**Data Analytics HW2**

**Q7) (a)** We have used Euclidean distance as the distance measure as it provides the best F- measure as compared to other metric such as manhattan distance. Minkowski also provides the similar results as it’s the generalized version of Euclidean. The results are shown as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Distance Metric** | **Accuracy** | **Precision** | **Recall** | **F-measure** |
| Euclidean | 43 | 33.3 | 86 | 48 |
| Minkowski | 43 | 33.3 | 86 | 48 |
| Manhattan | 43 | 0 | 93 | 0 |

**(b)** Yes, we can run KNN on our dataset. We have performed below steps to run KNN in our dataset.

* We have collected hurricane Harvey related tweets from twitter.
* We labelled the tweets in two classes as rescue tweets and non-rescue tweets. We labelled the rescue tweets as the tweets containing specific keywords such as **help, relief, hurricane, Houston** etc. and rest of them as non-rescue tweets.
* We preprocessed the tweets with following steps:
  + Lowercasing all the words and hashtags
  + Remove url’s, punctuation marks and white spaces.
  + Remove the commonly occurring stop words such as I, we, me etc.
  + Stem the words like helped to help.
* After preprocessing we have a set of cleaned tweets with words and hashtags only.
* We divided the dataset in training and testing set.
* We created the document feature matrix by considering training dataset as Bag of Words. All the tweets are represented as rows and each word/hashtag is represented as column. Entry in the cell is the occurrence of word/hashtag in the tweet.
* The feature set generated in training set are used in the testing set too so as to ignore the words/hashtags in the test set not present in the training set.
* We defined and chose the distance metric for KNN and applied them to determine the distance between the test set and training set instances.
* We applied Knn algorithm and label the tweets in the test set.
* We evaluate our classifier using various metrics.

We have divided our dataset in 80-20 ratio i.e. 80% training and 20% testing set. So out of 150 tweets, 120 are in training set and 30 are in testing set. Due to less training data, it might not perform well in this scenario.

**(c)** We run the knn classifier model for different values of k and chose the one with the highest F-measure.Yes, we can evaluate our KNN classifier with precision, recall and F-measure.

After running KNN in our dataset with k=3 we get below values during evaluation:

**Precision**: 33.3%

**Recall**: 86%

**F-measure**: 48%

**(d)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **k** | **Accuracy** | **Precision** | **Recall** | **F-measure** |
| 1 | 50 | 60 | 86 | 70.6 |
| 3 | 43 | 33.3 | 86 | 48 |
| 5 | 43 | 0 | 93 | 0 |
| 8 | 40 | 25 | 79 | 38 |
| 10 | 40 | 33.3 | 71 | 45 |

(e) Visualization of classification results

