7CCSMDM1 Data Mining

Coursework 2

March 2021

1 Text Mining

1.1

A initial interrogation of the data in order to find, (1) the possible sentiments that a tweet may have, (2) the second most popular sentiment in the tweets and (3) the date with the greatest number of extremely positive tweets yielded the results shown in Table 1

Possible sentiments of a tweet:	Neutral, Positive, Extremely Negative, Negative, Extremely Positive
Second most popular sentiment:	Negative
Date with greatest number of extremely positive tweets:	25-03-2020

Table 1: Text mining 1.1 findings

1.2

The tweets were tokenized, the total number of all words, distinct words including repetitions were counted and the top ten most frequently used words in the corpus were found. Following this using the sklearn package stop words were removed along with words composed of two or less characters. Once this was achieved, the number of all words including repetitions and the top ten most frequently used words in the corpus were recalculated. The findings are shown in Table 2.

Total number of all words (including repetitions):	1229802
Total number of all distinct words:	52144
10 most frequent words in the corpus:	the, to, and, covid, of, a, in, coronavirus, for, is
Total number of all words (including repetitions) (After stop words and words with $\leq$ 2 characters have been removed):	678561
10 most frequent words in the corpus (After stop words and words with $\leq$ 2 characters have been removed):	covid, coronavirus, prices, food, supermarket, store, grocery, people, amp, consumer

Table 2: Text mining 1.2 findings

1.2.1 Analysis of findings

Stop words can usually be defined as the most common words within a language. Once the stop words along with words which were $\leq$ 2 characters in length were removed, the total number of all words in the corpus including repetitions decreased by circa 45% from 1229802 to 678561. Additionally the most frequent words prior to the removal of stop words and words with length $\leq$ 2 characters were comprised predominantly of syncategorematic words, words which "do not stand by themselves" usually including articles, connectives, prepositions etc... After the removal the top ten most frequent words where in the majority comprised of nouns.

1.3

A line chart was plotted mirroring the words in the corpus and the frequency of their appearance. The x axis corresponds to words and the y axis indicates the fraction of documents in which a word appears. Figure 1 and Figure 2 below display the findings.

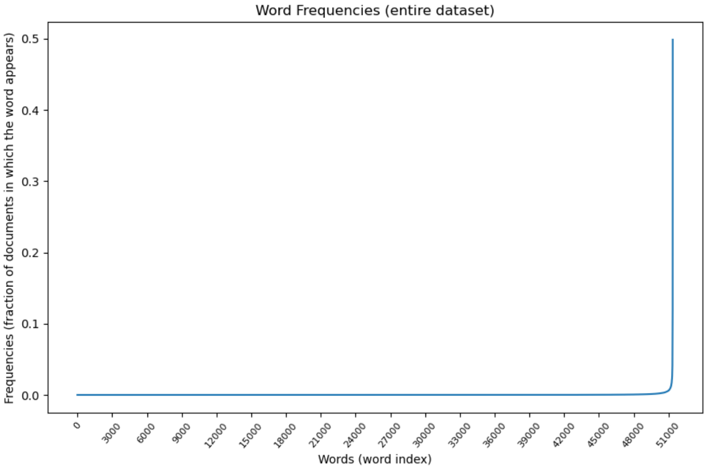

Figure 1: Word frequencies (entire dataset)

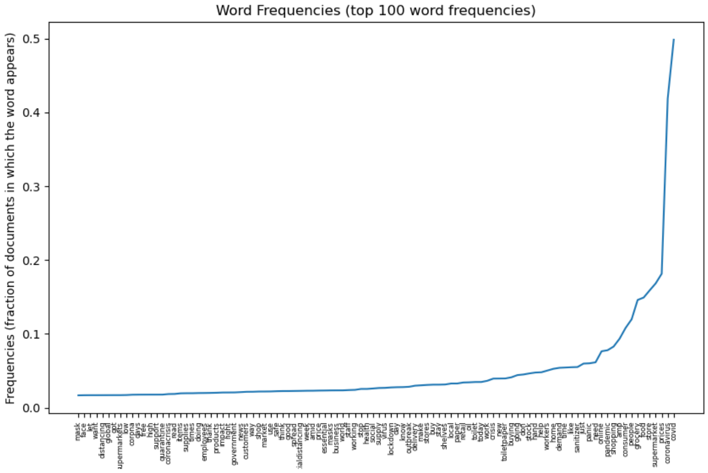

Figure 2: Word frequencies (top 100 words by frequencies)

1.3.1 Analysis of findings

A Term Document Matrix tracks the term frequency for each term by each document. Beginning with a Bag of Words representation of the documents and then for each document tracking the number of times a term exists. A Term Document Matrix can become a very large sparse matrix depending on the number of documents in the corpus and the number of terms within each document, which as a consequence becomes computationally expensive to mine.

Analysing the findings of Figure 1, it is apparent that the majority, circa 98% of the 52000 words within the set of distinct words of the corpus appear once or infrequently and therefore serve a reduced use for classification. Figure 2, suggests that 100 words of the set of distinct words appear with increased frequency. The the size of the term document matrix can be adapted to a size which reflects these findings, a term document matrix of size 1000 or circa 2% of the current set of distinct words would be inline with these findings.

1.4

Using the scikit-learn package, a Multinomial Naive Bayes classifier was produced for the Coronavirus Tweets NLP data. The accuracy and therefore the error rate was calculated using the training set exclusively. Table 3 reflects the results.

Multinomial Naive Bayes classifier error rate:	25.89%

Table 3: Text mining 1.4 findings

2 Image Processing

2.1

Retrieving the shape of the *avengers_imdb.jpg* image returned the data shown in Table 4.

Size of *avengers_image.jpg* **image:**	(1200, 630, 3)

Table 4: Image processing 1.1 findings

The data suggests that the *avengers_imdb.jpg* image has a size of 1200 rows, 630 columns and 3 colours suggesting RGB colour.

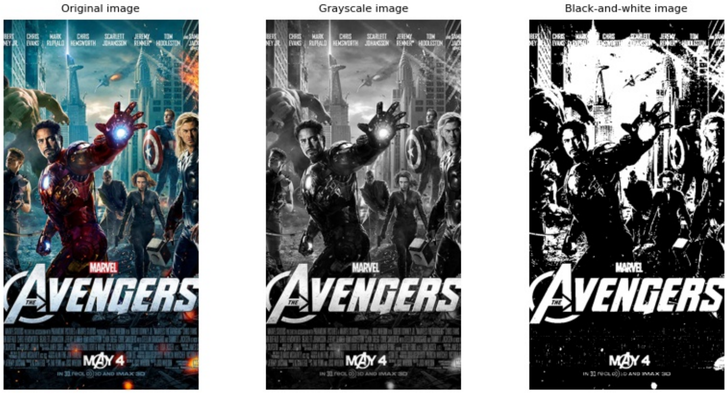

Figure 3: Avengers image transformation

Figure 3 depicts transformation of the avengers image, transforming the image to a grayscale and black-and-white representation. The two representations where undertaken using the skimage package, using specifically the rgb2gray and threshold_otsu methods accordingly.

2.2

The *bush_house_wikipedia.jpg* image was transformed three times, (1) Gaussian random noise (with variance 0.1) was added to the image, (2) filtering the image with a Gaussian mask (sigma of 1) and (3) apply a uniform smoothing mask (size 9x9). These transformation were applied sequentially to the output of the previous transformation. Figure 4 shows the original image and the resulting output from each transformation.

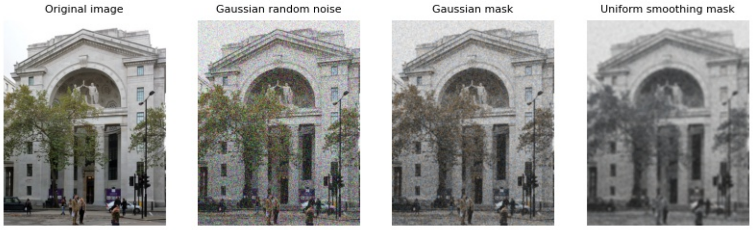

Figure 4: Bush House image transformation

2.3

The slic method from the skimage package was used to perform k-means clustering on the *forestry_commission_gov_uk.jpg* image with a goal of dividing the image into 5 segments. Figure 5 depicts the clustering output.

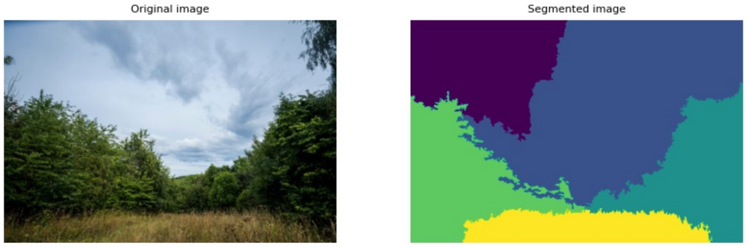

Figure 5: Forestry commission clustering output

2.4

Canny edge detection and Hough transform were applied to the *rolland* image. Probabilistic Hough Transform was the chosen Hough Transform method applied to the image. Inline with the process outlined in section 2.2 these two methods where applied sequentially taking as input the output of the previous step. The resulting outputs of these two methods are shown in Figure 6.

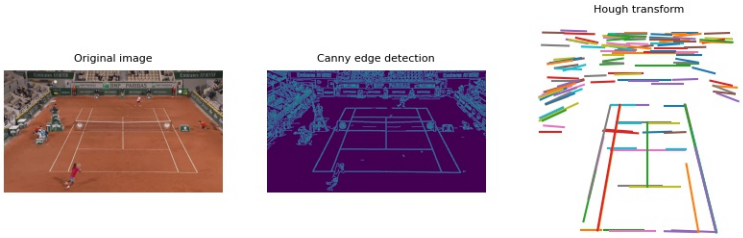

Figure 6: Forestry commission clustering output