CENG 463
Assignment 1

Cem Gündoğdu

cem.gundogdu@metu.edu.tr

December 16, 2021

Contents

1 Introduction

In this assignment, I used the `NaiveBayesClassifier` class from the NLTK library for Naive Bayes Classifier, and `SklearnClassifier` class from NLTK along with `SVC` class from scikit-learn for the Support Vector Classifier.

I have tried various preprocessing operations, and logged the accuracy at each step, along with the confusion matrix and metrics such as precision.

For the most of the assignment, I used the count of single words as features. In the end, I tried changing the features to bigrams. I list its results at the end of this report.

2 Different preprocessing operations

In this section, I list the accuracies for different preprocessing techniques. I haven't tried all combinations of them, since the number grows exponentially. However, I added each technique **incrementally** (added the preprocessing step without removing previous ones), with the hope that the change caused by the added technique will be representative of its usefulness.

These trials were made on the *dev* set, since I did them during the development phase. The evaluation in Evaluation will be done on the *test* set.

Complete logs from each try is available in the data directory. These include the confusion matrix, accuracy, and for each category the precision, recall and F_1 measure. I include only the last log here, to keep this report short.

Details of the calculation of precision, recall and F_1 measure are given in Calculation of metrics.

A table of the accuracies for each technique is in Table 1. I recorded these values during development, so these are for the *dev* set. As seen in the table, each techniques improved the accuracy in SVC. But for Naive Bayes, removing the punctuation and stemming caused a decrease in accuracy.

Table 1: Accuracies for different preprocessing techniques.

Preprocessing type	Accuracy (%)	
	Naive Bayes	Support Vector
Simplest version	64.20	64.09
Lowercase conversion	65.92	66.77
Punctuation removal	65.38	68.70
Stopword removal	67.63	70.74
Stemming	66.67	72.34
Removing short words	66.99	72.78

2.1 Simplest version

The words in title are counted twice — in this version and in all the following ones. Other than this, there is no processing in this version. The text is given to `nltk.word_tokenize()` and the resulting tokens are used to train the classifier.

2.2 Lowercase conversion

Converted all words to lowercase.

This improved accuracy in both SVC and NBC. I think this is because our corpus size is rather small. So, this change allows us to make better use of the limited data.

I think, on a larger corpus, keeping the case could be more beneficial, since it would allow us to distinguish the words in the title from the words in the body.

2.3 Punctuation removal

Removed the following characters from the text:

```
!"#$%&'()*+,./:;<=>?@[\]^_`{|}~'
```

I did not replace them with spaces, but simply removed them. This made, for example, the text *Sophies's* to become the single word *sophies*.

Removing the punctuation caused a decrease in the accuracy of NB classifier. I was expecting a decrease, since punctuation can actually be helpful in understanding the book's genre. For example, one could expect to see more question marks in a mystery book's description. However, you wouldn't expect lots of exclamation marks in a science book, unless the author got very excited about whatever scientific topic they were writing about.

SVC's accuracy increased.

2.4 Stopword removal

I used the corpus `nltk.corpus.stopwords.words('english')` from the NLTK library. I removed every word that occured in this list.

This improved the accuracy of both classifiers.

2.5 Stemming

I passed each word to the `nltk.stem.PorterStemmer()` from the NLTK library.

This decreased the accuracy of NB, and resulted in a significant improvement in SVC.

2.6 Removing short words

I thought stopword removal should have been enough, but I also tried removing words that were shorter than 3 characters. Surprizingly, this resulted in an increase in the accuracy of both classifiers.

3 Evaluation

I evaluated the classifiers with the best preprocessing operations based on the *dev* set. For SVC, I kept all the operations listed in Different preprocessing operations. For NB, I deleted the stemming and punctuation removal code.

3.1 Calculation of metrics

Recall is the ratio of true positives for a class to the number of input documents of that type. To find recall, we divide each diagonal entry by the sum of corresponding row.

Precision is the ratio of true positives for a class to the number of documents that are identified to be in that class. To calculate it, we divide diagonal entries by the sum in that column.

F_1 measure is the harmonic mean of precision and recall.

These three per-category metrics listed in Metrics for SVC and NBC were calculated by the following code:

```
precision = true_positives / (true_positives + false_positives)
recall = true_positives / (true_positives + false_negatives)
f1_measure = 2 * precision * recall / (precision + recall)
```

Confusion matrices were created by the `nltk.classify.util.ConfusionMatrix` class from the NLTK library.

3.2 Metrics for SVC and NBC

The confusion matrix and accuracy of the classifiers; along with the precision, recall and F_1 measure for each category are given in Figure 1 and Figure 2.

4 Comparison

NB and SVC have a similar accuracy with this choice of features. However, they respond to changes in the preprocessing step differently.

5 Discussion of confusion matrices

As the matrices in Figure 1 and Figure 2 show, some genres are more difficult to distinguish.

For example, horror and science-fiction books are not distinguished by NB very precisely. Similarly, many mystery books were classified by the NBC as horror books. But not as many horror books were not classified as mystery books. This shows that the classifier tends to choose horror more often, when the description is not very helpful.

There are also genres that are easily distinguished. For example, no mystery or horror books were classified wrongly as philosophy books by NBC. Similarly, three are very few false positives for the sports category.

6 Bigrams as features

I followed a "bag of words" approach in the tests mentioned above. I used the single words (or punctuation characters) as features, along with their counts.

In this section, I will show the results I got, when I used word pairs as features. This is similar to the 2-gram language models we discussed.

I implemented this change on the final versions of the code discussed above. That is, I combined the adjacent words after removing stopwords.

In Naive Bayes Classifier, this gave a decent result, but it wasn't as good as the results from the bag of words approach. In the SVC, results were horrible. About all documents were being classified as "horror", which was the first in the list of labels. I tried SVC without stemming, but it didn't help.

Results of this modification are shown in Figure 3 and Figure 4 with red background.

Figure 1: Confusion matrix and other metrics for the best NB version on test data set

NB - Bag of words

```
Loaded classifier from cache.
Accuracy: 0.6738197424892703
                |                           s     |
                |                           c     |
                |                           i     |
                |                           e     |
                |                           n     |
                |           p               c     |
                |           h               e     |
                |           i   r           -     |
                |       m   l   e   r   s   f     |
                |   h   y   o   l   o   c   i   s |
                |   o   s   s   i   m   i   c   p |
                |   r   t   o   g   a   e   t   o |
                |   r   e   p   i   n   n   i   r |
                |   o   r   h   o   c   c   o   t |
                |   r   y   y   n   e   e   n   s |
----------------+---------------------------------+
         horror |<164> 19   .   3  18   2  27   1 |
        mystery |  34<171>  .   4  14   .  15   2 |
     philosophy |   4   2<147> 32   3  30   9   1 |
      religion |  13   5  46<132>  2  13  19   . |
        romance |  19   9   1   .<148>  .  26  25 |
        science |   4   6  26  12   .<166> 16   . |
science-fiction |  35  11   6   3  11   8<165>  1 |
         sports |  14   1   1   2  41   6   6<163>|
----------------+---------------------------------+
(row = reference; col = test)



                 |  Precision  |     Recall  | F1-Measure
-----------------------------------------------------------
         horror  |     0.5714  |     0.7009  |     0.6296
        mystery  |     0.7634  |     0.7125  |     0.7371
     philosophy  |     0.6476  |     0.6447  |     0.6462
      religion   |     0.7021  |     0.5739  |     0.6316
        romance  |     0.6245  |     0.6491  |     0.6366
        science  |     0.7378  |     0.7217  |     0.7297
science-fiction  |     0.5830  |     0.6875  |     0.6310
         sports  |     0.8446  |     0.6966  |     0.7635
```

Figure 2: Confusion matrix and other metrics for the best SVC version on test data set

SVC - Bag of words

```
Loaded classifier from cache.
Accuracy: 0.7194206008583691
                |                           s     |
                |                           c     |
                |                           i     |
                |                           e     |
                |                           n     |
                |           p               c     |
                |           h               e     |
                |           i   r           -     |
                |       m   l   e   r   s   f     |
                |   h   y   o   l   o   c   i   s |
                |   o   s   s   i   m   i   c   p |
                |   r   t   o   g   a   e   t   o |
                |   r   e   p   i   n   n   i   r |
                |   o   r   h   o   c   c   o   t |
                |   r   y   y   n   e   e   n   s |
----------------+---------------------------------+
         horror |<172> 22   .   1  14   2  22   1 |
        mystery |  27<182>  .   2  18   .  11   . |
     philosophy |   5   .<153> 26   2  36   5   1 |
       religion |  13   2  36<150>  6   9  14   . |
        romance |  18   5   2   2<162>  .  20  19 |
        science |   4   5  21  10   .<175> 15   . |
science-fiction |  36  10   3   3  13   8<166>  1 |
         sports |   8   1   .   1  34   4   5<181>|
----------------+---------------------------------+
(row = reference; col = test)



                 |  Precision  |     Recall  | F1-Measure
-----------------------------------------------------------
         horror  |     0.6078  |     0.7350  |     0.6654
        mystery  |     0.8018  |     0.7583  |     0.7794
     philosophy  |     0.7116  |     0.6711  |     0.6907
       religion  |     0.7692  |     0.6522  |     0.7059
        romance  |     0.6506  |     0.7105  |     0.6792
        science  |     0.7479  |     0.7609  |     0.7543
science-fiction  |     0.6434  |     0.6917  |     0.6667
         sports  |     0.8916  |     0.7735  |     0.8284
```

Figure 3: NB with word pairs as features

NB - Word pairs

```
Loaded classifier from cache.
Accuracy: 0.5718884120171673
                |                           s     |
                |                           c     |
                |                           i     |
                |                           e     |
                |                           n     |
                |           p               c     |
                |           h               e     |
                |           i   r           -     |
                |       m   l   e   r   s   f     |
                |   h   y   o   l   o   c   i   s |
                |   o   s   s   i   m   i   c   p |
                |   r   t   o   g   a   e   t   o |
                |   r   e   p   i   n   n   i   r |
                |   o   r   h   o   c   c   o   t |
                |   r   y   y   n   e   e   n   s |
----------------+---------------------------------+
         horror | <95> 35   3  11  37   6  33  14 |
        mystery |  22<157>  2   4  35   4   8   8 |
     philosophy |   5   3<124> 40   2  42   8   4 |
      religion |   5   4  44<117> 12  27  12   9 |
        romance |  10  11   1   7<130>  3   9  57 |
        science |   3   2  21  17   3<173>  5   6 |
science-fiction |  34  20   9  10  33  21 <92> 21 |
         sports |   .   1   3   4  46   1   1<178>|
----------------+---------------------------------+
(row = reference; col = test)



                |  Precision  |     Recall  | F1-Measure
-------------------------------------------------------------
     philosophy |     0.5990  |     0.5439  |     0.5701
         sports |     0.5993  |     0.7607  |     0.6704
        mystery |     0.6738  |     0.6542  |     0.6638
       religion |     0.5571  |     0.5087  |     0.5318
        science |     0.6245  |     0.7522  |     0.6824
        romance |     0.4362  |     0.5702  |     0.4943
         horror |     0.5460  |     0.4060  |     0.4657
science-fiction |     0.5476  |     0.3833  |     0.4510
```

Figure 4: SVC with word pairs as features

SVC - Word pairs

```
Loaded classifier from cache.
Accuracy: 0.20386266094420602
                |                           s     |
                |                           c     |
                |                           i     |
                |                           e     |
                |                           n     |
                |           p               c     |
                |           h               e     |
                |           i   r           -     |
                |       m   l   e   r   s   f     |
                |   h   y   o   l   o   c   i   s |
                |   o   s   s   i   m   i   c   p |
                |   r   t   o   g   a   e   t   o |
                |   r   e   p   i   n   n   i   r |
                |   o   r   h   o   c   c   o   t |
                |   r   y   y   n   e   e   n   s |
----------------+---------------------------------+
         horror |<193> 15   .   1   4   .  21   . |
        mystery | 205  <8>  .   2   6   .  18   1 |
     philosophy | 133   . <23> 18   .  15  39   . |
       religion | 130   .  31 <10>  4   3  52   . |
        romance | 162   2   .   . <11>  .  43  10 |
        science | 110   2  16  10   3  <8> 81   . |
science-fiction | 109   3   3   1   5   2<117>  . |
         sports | 173   1   .   .  19   .  31 <10>|
----------------+---------------------------------+
(row = reference; col = test)



                |  Precision  |     Recall  | F1-Measure
----------------------------------------------------------
         horror |     0.1588  |     0.8248  |     0.2664
        mystery |     0.2581  |     0.0333  |     0.0590
     philosophy |     0.3151  |     0.1009  |     0.1528
       religion |     0.2381  |     0.0435  |     0.0735
        romance |     0.2115  |     0.0482  |     0.0786
        science |     0.2857  |     0.0348  |     0.0620
science-fiction |     0.2910  |     0.4875  |     0.3645
         sports |     0.4762  |     0.0427  |     0.0784
```

7 Logarithm of word counts in Naive Bayes

I also tried taking the logarithm of word counts before passing them to the Naive Bayes Classifier. I thought using logarithms was a good idea, since it is not that important, whether a word occurs 10 times or 20 times.

This method gave a slightly worse result. The results are shown in Figure 5 with green background.

Figure 5: Bayes with the logarithm of word counts

SVC - Word pairs

```
Created classifier and saved to cache.
Accuracy: 0.6625536480686696
                |                           s     |
                |                           c     |
                |                           i     |
                |                           e     |
                |                           n     |
                |           p               c     |
                |           h               e     |
                |           i   r           -     |
                |       m   l   e   r   s   f     |
                |   h   y   o   l   o   c   i   s |
                |   o   s   s   i   m   i   c   p |
                |   r   t   o   g   a   e   t   o |
                |   r   e   p   i   n   n   i   r |
                |   o   r   h   o   c   c   o   t |
                |   r   y   y   n   e   e   n   s |
----------------+---------------------------------+
         horror |<107> 42   .   4  47   7  17  10 |
        mystery |  17<179>  1   1  31   3   2   6 |
     philosophy |   1   .<120> 42   1  59   4   1 |
      religion |   3   4  32<151>  6  20   8   6 |
        romance |   2   3   .   4<157>  .  10  52 |
        science |   1   1  16   9   .<197>  4   2 |
science-fiction |  26  12   2   6  31  32<118> 13 |
         sports |   .   .   .   1  23   4   .<206>|
----------------+---------------------------------+
(row = reference; col = test)



                 |  Precision  |     Recall  | F1-Measure
-------------------------------------------------------------
     philosophy  |     0.7018  |     0.5263  |     0.6015
         sports  |     0.6959  |     0.8803  |     0.7774
        mystery  |     0.7427  |     0.7458  |     0.7443
       religion  |     0.6927  |     0.6565  |     0.6741
        science  |     0.6118  |     0.8565  |     0.7138
        romance  |     0.5304  |     0.6886  |     0.5992
         horror  |     0.6815  |     0.4573  |     0.5473
science-fiction  |     0.7239  |     0.4917  |     0.5856
```