CENG 463
Assignment 1

Cem Gündoğdu
cem.gundogdu@metu.edu.tr

December 15, 2021

Contents

1 Naive Bayes Classifier

1.1 Simplest version

No case conversion, no punctuation removal, no stopword removal, no stemming. Used the default tokenizer from NLTK.

1.1.1 Confusion Matrix

```
Accuracy: 0.6538049303322615
                |                           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 | <43> 22   .   4  29   5   8   7 |
        mystery |   7 <93>  .   .  13   .   1   6 |
     philosophy |   .   2 <45> 21   .  45   1   . |
       religion |   .   4  13 <76>  5  11   2   4 |
        romance |   2   1   .   3 <88>  1   3  16 |
        science |   .   .   1   3   .<106>  4   1 |
science-fiction |  10   8   .   3  26  12 <55>  6 |
         sports |   .   .   1   .  10   2   .<104>|
----------------+---------------------------------+
(row = reference; col = test)
```

1.1.2 Recall and Precision

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.

	Recall	Precision
horror	$\frac{43}{43+22+4+29+5+8+7} = 0.36440678$	$\frac{43}{43+7+2+10} = 0.693548387$