wrangle_report

September 18, 2019

1 Project 1

1.1 Wrangling & Analysing

1.1.1 Introduction

This document briefly describes my experience when it comes to wrangling and analyses.

1.1.2 Reading Data

After completing the course "Introduction to Python", reading data from a text file is pretty simple, especially if you make use of pandas read_csv function.

Similarly, there is no complexity in reading a file from the web. It's much like reading a local file.

However, I got much satisfaction from gathering my own data from Twitter and reading that information into a file, before reading it into a data frame. The application of using an API to gather more data has many possibilities, the greatest of which is enriching data that we already have within RBS. The task of pulling data from a site like Twitter is not complicated. In fact, the whole process of creating the authorisation to pull the data was the most involved part of the whole process.

1.1.3 Assessing Data

Now that I had all the data I needed it was time to move on to assessing just how clean (or dirty and untidy) the data was.

This was probably the most time consuming part of the process (although cleaning data came very close). I often found myself asking how I wanted the data to be presented as this would inform how I should go about analysing it. For example, did I want to normalise the data I had or combine it all into one data structure. At times I wasn't sure how to answer it.

The flat file we had to read in (twitter_archive_enhanced.csv) took up most of my. I found numerous examples of ways to improve the data for analysis. Here are a few examples: * 4 columns used to describe the dog stage * html tags in the source column * multiple urls in the expanded_urls column

The image_prediction.csv file also had similar issues to the twitter_archive_enhanced.csv file. Namely, they used values as column names.

The final file was the simplest of all 3 and contained very minor issues. This took very little time to fix.

1.1.4 Cleaning Data

Just like assessing data this was a lengthy process, partly because I didn't have much knowledge around what to do. I often debated whether something needed cleaning or not and in a lot of cases I had no idea how to go about it programmatically.

This is where Google became a close friend. I often found myself looking at a problem and then typing that problem into the search engine hoping for an answer. I rarely got a straight forward answer and I had to spend a lot of time reading through blogs, panda documentation and forums - piecing together information from multiple sources.

I also referred to the course notes for help but I often spent to much time clicking through the lessons looking for that bit of information I needed, when it was easier to just Google it.

Once I found how to do something I discovered I would need it more than once and this was invaluable as I cemented the knowledge I was gaining.

I cleaned all the issues I found, however it was not straight forward. During the assessing phase I might have decided to clean a column one way, but as I progressed, I often found myself changing my mind.

1.1.5 Insight Analysis

Analysing the data for insights is probably where I am weakest at the moment. It is also the part I probably enjoy the least. Don't get me wrong, I do like discovering interesting bits about the data but I enjoy it if the process is more organic.

Having to purposely look for insight can sometimes feel like a chore for me. It's probably because I have had no real experience or knowledge. I tried to keep it simple and used basic algorithms to help me discover interesting bits about the data.

1.1.6 Visualising Data

Completely opposite to insight analysis, I actually enjoy trying to visualise the data. Just like my skills on insight analysis, I have very limited knowledge in the best way to visualise data.

In the end I had a good go at creating line, bar and word clouds to visualise my data.