Group Report:

This report details the completeness of our python twitter data analysis notebook with regards to the basic and additional requirements, the design and implementation of the notebook and any known problems with the final submission. In addition to this report, each member has submitted an individual report detailing their team contributions, summary of provenance stating which fragments of code they wrote and any individual problems they experienced and how they solved them.

Summary of requirements:

Most of the requirements are satisfied in the python notebook “Digifest-2016-Twitter-analysis-final.ipynb”

Basic:

- **Refine the dataset**
    - We refine the dataset using regular expressions and creating new columns with useful information in the pandas dataframe. See notebook for more details.
- **Analysis and visualisations**
    - The total number of tweets, retweets and replies in the dataset is analysed in part 2 “Structure of the dataset” in the notebook. The visualisation for this is done using a pie chart.
    - The means and standard deviations for the number of tweets, retweets and replies sent by each user is analysed in part 3 “Means and standard deviations.”
    - Hashtag cloud in notebook part 5 “Hashtag cloud”

Easy:

- **Applications used to send tweets**
    - Analysis in the notebook part 6 “Applications used to send tweets”
- **Extended analysis**
    - Extended analysis of means and standard deviations found in notebook part 3 “Means and standard deviations”

Medium:

- **User activity over the period of time**
    - Found in the notebook part 4 “Timeline of activity”
- **Interaction between users**

    - Found in the notebook part 7 “Networkx and Graphviz”
    - Done using networkx to creating a graph of nodes and connections between the users to see their interaction.

Hard:

- **Analysing other datasets**
    - We created a python script named “`json-csv-converter.py`”. This takes raw twitter data in JSON format and converts it to a csv file with the same columns as digifest2016dataset.csv, details of its use can be found in the README. This means we can reuse parts of our notebook for creating graphs for the data generated from the script.
    - The file “`Twitter-eval-reuse.ipynb`” is a slightly changed version of our main notebook for use with files generated by “`json-csv-converter.py`”. It is designed for use with JSON twitter files found in “`studres/Library/TwitterSample`”. The “`TwitterSample`” directory contains many sub directories containing JSON files each with a single minute of raw twitter data from July 2015. As the sample is smaller and the time scale is much shorter analysis of how the data changes over time and the 3D graph have been removed from this notebook file. By changing the file read in at the start of “`Twitter-eval-reuse.ipynb`” it should be able to generate graphs for any csv converted from the data on studres with our script. One of the twitter files from studres, “`00.json`” has been submitted with this practical along with the converted csv file “`twitter_data.csv`” which can be used for running “`Twitter-eval-reuse.ipynb`” but any converted file produced by “`json-csv-converter.py`” should work.

  3D graph
    - draws comparisons between 3 different datasets, variable can be used to change the viewing angle and elevation to see the interactions between the 3 datasets. Found in notebook part 8 “Followers, friends and retweets”.

Known problems:

Error in creating new columns in dataframe

The python notebook gives red cells on line 6 when we add new columns to the dataframe. The new columns are still made with no problem but the errors are still there on the notebook. These cells create columns in the same way as demonstrated in lectures so we assumed it was fine.

OSError when creating networkx graphs

There is occasionally an OSError when saving the images that networkx generates. We don't know what the problem is but it doesn't seem to be anything to do with our code. In case the networkx graphs have this error, try running those cells again. The pngs for those graphs will also be provided in the event they cannot be reproduced.