Assignment 4

Answer 1:

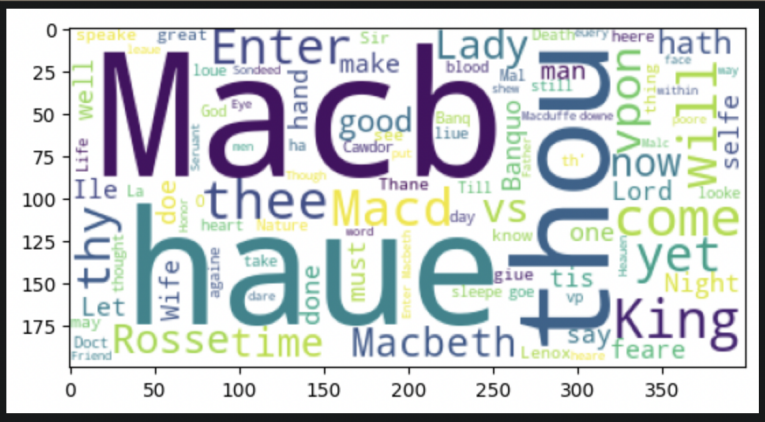
The average word length is 3.97.

The average sentence length is 4.4.

The number of times the word “King” appears is 15.

Word Cloud for top 100 tf-idf score

```
In [114]: import matplotlib.pyplot as mpLib
          mpLib.imshow(wordcloud)
          mpLib.axis("on")
          mpLib.show()
```



Answer 2:

Done the binary classification for ‘MCAT’ positive class.

Accuracy and confusion matrix with logistic regression:-

```
In [25]: from sklearn.metrics import accuracy_score, hamming_loss, f1_score

         print('Accuracy Score is {}'.format(accuracy_score(rcv1_target_test, log_reg_predicti

         Accuracy Score is 0.9998549667666705
```

```
In [26]:
         from sklearn.metrics import confusion_matrix

         #confusion matrix for easy visualization
         matrix = confusion_matrix(rcv1_target_test, log_reg_prediction)

         print(matrix)

         [[241289      0]
          [    35      0]]
```

Accuracy and confusion matrix with SVM:-

```
In [29]: from sklearn.metrics import accuracy_score

         print('Accuracy Score is {}'.format(accuracy_score(rcv1_target_test, svm_prediction)))

         Accuracy Score is 0.9998549667666705
```

```
In [30]:
         from sklearn.metrics import confusion_matrix

         #confusion matrix for easy visualization
         matrix = confusion_matrix(rcv1_target_test, svm_prediction)
         print(matrix)

         [[241287      2]
          [    33      2]]
```