Introduction to Data Science
Coursework
Deadline: 12.00, Wednesday, 24th November, 2021
Submission: TBD

This assessment is worth 60% of the overall mark. This is an individual assessment. You are reminded of the University's regulations on plagiarism.

Brief

In this coursework you are asked to explore a large dataset of tweets collected from the Twitter API. Most of the tasks below should be possible to accomplish using the skills you have learned in this course. The assignment tests your ability to apply your skills and knowledge of data science to an unseen dataset. You may also need to do some self-guided learning to work out how to complete some of the tasks – this is an important part of data science in the real world!

There are several data analysis and critical reflection tasks to complete. Your submission should be in the form of a single pdf containing code, results, figures and text where appropriate. Do not simply copy paste your code into a document, include only important code snippets, the majority of the document should be text and images. Please separate and label the different tasks clearly to aid assessment.

The main aim of the coursework is to use the data supplied to find real-world events that happened in Europe during a specified period. As well as working through the technical challenges of handling a large and complex dataset, you will undertake several data analysis tasks to identify and characterise events.

Dataset

The dataset you will use consists of tweets collected from the Twitter API during the period June 1st to June 30th 2021. These tweets were collected using a geographical filter specifying a rectangular box over Europe. The bounding box used is specified by (longitude,latitude) coordinates. The lower-left corner is at (-24.5, 34.8) and the upper-right is at (69.1, 81.9). No keyword or other thematic filters were applied, so the dataset should contain all tweets that Twitter can identify as originating from the specified region, irrespective of their topic/content.The data is available for download as a number of compressed files each covering 1 hour of data. The whole dataset contains millions of tweets.

Each file contains a tweet on every line. Tweets are stored as JSON objects, as described in the Twitter developer documentation. In particular, as these are located tweets, pay attention to the difference between “place” tags and “coordinates” tags.

Files can be downloaded at the link below (you will need to login via your University account):
https://universityofexeteruk-my.sharepoint.com/:f:/g/personal/r_arthur_exeter_ac_uk/Et7hlFf8BE9PheegECWvu_gBeazWjyoHuHwz2p3aDvlI1w?e=qVjDiy

WARNING: With careful management it should be possible to do this analysis without a huge hard drive. The optimal strategy is not to simply decompress all the files.

HINT: Look at the *json* and *zipfile* modules in Python for processing the tweets.
HINT: One of the challenges with this coursework is handling large data files. You may wish to think about your strategy. The JSON files contain a lot of redundant information. Maybe make smaller JSON or CSV files by writing out just the fields you are interested in from the raw JSON files? Think about what you will need to do with the data to complete the coursework. A good strategy will require less storage and make the files quicker to process.
HINT: Process the JSON objects carefully, check things like timestamps and mandatory fields for consistency.

Tasks

Part 1. Basic Stats (20 marks)

1. Count the total number of tweets, describing how you deal with duplicates or other anomalies in the data set. [5 marks]
2. Plot a time-series of the number of tweets by day using the whole dataset and comment on what you see. [5 marks]
3. Using a box and whisker diagram (https://en.wikipedia.org/wiki/Box_plot) compare the average number of tweets on the weekdays in the dataset to the numbers for weekend days. Are there statistically significant differences between the number of tweets on weekdays and weekends? [5 marks]
4. Plot the average number of tweets at each hour of the day for weekdays and weekends and comment. You should have two plots where the x-axis is time of day (from midnight to midnight) and the y-axis shows the number of tweets. [5 marks]

Part 2. Mapping (20 marks)

1. Draw a map of Europe showing the location of the GPS-tagged tweets - these are tweets which have a “coordinates” field in the metadata. The exact form of the map is up to you: marks will be given for accuracy, clarity and presentation. [15 marks]
2. Explain any patterns you observe. [5 marks]

   HINT: Some useful libraries are geoPandas (https://geopandas.org/) and Folium (https://python-visualization.github.io/folium/), though you may use any libraries you like.

   HINT: Think carefully about issues like resolution and any other relevant information that should be on or attached to the map.

Part 3. Users (20 marks)

1. Make a histogram of tweets per user with number of users on the y-axis and number of tweets they make on the x-axis. Discuss the distribution that you see. All the users in the data set should be included! [5 marks]

2. Find the top-5 users by total number of tweets. Do you think any are automated accounts (aka. bots)? Justify your answer. [5 marks]
3. Find the 5 users who receive the most mentions and comment on this. [5 marks]
4. Calculate how often users in the UK, France, Germany, Italy and Turkey mention users in each of the other 4 countries. You should compute 25 numbers e.g. UK mentions UK, UK mentions France, France mentions UK etc. Comment on any patterns you observe. [5 marks]

Part 4. Events (20 marks)

1. Identify 3 days with unusually high activity in 3 different countries of your choosing. For example you could choose one day in the UK, one in France and one in Turkey. Describe and justify how you identify 'unusual' days. [5 marks]
2. Characterise each of these three days. Exactly how you do this is up to you, but for example you could:
    Display some indicative Tweets.
    Make a word cloud from the tweet text.
    Plot tweets locations on a map.
   Validate your conclusions with some other source of data e.g. government or news reports. [15 marks]

Part 5. Reflection (20 marks)

Using social media to study the real world is very common in academia, media and industry. Now that you have some experience analysing Twitter data discuss:

1. The strengths and weaknesses of Twitter as a data source from a technical/statistical perspective. [5 marks]
2. Biases in Twitter data and how they might be mitigated. [5 marks]
3. Ethical and legal concerns about using Twitter data. [5 marks]

Keeping in mind the issues raised discuss:

4. the use of Twitter to study the effectiveness of lockdown policies [5 marks]

Write no more than 300 words for each point.

Presentation

Documents which are extremely hard to navigate, messy or otherwise poorly presented will be penalised up to 10%.

Marking Guide

This is a general marking guide to how your document will be assessed. All criteria apply to all relevant questions.

Important:

- If 5 marks are available for a question, there may be 3 for the numeric or graphical output and 2 for the discussion. The ratio may be different for different questions.
- Partial marks are available for correct methods, so include working. **This does not mean including every line of code - summarise in words and code snippets what you did. Including large chunks of code will be penalised as bad presentation.**

Criterion	**What is expected for a good mark?**
Writing	Your writing should be clear, well-structured and concise.
Structure of the document	The structure should be clear, easy to navigate and with useful headings.
Presentation	Your document should conform to a clear and consistent visual style, be well-spaced, with appropriate font sizes and consistent and complete labelling and captions.
Code	We are interested in seeing relevant short code snippets that add meaning and context to your document. Your code should be well-structured and readable, with consistent naming conventions, good use of object-oriented or functional programming principles where needed and with an appropriate level of commenting to add context or explanation where it is needed.
Graphs and maps	Your graphic outputs should be well labelled and captioned, readable, meaningful and relevant.
Analysis	This is the most important area in which you will be assessed. Your analysis of the data should be thorough and we are looking to be impressed by your background research, verification of conclusions and exploration of the available techniques.

Explanation	Your methods and approaches should be clearly described and justified, and your comments and conclusions should be robust, valid and verified against additional sources where possible.

Submission

Your work must be submitted by 12pm (noon) on the hand-in date shown at the top of this descriptor. Please allow time for the submission process.

You should submit one PDF document containing all of your answers. Instructions for submission will be made available through the ELE page closer to the submission date.