Python Project Data Analysis with Pandas

Aayaan Ilyas

1 Overview

The objective of this project is to get an experience of using Python for data analysis. By completing this project, you should further enhance your Python programming skills and get more detailed insights into components of the Python ecosystem such as tools for data analytics and visualisation, and Jupyter notebooks, as well as into the practices of reproducible research using Jupyter notebooks.

2 The Dataset

The dataset contains information on more than 70,500 tweets with the hashtag #CometLanding used for the landing of the Rosetta's Philae lander on comet Churyumov-Gerasimenko on 12th November 2014. It was collected by Ernesto Priego (Centre for Information Science, City University London.

The dataset is provided as a .xlsx file and may be downloaded from [4]. In addition, the tab with the actual dataset has been exported into a CSV file. It is recommended to use this CSV file as input data and use the. xlsx only for reference.

3 Basic Requirements

In this project, you have to:

1. Check the consistency of the initial (raw) data.
2. In case of any inconsistencies, output refined data for further analysis.
3. Provide an executable Python script to automate the two steps above.
4. Implement unit tests to check at least some of the auxiliary functions.
5. Carry out certain data analysis and visualisation tasks.
6. Provide an executable Python script to regenerate all images and save them in a directory.

7. Provide a reproducible Jupyter notebook combining your report and data analysis.

8. Organise code with minimal duplication for its reuse in Jupyter notebook(s) and Python scripts. Your project should demonstrate the following characteristics

9. Correctness: your calculations and visualisations should be free from errors and describe the data in an accurate way;

10. Repeatability: you should be able to rerun easily the whole procedure from checking the consistency of the raw data to getting the final outcomes on the computer you use to work on the project;

You should first understand the structure of the raw data and the meaning of all information contained in the dataset. You do not have to make actual use of TAGS (https://tags.hawksey) and the Twitter API in case you may need to consult with it. You may start from opening the CSV file using the Jupyter notebook provided in Practicals/Python/notebooks/CometLanding.ipynb, As said, "There are several duplicates in this dataset. Requires refining", so you should refine it first.

You should use the following libraries:

- pandas (http://pandas.pydata.org/) - Python Data Analysis Library
- Matplotlib (http://matplotlib.org/) - Python plotting library

Keeping Jupyter notebooks under version control, you may discover that standard tools to inspect changes and perform merges do not work well with Jupyter notebooks. To have a smooth collaboration workflow, you will have to decide how to split your code between. py files and Jupyter notebooks in an optimal way (this will also facilitate code reuse in executable scripts and unit testing). You may use other specialised libraries as well. You can use Venv or virtualenv to have your own setup.

The minimal requirement for the project is to:

- Develop a procedure to check that the data matches expected format and remove duplicates
- In case any inconsistencies and/or duplicates are found, produce a new file with refined data to be used in the subsequent analysis
- This step must be automated to the point when it can be run with a single shell command to call an executable Python script specifying necessary arguments
- Develop a set of unit tests to cover at least some of the auxiliary functions you wrote, in order to increase the robustness of your code

- The refinement process should be documented in case one may need to modify and re-run it (although it's not necessary to repeat it each time while re-running the analysis)
- Perform the descriptive analysis of the dataset:
- Calculate the total number of tweets, retweets, and replies in the dataset
- Calculate the number of different users tweeting in this dataset
- Calculate the average number of tweets, retweets, and replies sent by a user
- Identify the most popular hashtags
- Build plots/visualisations for:
- The structure of the dataset (tweets/retweets/replies)
- The timeline of the tweets activity
- The word cloud for all other hashtags used in the tweets from the dataset
- Provide the Jupyter notebook to re-run the analysis, starting from refined data
- Provide an executable Python script to regenerate all images and save them in the images subdirectory

4 Additional Requirements

Note: It is strongly recommended to ensure that you have completed the Basic Requirements before you attempt to deal with the Additional Requirements.

- Easy: Analyse applications used to send tweets.
- Easy: Extend the descriptive analysis, for example, by calculating the average number of times each user is retweeted and the average number of times each user replies. - Medium: Analyse patterns of user activity over the period covered by the dataset.
- Medium to Hard: Analyse interactions between users by constructing, visualising (for example, using the networkx library, and analyzing the graph with vertices corresponding to users and edges corresponding to their connections by means of retweets, replies, and mentions. Determine some interesting properties of this graph and produce some visualisations.
- Hard: Implement interactive visualisiations (e.g. using sliders to control the parameters of the graph of user interactions).

5 Finally

The project is very open-ended, so be creative, have fun, and produce something that you would be interested in making and would be proud of! Inspiration taken from CS2006 Module