7CCSMDM1 - Coursework 1

Adam Toner 1786775

1 - Text Mining

1.1

There are 5 possible sentiments that a tweet may have.

The second most popular tweet sentiment is Negative.

The day with the most extremely positive tweets was 25-03-2020.

1.2

With stop words, there are 1350959 words total, 80071 of which are unique.

The ten most frequently used words are: the, to, t, co, and, https, covid, of, a, in

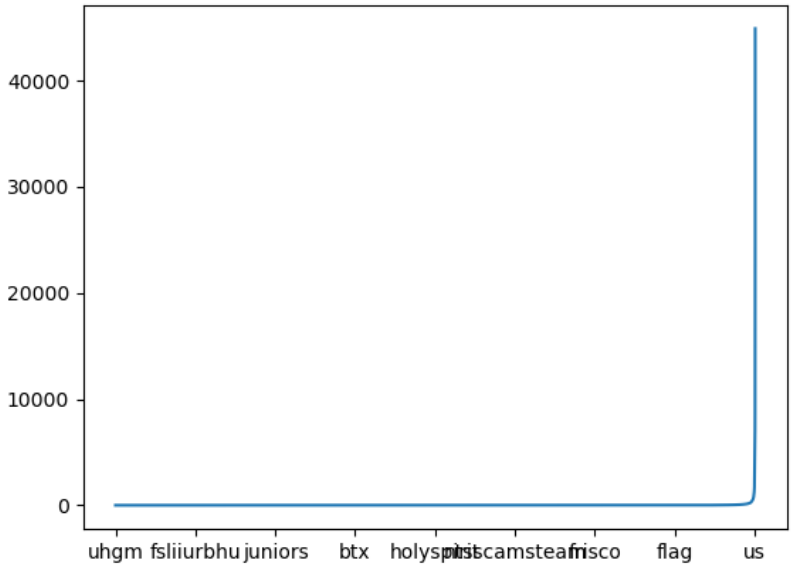
Without stop words, there are 1045810 words total, 79369 of which are unique.

The ten most frequently used (non-stop) words are: the, and, https, covid, coronavirus, for, are, you, this, prices

It is clear that there is a large part of the corpus that consists of stop words, relatively few of which are unique.

1.3

This is a graph showing the frequency of the words in the corpus, including the stop words.



It is clear that very few of the words make up a large proportion of the corpus. It is worth noting that the graph shown here do not show how many messages the words appear in, instead it shows the frequency of each word. This can be used to infer the size of the term-document matrix as it allows us to see the relative frequencies of the words, showing that a lot of the words appear a low number of times and so most of the term-document matrix will be mostly empty values. The dense term-document matrix would have one row for each tweet and one column for each non-stop word. Therefore the term-document matrix would have 79369 columns and 41157

1.4

The error rate of the classifier is 25.25%.

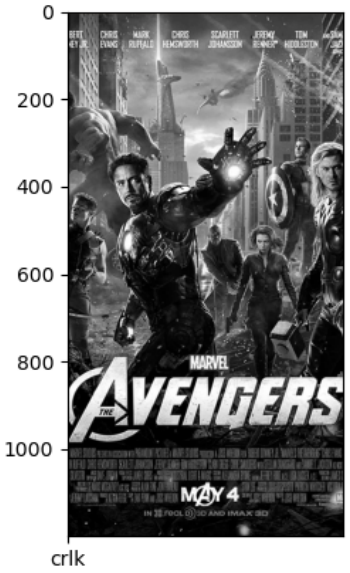
The total runtime for section 1 was 3.43 seconds.

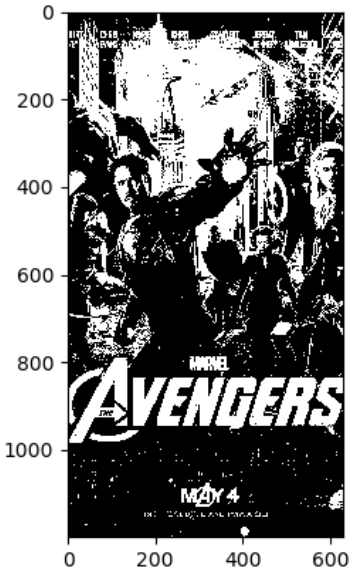
2 - Image Processing

2.1

The size of avengers_imdb.jpg is (1200, 630).

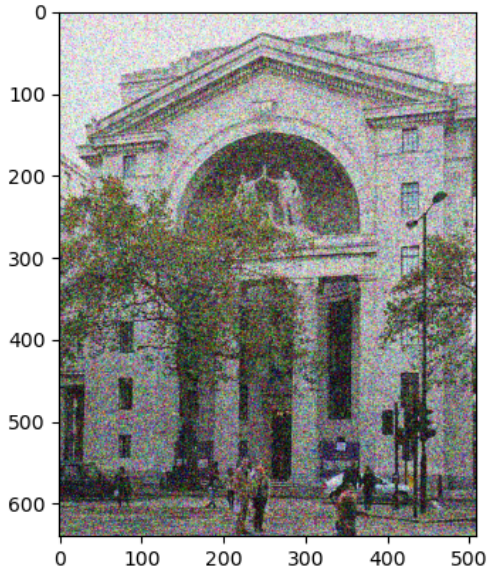
Here are the altered images:



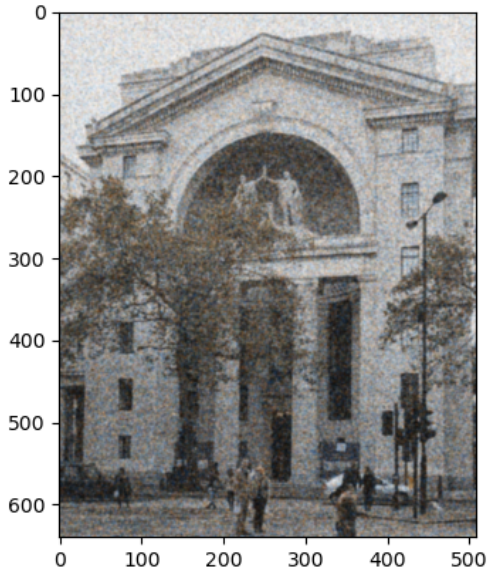


2.2

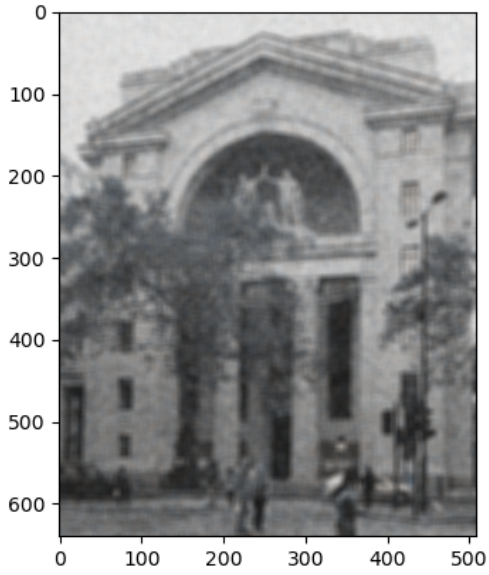
bush_house_wikipedia.jpg with gaussian noise:



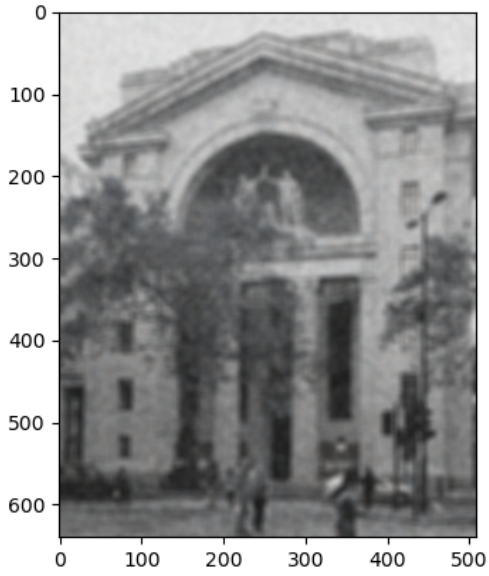
bush_house_wikipedia.jpg with gaussian blue:



bush_house_wikipedia.jpg with universal smoothing:

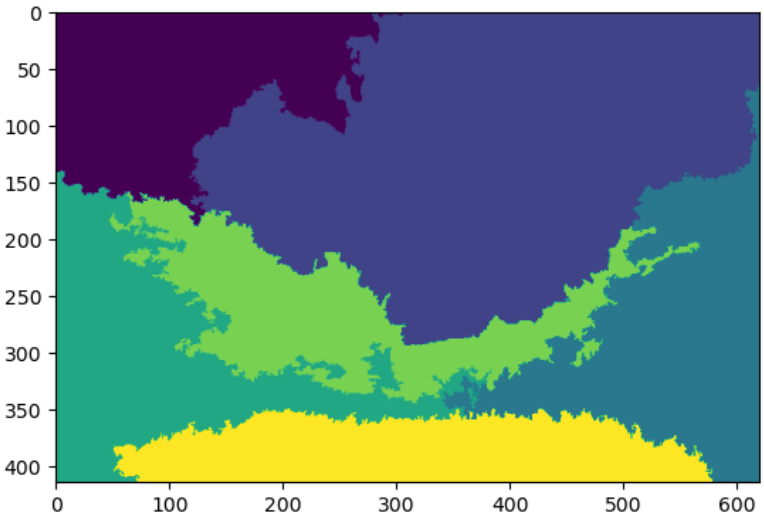


bush_house_wikipedia.jpg with gaussian blur and universal smoothing:



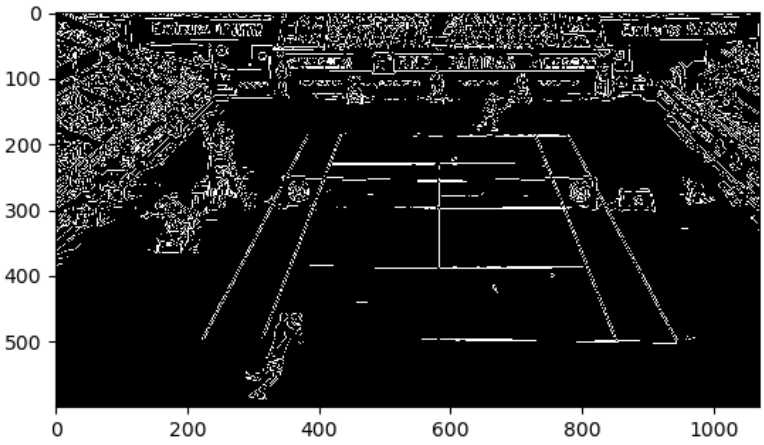
2.3

forestry_commission_gov_)uk.jpg split into 5 segments with k-mean segmentation:

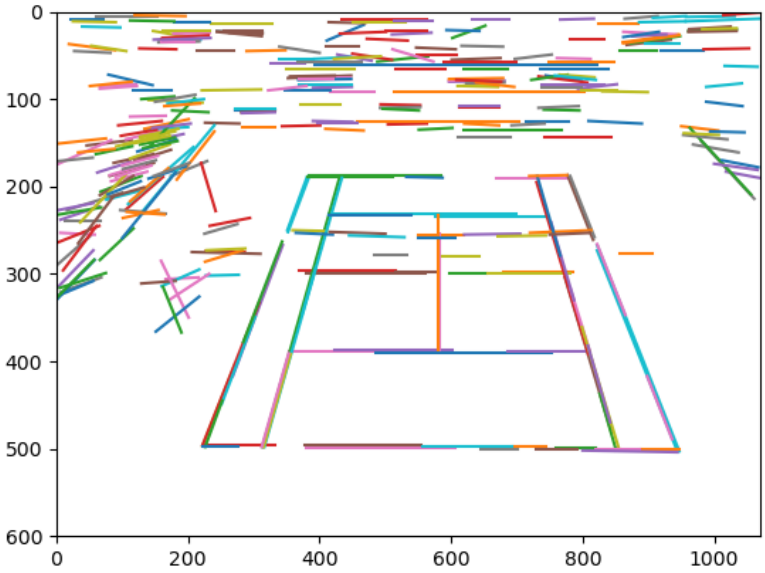


2.4

rolland_garros_tv5monde.jpg with Canny Edge Detection:



rolland_garros_tv5monde.jpg with Hough Transform:



Thank you for taking the time to read my project.