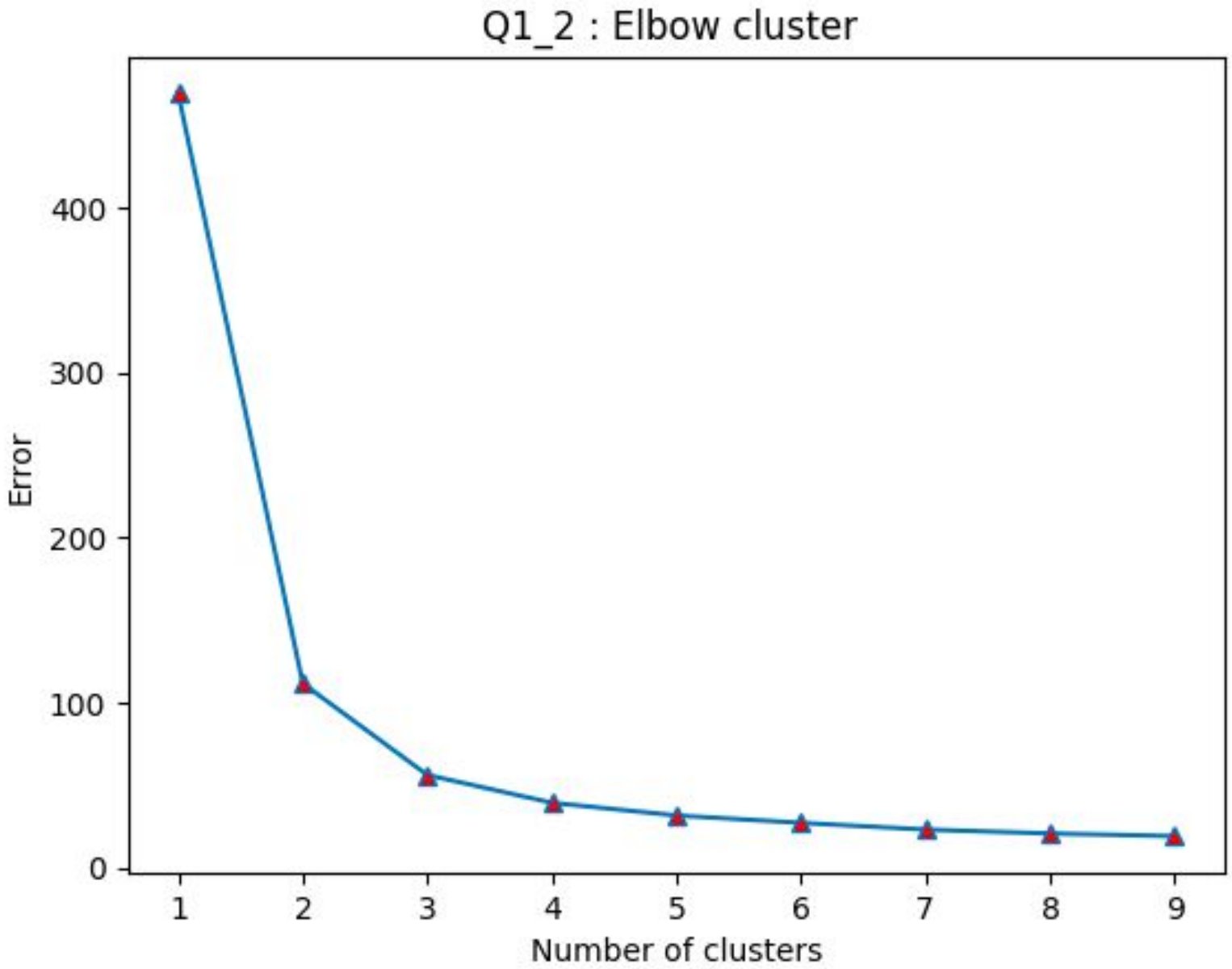
ML - Assignment 4

-Harkishan Singh (2017233)

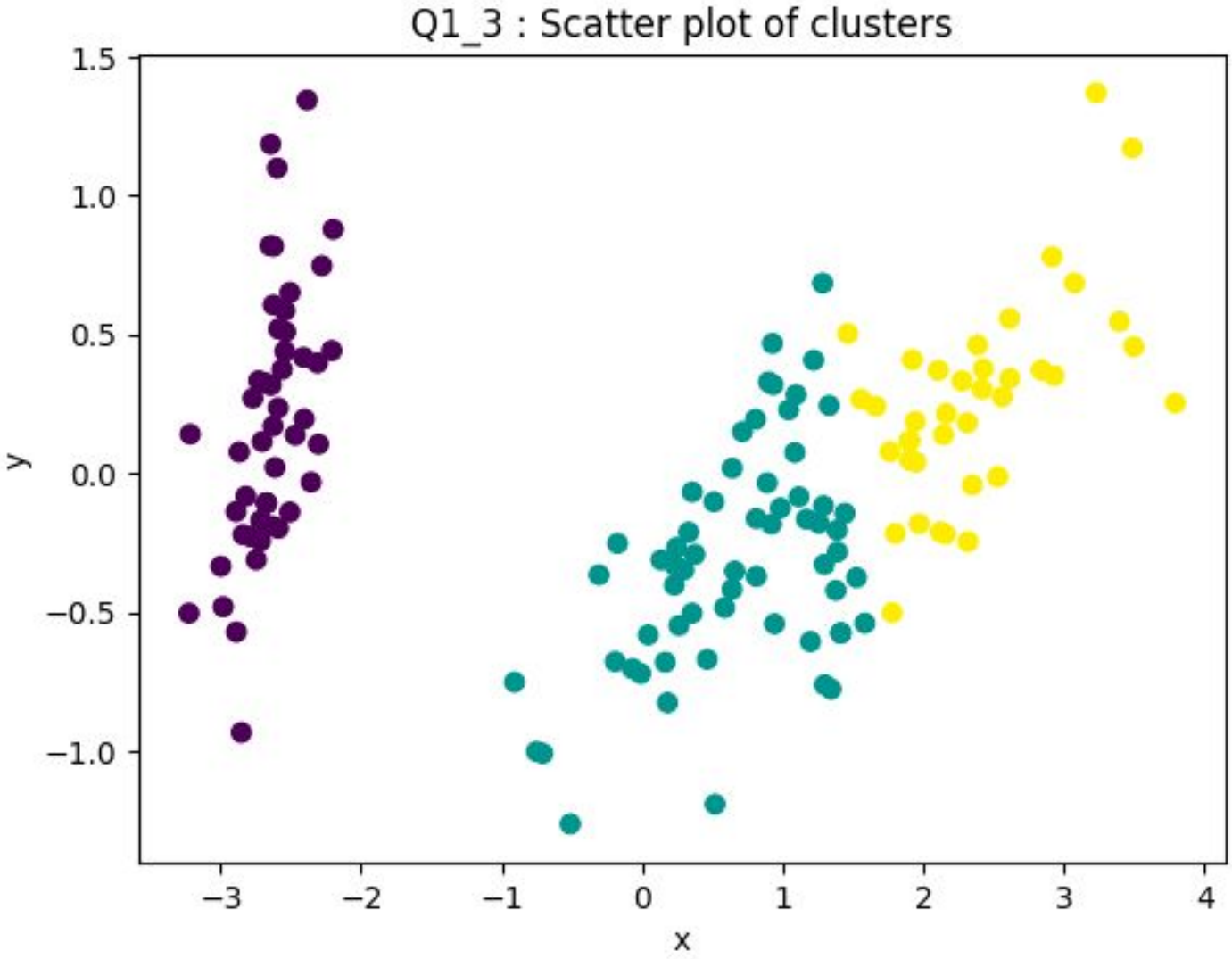
Question 1 : KMeans

1) See file ‘Q1_1.py’. Data loaded and splitted according to the question.
2) See file ‘Q1_2.py’. I have plotted the graph of the error I am getting at each cluster point. Below is the graph :



It is clear from the graph that at clusters 2 and 3, we can see a flattening and the graph is making an elbow at **k = 3**. So, 3 clusters is the optimal value of clusters for iris data.

3) For this question, I have first reduced the dimension of the data to 2 using PCA, so that I can plot the scatter plot. Below is the scatter plot:



4) Following are the train and test accuracy :

```
Train Accuracy =  0.9047619047619048
Test/Validation Accuracy =  0.8666666666666667
```

Train accuracy : 90% and test accuracy : 86%

We have labelled data and I have just compared the outcomes of KMeans with the labelled data.

Question 2 : Naive Bayes

1) See file “Q2_1.py”. In this question, firstly I loaded the data “yelp_labelled.txt”.
    a) Loaded the data while lowering all the documents (sentences).
    b) Removed punctuations from documents (function : remove_punctuations()).
    c) Tokensied document so that each string is in the form of a list.
    d) Removed stopword from the data (function : remove_stopwords())
    e) Returned splitted data. (70 : 30)
2) Done in 1st part
3) See file “Q2_3.py”.
    a) Created a vocabulary of unique words from the train data.
    b) Created a matrix with element (di * wj) ⇒ di is document i and wi is the frequency of word wj in document di.
4) Following are the results :

```
Train Accuracy =  0.9657142857142857
Test/Validation Accuracy =  0.73
```

Train accuracy is 96.5% and test accuracy is 73%

Following are some of the misclassified documents from test data :

.

- ➔ Not my thing. ⇒ ['thing'] 0
- ➔ 2 Thumbs Up!! ⇒ ['2', 'thumbs'] 1
- ➔ After one bite, I was hooked. ⇒ ['one', 'bite', 'hooked'] 1
- ➔ A FLY was in my apple juice.. A FLY!!!!!!!! ⇒ ['fly', 'apple', 'juice', 'fly'] 0

Potential reason for misclassification is that T

- ➔ This method is losing the context of a statement and only looking at the position of words and by that it is trying to predict the test cases.
- ➔ Another reason can be that the vocabulary which I made from the training data is not a complete vocabulary and some of the words from the testing data is not present in the vocabulary.
- ➔ The training set could be noisy as we only see 96% accuracy on the training data.