Team 4: Bi-Weekly Report 3

Project Title: PEACH Data Mining Date: 17 November 2017

Gavin Shek, Saqib Jahangir, David Stepanovs

Overview

Having now completed all the HCI aspects of the project, over the past 2 weeks we have been researching open-source data anonymization tools and libraries. Alongside this, we have also looked at the previous data generator tool produced by the previous team. This was available to us via the PEACH GitLab. We also gained an understanding of what is important and required when anonymising confidential patient data by reading through various articles and documents.

Meetings

Numerous Meetings During and Outside of Lab Sessions

During our meetings, we ensured that as a team, we were researching as many different technologies and solutions as possible. We made sure we were able to compare these different pieces of software and outlined what we wanted. This included something which was open source, had a clear structure and was effective in anonymizing data.

We also researched into the requirements of data anonymizing including what was required by law, and what was required by researchers. Discovering a payoff between anonymising too much data and created vague results, we considered these aspects when comparing open source solutions.

Completed Tasks

- Researched many different data anonymization tools
- Compared data anonymization tools
- Found a good open-source data anonymization library - ARX
- Understood what was required by law when anonymizing confidential client data

Problems to be Solved

- Understand how ARX works
- Research into effectiveness of other more complex anonymization techniques
    - Genetic algorithms
    - Machine learning

Plan

- Understand how ARX works
- Carry out further research into machine learning ideas and how to incorporate this into our project

Individual Section

Gavin Shek

Over these past couple of weeks, I researched into what was required of data anonymization tools to both ensure detailed enough results but also processed data which isn't able to identify people. Using this, I was able to quantify the effectiveness of different existing solutions and then began looking at the existing data generator produced by previous teams.

Saqib Jahangir

I compared multiple data anonymization tools and found that the ARX library would suit us best, because it's open source and produces good anonymous datasets.

David Stepanovs

I found multiple open source data anonymization tools and looked at the code to understand how they work.