisation, communication, programming
Weaknesses: Pro-active documentation

4.2 Yansong (Mike) Liu

Yansong has been invaluable to this team. His has thrown himself into the most technical aspects of this project and handled them incredibly well. He is responsible for the bulk of work surrounding the BERT algorithm. He continued to come up with great solutions to problems we were facing, especially regarding BERT. I absolutely cannot fault his work ethic and commitment to this project. He continued to be best suited to the programming aspects of our project. Yansong has definitely improved in terms of his contributions to meetings since prototype 1 and has engaged much more in them. In terms of weaknesses, looking back I think Yansong could have helped contribute more to documentation. He still contributed more than enough work overall, but I think it would have been beneficial to have seen more of his perspective come through in the various report submissions we have made.

Strengths: Programming, problem solving
Weaknesses: Documentation

4.3 Mark Anson (Myself)

Overall, I am very happy with the work I have contributed during this project. Beyond the things I can mention specifically, I spent huge amounts of my time discussing ideas and providing input and assistance to my teammates to the best of my ability. I have made plenty of important contributions to this project, having built up a large codebase for various purposes, including all of my now defunct work for our initial version of the PDF extraction algorithm, my table extraction work, my database insertion work for reportQuery, and my work on database 1. I was involved heavily the report elements of the project including creating and editing videos and the project website, as well as research, requirement analysis, and client liaison throughout the project.

Following on from what I wrote in my first individual report, I think I have gotten much better at contributing to meetings, especially with clients and experts. A great example of this was in a meeting with John Booth, the senior data steward at GOSH DRIVE, I was able to help provide clarity regarding data privacy concerns he had relating to the use of our reportQuery system within the NHS. I think I have also gotten much better at coming up with solutions to the programming problems we were facing, for example coming up with a solution to finding NGO names with my name extraction algorithm.

I think within this project I was well suited to working in programming and documentation, as I believe that they are my two strongest areas. In terms of my weaknesses, I think one thing I need to work on is the robustness of my code, especially if it is going to be integrated into another system. A few times I put a little too much trust in my ability to write perfect code the first time. For instance, during integration testing for the database insertion code I'd written, there were a few bugs that I could have squashed beforehand if I had spent more time testing.

Strengths: Documentation, communication, programming
Weaknesses: Stringent testing

4.4 Improvements in general areas of weakness

In my first individual report, I mentioned that our team needed to focus on our timeliness. This is certainly something that has improved as the project has continued. I found that we got much better at meeting at prescribed times and getting down to work. This was incredibly beneficial especially as the project neared its conclusion. During the onset of COVID-19 we quickly created a routine of daily group calls in the morning to allow Yansong to join in from China.

4.5 Scoring

My approach to scoring from my first report has not changed. I think, given what has been discussed throughout our various reports, it is very reasonable to give 10s all around. As before, the weaknesses I described feel rather like nit-picking.

Rachel:	10
Yansong:	10
Mark:	10

5. Conclusion

At the end of this report, I would like to take a moment to thank the various members of staff that have helped us get to our final submission. In particular, our TA, Sheena Visram, has been an incredible help to not just our team but many others. Her useful advice and ever-present optimism kept morale high and kept the team focused in times where we were struggling to make sense of what to do next. Dean Mohamedally has provided us with great insight into the nature of our client's requirements as well as pointing us in the direction of experts such as Joseph Greener. He has worked tirelessly this year to keep everyone's projects moving forwards.

The Systems Engineering module has provided me with a huge learning experience. It has allowed me to improve teamwork, communication, client liaison, and time management skills while also learning a lot about emerging technologies in machine learning. This has not been an easy journey, but it is great to have worked on a project that we believe in time will provide real benefit to the ANCSSC and maybe even other sectors as well.