Machine Learning Project
Assignment 1

University of Groningen

Benjamin Kleppe
S3151522
Machine Learning Project
A. Toral
17-02-2020

1)

```
##### Evaluation...
  Accuracy: 0.540037
  |-----------|-----------|-----------|-----------|
  |category   |precision  |recall     |F-measure  |
  |-----------|-----------|-----------|-----------|
  |BINNENLAND |1.000000   |0.015152   |?          |
  |BUITENLAND |1.000000   |0.250000   |?          |
  |ECONOMIE   |1.000000   |0.282051   |?          |
  |INTERVIEW  |0.337243   |0.991379   |?          |
  |KUNST      |0.285714   |0.080000   |?          |
  |RECENSIE   |0.891892   |0.795181   |?          |
  |SPORT      |1.000000   |0.166667   |?          |
  |-----------|-----------|-----------|-----------|
```

2)

```
##### Splitting datasets...
  |-----------|-----------|-----------|-----------|
  |category   |precision  |recall     |F-measure  |
  |-----------|-----------|-----------|-----------|
  |BINNENLAND |NA         |NA         |NA         |
  |BUITENLAND |1.000000   |0.361702   |0.53125    |
  |ECONOMIE   |1.000000   |0.089286   |0.163934   |
  |INTERVIEW  |0.294872   |0.948454   |0.449878   |
  |KUNST      |0.375000   |0.136364   |0.2        |
  |RECENSIE   |0.889503   |0.851852   |0.87027    |
  |SPORT      |1.000000   |0.225806   |0.368421   |
  |-----------|-----------|-----------|-----------|
```

3)

```
##### Analysis...
Most Informative Features
                Reuter = True           BUITEN : RECENS =    276.6 : 1.0
                   dat = None            SPORT : INTERV =    260.4 : 1.0
                   een = None            SPORT : RECENS =    153.0 : 1.0
                     = True           RECENS : BINNEN =    141.5 : 1.0
                    en = None            SPORT : INTERV =    135.5 : 1.0
                   met = None            SPORT : INTERV =    125.1 : 1.0
                   AFP = True           BUITEN : INTERV =    118.2 : 1.0
                  ISBN = True           RECENS : BUITEN =    111.2 : 1.0
              toernooi = True            SPORT : RECENS =    107.5 : 1.0
                    WK = True            SPORT : RECENS =     96.6 : 1.0
```

4)

```
##### Splitting datasets...
0.5130597014925373
0.5186567164179104
0.5
0.5111940298507462
0.527001862197393
0.5186567164179104
0.5186567164179104
0.5223880597014925
0.5149253731343284
0.5018656716417911
0.51464
```

5)

```
0.8432835820895522
0.8507462686567164
0.8022388059701493
0.8470149253731343
0.8584729981378026
0.8190298507462687
0.8544776119402985
0.835820895522388
0.8264925373134329
0.8451492537313433
0.8382726729481085
```

6)

Lowercase	Porter	Lancaster
0.8227611940298507	0.8432835820895522	0.8376865671641791
0.8152985074626866	0.8097014925373134	0.8526119402985075
0.8246268656716418	0.8339552238805971	0.8022388059701493
0.8432835820895522	0.8302238805970149	0.8302238805970149
0.8379888268156425	0.8361266294227188	0.813780260707635
0.832089552238806	0.8470149253731343	0.8488805970149254
0.8600746268656716	0.8600746268656716	0.8638059701492538
0.8488805970149254	0.8264925373134329	0.8563432835820896
0.8339552238805971	0.8526119402985075	0.8339552238805971
0.8283582089552238	0.8208955223880597	0.8339552238805971
0.8347317185024599	0.8360380360766003	0.8373481753244949

Lowercase lowered the accuracy, Porter was either increasing or decreasing the accuracy a little bit. Lancaster was the only parameter which increased the accuracy consistently.
I used the bag_of_non_stopwords function in order to decrease the amount of feats.

I used the linear kernel because it is the fastest (instead of svc, so this is my first tweek).
Also, I increased C to 10, to get a smaller-margin hyperplane in order to achieve a low testing error.
This resulted in the following results:

```
0.8526119402985075
0.8507462686567164
0.8600746268656716
0.832089552238806
0.8659217877094972
0.8488805970149254
0.8395522388059702
0.8283582089552238
0.8544776119402985
0.8507462686567164
0.8483459101142333
```

7)

```
        Welch Two Sample t-test

data:  old and new
t = -1.5305, df = 15.731, p-value = 0.1458
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.024467334  0.003967334
sample estimates:
mean of x mean of y
  0.83810   0.84835
```

old = data from assignment 5.
new = data from assignment 6.

H0 = mean of old is equal to the mean of new
H1 = mean of old is not equal to the mean of new

With t = -1.5305, degrees of freedom = 15.731 and the p-value(0.1458), we cannot reject H0, since $p > 0.05$.

The two scores are not statistically different.