

**Postgraduate Certificate in Software Design with Artificial Intelligence**

**Data Visualisation - (AL\_KSAIG\_9\_1)**

**Major Assignment – Visualising Twitter Data**

Student ID: A00267948

Student Name: Daniel Foth

GIT: <https://github.com/DanielsHappyWorks/Visualising-Twitter-Data>

Brief Description:

The assignment should contain your visualisation of the data obtained from David's assignment with

* an introduction explaining your data, its source, etc.
* a section outlining what software you used to create the visualisation
* the visualisations themselves with a short analysis for each visualisation.
* a conclusion bringing all your results together.

Contents

[Introducing the data 3](#_Toc38218548)

[Program Design 4](#_Toc38218549)

[Design Diagram 4](#_Toc38218550)

[The Program Files 4](#_Toc38218551)

[code/Tweet\_scraper.py 4](#_Toc38218552)

[code/Model\_generator.py 4](#_Toc38218553)

[code/Graph\_data.py 5](#_Toc38218554)

[code/Tweet.py 5](#_Toc38218555)

[Visualisations 5](#_Toc38218556)

[Conclusion 5](#_Toc38218557)

# Introducing the data

The data set is stored as data.csv in the code directory. It contains 513 of the 1500 tweets scraped by the tweet\_scraper.py script using the Twitter API. Some tweets have been removed as they were illegible and manually annotating that many tweets is time consuming, so the data had to be limited.

The tweet data that was saved consists of:

1. *Tweet ID*
2. *Creation Time*
3. *Retweet Count*
4. *Favourite Count*
5. *Source*
6. *Username*
7. *User Location*
8. *Content*
9. *Basic Content – this is the same as the content but stripped of special characters so it can be used with a variety of models.*

The two extra columns annotated manually are:

1. *Emotion – defines the emotion that stands out most in the tweet*
2. *Polarity – a rating of weather the tweet is Positive, Neutral or Negative*

Once models are created and data gets predicted, thses get stored as .pkl files in the /output/model\_data directory. They contain the tweet, model prediction, and Vader sentiment analysis prediction.

The unseen data used for testing the model predictions against Vader is always retrieved live. Once predictions are made, they get stored as .pkl files for graphing. Alternatively, they can be opened using the pandas “read\_pickle” function. These have the same properties as data.csv

The topics of interest include:

Politics:

1. *Donald Trump*
2. *Boris Jonson*

State of the world:

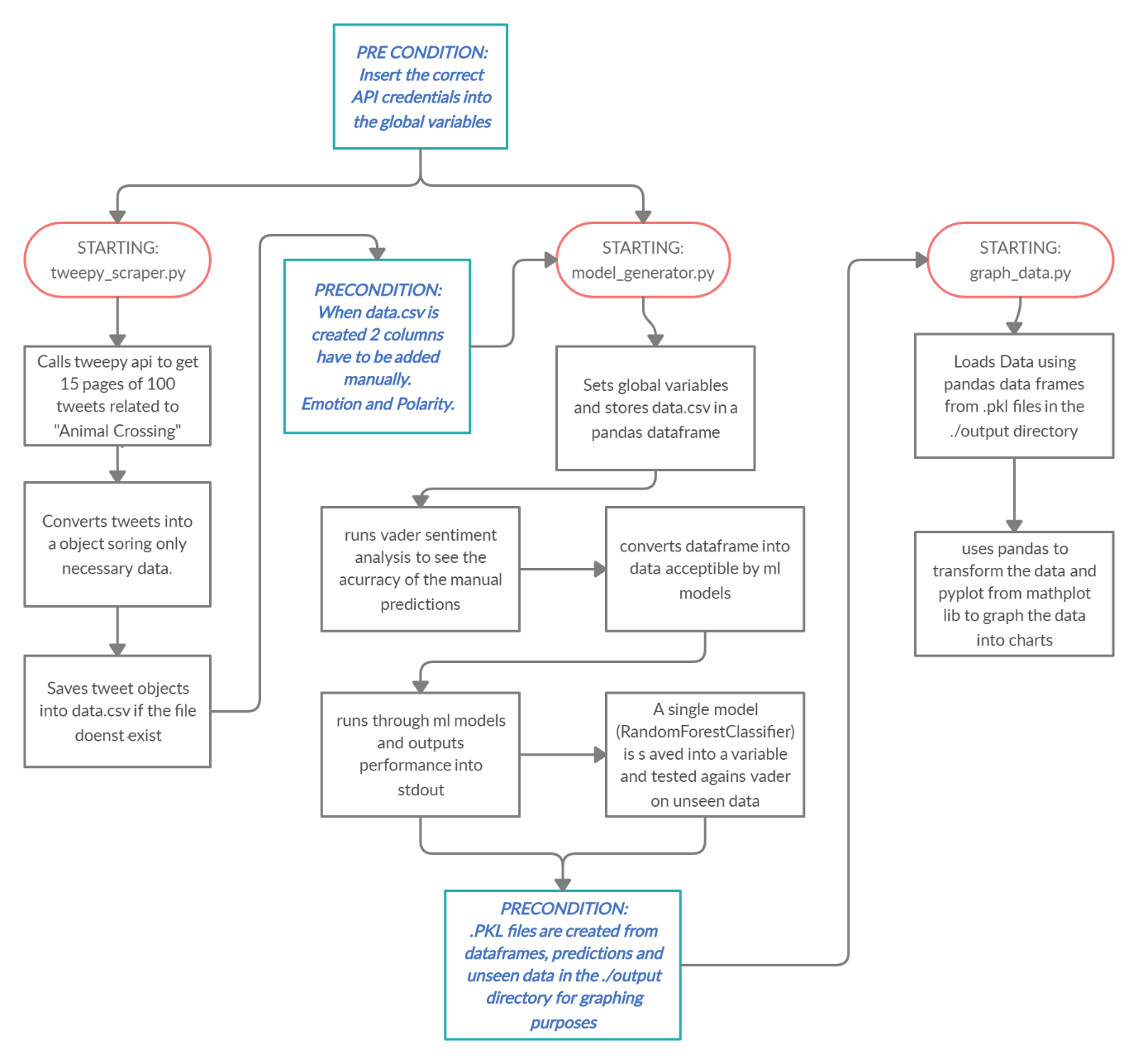
1. *Covid-19*
2. *Isolation*

New Technology

1. *Virtual Reality*
2. *Ps5*
3. *Xbox Series X*

# Program Design

## Design Diagram



This figure describes the general flow of the program from start to finish.

(Created with <https://creately.com/>)

## The Program Files

### code/Tweet\_scraper.py

It is responsible for scraping tweets from the tweepy API. It can scrape up to 100 tweets per page. When run it will generate a data.csv file with 1500 tweets on the topic of “Animal Crossing” if it doesn’t already exist.

### code/Model\_generator.py

It will create models and run one against multiple categories on unseen data. It uses the data.csv file. It requires that 2 extra columns are added in manually. The Emotion and Polarity column. The predictions made using the scikit learn library, get stored as data frames into .pkl files along with the tweet data for graphing purposes.

### code/Graph\_data.py

It graphs the data in the ./output/ directory and stores the graphs as png files under ./output/graphs.

The tools used for graphing are pandas data frames to format and audit the data from .pkl files and matplotlib which is a charting library for python.

Matplotlib can plot many kinds of graphs including bar, histograms, line and scatter plots.

### code/Tweet.py

The Tweet class processes a line of tweet data from the API into data that can be used as part of the model creation and graphing. See the data section to understand what is extracted.

# Visualisations

# Conclusion