Visualisation of Mobile App Usage

Matthew James O'Hare - 2255357o

December 20, 2019

1 Status report

1.1 Proposal

1.1.1 Motivation

Comparing the usage of mobile phones in this day and age is a very popular subject. In a world where there a thousands of possible applications to be downloaded, the specific ones that users download and use for different reasons is of great interest to researchers. Usage patterns especially as it helps researches understand the link between applications and further research of whether it is possible to predict or identify a user based on their usage. All previous analysis of this subject has involved statistical methods however this project looks at extending these methods and developing more visual models in the form of spring models or force-directed graph layouts. They take high dimensional data and display it (using a specified metric) in a two and three dimensional space to allow easy visual comparison of data. The hope is that this project will assist this and help us to visualise the differences (or similarities) between iPhone users.

1.1.2 Aims

The project has various aims throughout it. The project will take data that has previously been used with statistical models and attempt to display them as visualisations, specifically spring models or force-directed graph layouts. This is the end goal.

Aims leading up to this include exploring the data as much as possible and identifying key features that can help identify a user by their application usage, this includes removing anomalies and using the correct characteristics to help create an accurate model. The end result of this analysis would be a modified data set which has a valuable set of users and their features.

From this, analysis will be carried out to find the best metric for comparing rows in this data (each row represents a user) such as hamming distance, cosine similarity and others.

The project will then explore the possible spring model algorithms on offer such as Chalmers 96, PCA, t-SNE and UMAP. This modified data set along with the best metric will be used with these algorithms to create a spring model layout to help us visualise this high dimensional data in two and three dimensional space.

1.2 Progress

- Language chosen: The project will be implemented in Python 3.
- Background research carried out on spring model algorithms such as PCA, t-SNE and UMAP.

- Background research carried out on where the initial app usage data came from and how it was harvested.
- In depth analysis has been carried out to evaluate the characteristics of the data. Also to find what data is useful in conjunction with others and any possible anomalies in the data (ie. users that have not used many apps may not be useful).
- At the time of writing, the current metric being used to compare the data is hamming distance however is likely to change or be improved.
- At the time of writing, the spring model algorithm to be used has not been decided while the best metric is still being chosen.

1.3 Problems and risks

1.3.1 Problems

- Due to the large data set (over 289,000 rows), runtimes are long for both importing the data and carrying out analysis on it.
- When comparing the data, the outputs change dramatically based on what metric is used.

1.3.2 Risks

- Due to the large data set (over 289,000 rows), runtimes are long for both importing the data and carrying out analysis on it. **Mitigation:** will use only a sub-sample of the data set when carrying out comparisons of the data
- When using the spring model algorithms, the layouts of them change dramatically based on what metric is used. **Mitigation:** Analysis will be carried out on a sub-sample of the data to find the best metric therefore when the time comes to run the spring model algorithm, I know that the metric used will be the most efficient.
- Unclear whether the spring model algorithm will create an accurate representation of the differences between users. **Mitigation:** will use or develop a metric that fits the data appropriately and will create the best comparison of it.

1.4 Plan

Semester 2

- Week 13-16: Continue the analysis of finding the most appropriate metric with the best spring model algorithm. **Deliverable:** A written process of this analysis followed by the actual implementation of it to be viewed.
- Week 16-18: Full analysis and evaluation of the created visualisation. **Deliverable:** Detailed evaluation document detailing the result of the process, why this was the result and any potential issues or changes to the process.

- Week 18-20: Potentially using this process on a different data set. **Deliverable:** Evaluation document detailing the result of using this process with a different data set and the comparisons between both experiments.
- Week 20-24: Write the dissertation. **Deliverable:** Multiple drafts of the dissertation submitted to my supervisor with plenty of time before the deadline

1.5 Ethics and data

This project does not involve human subjects or data. No approval required.