CSCE 470

Programming Assignment #2

Jason Gilman

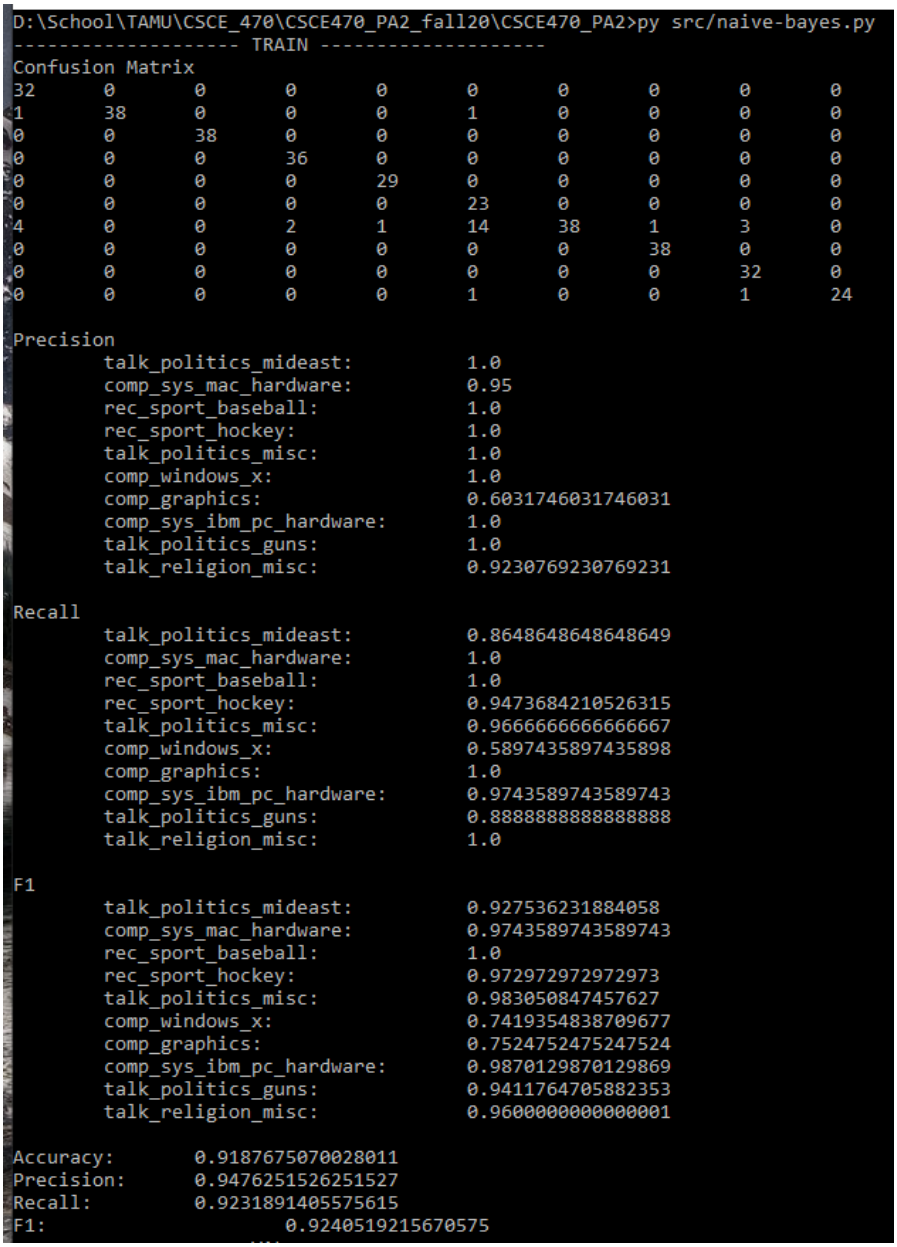
10-18-2020

Introduction

This report will describe the design and results for three text classification algorithms: Naïve-Bayes, KNN, and Rocchio. Finally, explainations for time complexity will be given for each algorithm.

Results

Naïve-Bayes - Full set



```
D:\School\TAMU\CSCE_470\CSCE470_PA2_fall20\CSCE470_PA2>py src/naive-bayes.py
-------------------- TRAIN --------------------
Confusion Matrix
32      0       0       0       0       0       0       0       0       0
1       38      0       0       0       1       0       0       0       0
0       0       38      0       0       0       0       0       0       0
0       0       0       36      0       0       0       0       0       0
0       0       0       0       29      0       0       0       0       0
0       0       0       0       0       23      0       0       0       0
4       0       0       2       1       14      38      1       3       0
0       0       0       0       0       0       0       38      0       0
0       0       0       0       0       0       0       0       32      0
0       0       0       0       0       1       0       0       1       24

Precision
        talk_politics_mideast:          1.0
        comp_sys_mac_hardware:          0.95
        rec_sport_baseball:             1.0
        rec_sport_hockey:               1.0
        talk_politics_misc:             1.0
        comp_windows_x:                 1.0
        comp_graphics:                  0.6031746031746031
        comp_sys_ibm_pc_hardware:       1.0
        talk_politics_guns:             1.0
        talk_religion_misc:             0.9230769230769231

Recall
        talk_politics_mideast:          0.8648648648648649
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             1.0
        rec_sport_hockey:               0.9473684210526315
        talk_politics_misc:             0.9666666666666667
        comp_windows_x:                 0.5897435897435898
        comp_graphics:                  1.0
        comp_sys_ibm_pc_hardware:       0.9743589743589743
        talk_politics_guns:             0.8888888888888888
        talk_religion_misc:             1.0

F1
        talk_politics_mideast:          0.927536231884058
        comp_sys_mac_hardware:          0.9743589743589743
        rec_sport_baseball:             1.0
        rec_sport_hockey:               0.972972972972973
        talk_politics_misc:             0.983050847457627
        comp_windows_x:                 0.7419354838709677
        comp_graphics:                  0.7524752475247524
        comp_sys_ibm_pc_hardware:       0.9870129870129869
        talk_politics_guns:             0.9411764705882353
        talk_religion_misc:             0.9600000000000001

Accuracy:       0.9187675070028011
Precision:      0.9476251526251527
Recall:         0.9231891405575615
F1:                     0.9240519215670575
```

Naïve-Bayes – Full Set

```
-------------------- VAL --------------------
Confusion Matrix
4       0       0       0       0       0       0       0       0       0
1       8       0       1       0       0       0       3       0       0
0       0       5       0       0       0       0       0       0       0
0       0       0       5       0       0       0       0       0       0
0       0       0       0       2       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0
3       2       3       2       3       10      10      1       3       1
0       0       0       0       0       0       0       6       0       0
0       0       0       0       0       0       0       0       0       0
2       0       2       2       3       0       0       0       6       6

Precision
        talk_politics_mideast:          1.0
        comp_sys_mac_hardware:          0.6153846153846154
        rec_sport_baseball:             1.0
        rec_sport_hockey:               1.0
        talk_politics_misc:             1.0
        comp_windows_x:                 0.0
        comp_graphics:                  0.2631578947368421
        comp_sys_ibm_pc_hardware:       1.0
        talk_politics_guns:             0.0
        talk_religion_misc:             0.2857142857142857

Recall
        talk_politics_mideast:          0.4
        comp_sys_mac_hardware:          0.8
        rec_sport_baseball:             0.5
        rec_sport_hockey:               0.5
        talk_politics_misc:             0.25
        comp_windows_x:                 0.0
        comp_graphics:                  1.0
        comp_sys_ibm_pc_hardware:       0.6
        talk_politics_guns:             0.0
        talk_religion_misc:             0.8571428571428571

F1
        talk_politics_mideast:          0.5714285714285715
        comp_sys_mac_hardware:          0.6956521739130435
        rec_sport_baseball:             0.6666666666666666
        rec_sport_hockey:               0.6666666666666666
        talk_politics_misc:             0.4
        comp_windows_x:                 0
        comp_graphics:                  0.4166666666666667
        comp_sys_ibm_pc_hardware:       0.7499999999999999
        talk_politics_guns:             0
        talk_religion_misc:             0.42857142857142855

Accuracy:       0.48936170212765956
Precision:      0.6164256795835743
Recall:         0.49071428571428566
F1:                     0.45956521739130435
```

Naïve-Bayes – Half Set

```
D:\School\TAMU\CSCE_470\CSCE470_PA2_fall20\CSCE470_PA2>py src/naive-bayes.py train_half
-------------------- TRAIN --------------------
Confusion Matrix
11      0       0       0       0       0       0       0       0       0
0       19      0       0       0       0       0       0       0       0
0       0       18      0       0       0       0       0       0       0
0       0       0       15      0       0       0       0       0       0
0       0       0       0       14      0       0       0       0       0
0       0       0       0       0       2       0       0       0       0
7       0       1       4       1       17      19      1       9       0
0       0       0       0       0       0       0       18      0       0
0       0       0       0       0       0       0       0       9       0
0       0       0       0       0       0       0       0       0       12

Precision
        talk_politics_mideast:          1.0
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             1.0
        rec_sport_hockey:               1.0
        talk_politics_misc:             1.0
        comp_windows_x:                 1.0
        comp_graphics:                  0.3220338983050847
        comp_sys_ibm_pc_hardware:       1.0
        talk_politics_guns:             1.0
        talk_religion_misc:             1.0

Recall
        talk_politics_mideast:          0.6111111111111112
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             0.9473684210526315
        rec_sport_hockey:               0.7894736842105263
        talk_politics_misc:             0.9333333333333333
        comp_windows_x:                 0.10526315789473684
        comp_graphics:                  1.0
        comp_sys_ibm_pc_hardware:       0.9473684210526315
        talk_politics_guns:             0.5
        talk_religion_misc:             1.0

F1
        talk_politics_mideast:          0.7586206896551725
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             0.972972972972973
        rec_sport_hockey:               0.8823529411764706
        talk_politics_misc:             0.9655172413793104
        comp_windows_x:                 0.1904761904761905
        comp_graphics:                  0.4871794871794871
        comp_sys_ibm_pc_hardware:       0.972972972972973
        talk_politics_guns:             0.6666666666666666
        talk_religion_misc:             1.0

Accuracy:       0.7740112994350282
Precision:      0.9322033898305084
Recall:         0.7833918128654972
F1:                     0.7896759162479243
```

Naïve-Bayes – Half Set

```
-------------------- VAL --------------------
Confusion Matrix
0       0       0       0       0       0       0       0       0       0
0       3       0       0       0       0       0       0       0       0
0       0       1       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0
10      7       9       10      8       10      10      8       9       5
0       0       0       0       0       0       0       2       0       0
0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       2

Precision
        talk_politics_mideast:          0.0
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             1.0
        rec_sport_hockey:               0.0
        talk_politics_misc:             0.0
        comp_windows_x:                 0.0
        comp_graphics:                  0.11627906976744186
        comp_sys_ibm_pc_hardware:       1.0
        talk_politics_guns:             0.0
        talk_religion_misc:             1.0

Recall
        talk_politics_mideast:          0.0
        comp_sys_mac_hardware:          0.3
        rec_sport_baseball:             0.1
        rec_sport_hockey:               0.0
        talk_politics_misc:             0.0
        comp_windows_x:                 0.0
        comp_graphics:                  1.0
        comp_sys_ibm_pc_hardware:       0.2
        talk_politics_guns:             0.0
        talk_religion_misc:             0.2857142857142857

F1
        talk_politics_mideast:          0
        comp_sys_mac_hardware:          0.4615384615384615
        rec_sport_baseball:             0.18181818181818182
        rec_sport_hockey:               0
        talk_politics_misc:             0
        comp_windows_x:                 0
        comp_graphics:                  0.20833333333333334
        comp_sys_ibm_pc_hardware:       0.33333333333333337
        talk_politics_guns:             0
        talk_religion_misc:             0.4444444444444445

Accuracy:       0.19148936170212766
Precision:      0.41162790697674423
Recall:         0.18857142857142856
F1:                     0.16294677544677544
```

KNN – Full Set

```
D:\School\TAMU\CSCE_470\CSCE470_PA2_fall20\CSCE470_PA2>py src/knn.py
-------------------- TRAIN --------------------
Confusion Matrix
32      2       5       6       6       4       6       3       4       5
0       31      0       1       0       0       0       0       0       0
1       0       28      2       0       0       0       0       0       0
0       0       0       24      0       0       1       0       0       0
1       0       1       2       17      0       0       1       2       3
0       0       0       0       0       34      1       0       0       0
0       0       0       0       0       0       24      0       0       0
0       1       0       0       0       1       5       35      0       0
3       4       4       3       7       0       1       0       30      3
0       0       0       0       0       0       0       0       0       13

Precision
        talk_politics_mideast:          0.4383561643835616
        comp_sys_mac_hardware:          0.96875
        rec_sport_baseball:             0.9032258064516129
        rec_sport_hockey:               0.96
        talk_politics_misc:             0.6296296296296297
        comp_windows_x:                 0.9714285714285714
        comp_graphics:                  1.0
        comp_sys_ibm_pc_hardware:       0.8333333333333334
        talk_politics_guns:             0.5454545454545454
        talk_religion_misc:             1.0

Recall
        talk_politics_mideast:          0.8648648648648649
        comp_sys_mac_hardware:          0.8157894736842105
        rec_sport_baseball:             0.7368421052631579
        rec_sport_hockey:               0.631578947368421
        talk_politics_misc:             0.5666666666666667
        comp_windows_x:                 0.8717948717948718
        comp_graphics:                  0.631578947368421
        comp_sys_ibm_pc_hardware:       0.8974358974358975
        talk_politics_guns:             0.8333333333333334
        talk_religion_misc:             0.5416666666666666

F1
        talk_politics_mideast:          0.5818181818181818
        comp_sys_mac_hardware:          0.8857142857142857
        rec_sport_baseball:             0.8115942028985507
        rec_sport_hockey:               0.7619047619047619
        talk_politics_misc:             0.5964912280701755
        comp_windows_x:                 0.9189189189189189
        comp_graphics:                  0.7741935483870968
        comp_sys_ibm_pc_hardware:       0.8641975308641975
        talk_politics_guns:             0.6593406593406592
        talk_religion_misc:             0.7027027027027027

Accuracy:       0.7507002801120448
Precision:      0.8250178050681255
Recall:         0.7391551774446512
F1:                     0.7556876020619531
```

KNN – Full Set

```
-------------------- VAL --------------------
Confusion Matrix
9       2       2       2       4       1       2       2       2       3
0       2       0       0       0       0       0       0       0       0
0       0       3       3       0       0       0       0       0       0
0       0       0       3       0       0       0       0       0       0
0       0       1       1       1       1       0       1       1       1
0       0       0       0       0       4       1       0       0       0
0       0       0       0       0       0       3       1       0       0
0       5       1       0       0       3       3       3       1       0
1       1       2       1       3       1       1       3       4       3
0       0       1       0       0       0       0       0       1       0

Precision
        talk_politics_mideast:          0.3103448275862069
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             0.5
        rec_sport_hockey:               1.0
        talk_politics_misc:             0.14285714285714285
        comp_windows_x:                 0.8
        comp_graphics:                  0.75
        comp_sys_ibm_pc_hardware:       0.1875
        talk_politics_guns:             0.2
        talk_religion_misc:             0.0

Recall
        talk_politics_mideast:          0.9
        comp_sys_mac_hardware:          0.2
        rec_sport_baseball:             0.3
        rec_sport_hockey:               0.3
        talk_politics_misc:             0.125
        comp_windows_x:                 0.4
        comp_graphics:                  0.3
        comp_sys_ibm_pc_hardware:       0.3
        talk_politics_guns:             0.4444444444444444
        talk_religion_misc:             0.0

F1
        talk_politics_mideast:          0.4615384615384615
        comp_sys_mac_hardware:          0.33333333333333337
        rec_sport_baseball:             0.37499999999999994
        rec_sport_hockey:               0.4615384615384615
        talk_politics_misc:             0.13333333333333333
        comp_windows_x:                 0.5333333333333333
        comp_graphics:                  0.4285714285714285
        comp_sys_ibm_pc_hardware:       0.23076923076923075
        talk_politics_guns:             0.2758620689655173
        talk_religion_misc:             0

Accuracy:       0.3404255319148936
Precision:      0.48907019704433496
Recall:         0.32694444444444437
F1:                     0.32332796513830997
```

KNN – Half Set

```
D:\School\TAMU\CSCE_470\CSCE470_PA2_fall20\CSCE470_PA2>py src/knn.py train_half
-------------------- TRAIN --------------------
Confusion Matrix
17      3       4       4       3       2       2       2       2       4
0       11      0       0       0       0       0       0       0       0
0       0       11      0       0       0       0       0       0       0
0       0       0       12      0       0       1       0       0       0
0       0       0       0       4       0       0       0       0       0
0       0       0       0       0       17      0       0       0       0
0       0       0       0       0       0       11      0       0       0
0       3       0       0       0       0       2       16      0       0
1       2       4       3       8       0       3       1       16      6
0       0       0       0       0       0       0       0       0       2

Precision
        talk_politics_mideast:          0.3953488372093023
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             1.0
        rec_sport_hockey:               0.9230769230769231
        talk_politics_misc:             1.0
        comp_windows_x:                 1.0
        comp_graphics:                  1.0
        comp_sys_ibm_pc_hardware:       0.7619047619047619
        talk_politics_guns:             0.36363636363636365
        talk_religion_misc:             1.0

Recall
        talk_politics_mideast:          0.9444444444444444
        comp_sys_mac_hardware:          0.5789473684210527
        rec_sport_baseball:             0.5789473684210527
        rec_sport_hockey:               0.631578947368421
        talk_politics_misc:             0.26666666666666666
        comp_windows_x:                 0.8947368421052632
        comp_graphics:                  0.5789473684210527
        comp_sys_ibm_pc_hardware:       0.8421052631578947
        talk_politics_guns:             0.8888888888888888
        talk_religion_misc:             0.16666666666666666

F1
        talk_politics_mideast:          0.5573770491803278
        comp_sys_mac_hardware:          0.7333333333333334
        rec_sport_baseball:             0.7333333333333334
        rec_sport_hockey:               0.7499999999999999
        talk_politics_misc:             0.4210526315789474
        comp_windows_x:                 0.9444444444444444
        comp_graphics:                  0.7333333333333334
        comp_sys_ibm_pc_hardware:       0.8
        talk_politics_guns:             0.5161290322580644
        talk_religion_misc:             0.2857142857142857

Accuracy:       0.6610169491525424
Precision:      0.8443966885827351
Recall:         0.6371929824561404
F1:                     0.6474717443176069
```

KNN – Half Set

```
------------------- VAL --------------------
Confusion Matrix
9       2       3       0       3       2       2       4       2       5
0       1       0       0       0       0       1       0       0       0
0       0       1       2       0       0       0       0       0       0
0       0       0       5       0       0       0       0       0       0
0       0       1       0       0       0       0       0       0       1
0       0       0       0       0       4       2       1       0       0
0       0       1       0       0       0       0       1       0       0
0       5       1       0       0       3       2       3       0       0
1       2       3       3       5       1       3       1       7       1
0       0       0       0       0       0       0       0       0       0

Precision
        talk_politics_mideast:          0.28125
        comp_sys_mac_hardware:          0.5
        rec_sport_baseball:             0.3333333333333333
        rec_sport_hockey:               1.0
        talk_politics_misc:             0.0
        comp_windows_x:                 0.5714285714285714
        comp_graphics:                  0.0
        comp_sys_ibm_pc_hardware:       0.21428571428571427
        talk_politics_guns:             0.25925925925925924
        talk_religion_misc:             0.0

Recall
        talk_politics_mideast:          0.9
        comp_sys_mac_hardware:          0.1
        rec_sport_baseball:             0.1
        rec_sport_hockey:               0.5
        talk_politics_misc:             0.0
        comp_windows_x:                 0.4
        comp_graphics:                  0.0
        comp_sys_ibm_pc_hardware:       0.3
        talk_politics_guns:             0.7777777777777778
        talk_religion_misc:             0.0

F1
        talk_politics_mideast:          0.4285714285714286
        comp_sys_mac_hardware:          0.16666666666666669
        rec_sport_baseball:             0.15384615384615383
        rec_sport_hockey:               0.6666666666666666
        talk_politics_misc:             0
        comp_windows_x:                 0.47058823529411764
        comp_graphics:                  0
        comp_sys_ibm_pc_hardware:       0.25
        talk_politics_guns:             0.3888888888888889
        talk_religion_misc:             0

Accuracy:       0.3191489361702128
Precision:      0.3159556878306878
Recall:         0.30777777777777776
F1:                     0.2525228039933922
```

Rocchio – Full Set

```
D:\School\TAMU\CSCE_470\CSCE470_PA2_fall20\CSCE470_PA2>py src/rocchio.py
-------------------- TRAIN --------------------
Confusion Matrix
34      0       1       1       1       2       1       1       0       0
1       35      0       1       0       0       3       3       0       0
1       0       34      0       0       0       0       0       0       0
1       0       0       35      0       0       0       0       0       0
0       1       0       0       25      0       2       1       0       1
0       0       0       0       0       35      0       0       0       0
0       0       0       1       0       0       30      1       0       0
0       0       0       0       0       1       1       31      0       0
0       2       1       0       3       1       1       2       35      3
0       0       2       0       1       0       0       0       1       20

Precision
        talk_politics_mideast:          0.8292682926829268
        comp_sys_mac_hardware:          0.813953488372093
        rec_sport_baseball:             0.9714285714285714
        rec_sport_hockey:               0.9722222222222222
        talk_politics_misc:             0.8333333333333334
        comp_windows_x:                 1.0
        comp_graphics:                  0.9375
        comp_sys_ibm_pc_hardware:       0.9393939393939394
        talk_politics_guns:             0.7291666666666666
        talk_religion_misc:             0.8333333333333334

Recall
        talk_politics_mideast:          0.918918918918919
        comp_sys_mac_hardware:          0.9210526315789473
        rec_sport_baseball:             0.8947368421052632
        rec_sport_hockey:               0.9210526315789473
        talk_politics_misc:             0.8333333333333334
        comp_windows_x:                 0.8974358974358975
        comp_graphics:                  0.7894736842105263
        comp_sys_ibm_pc_hardware:       0.7948717948717948
        talk_politics_guns:             0.9722222222222222
        talk_religion_misc:             0.8333333333333334

F1
        talk_politics_mideast:          0.8717948717948718
        comp_sys_mac_hardware:          0.8641975308641974
        rec_sport_baseball:             0.9315068493150684
        rec_sport_hockey:               0.9459459459459458
        talk_politics_misc:             0.8333333333333334
        comp_windows_x:                 0.945945945945946
        comp_graphics:                  0.8571428571428572
        comp_sys_ibm_pc_hardware:       0.8611111111111112
        talk_politics_guns:             0.8333333333333333
        talk_religion_misc:             0.8333333333333334

Accuracy:       0.8795518207282913
Precision:      0.8859599847433086
Recall:         0.8776431289589185
F1:                     0.8777645112119996
```

Rocchio – Full Set

```
-------------------- VAL --------------------
Confusion Matrix
8       0       0       0       1       0       1       0       0       1
0       6       1       0       0       0       0       0       1       0
2       0       6       0       0       0       0       0       1       0
0       1       1       8       0       0       0       0       0       0
0       0       0       0       3       1       0       0       0       1
0       0       0       0       0       6       1       1       0       0
0       1       2       1       0       0       6       1       1       0
0       0       0       0       0       1       1       6       0       0
0       2       0       0       2       2       1       2       4       3
0       0       0       1       2       0       0       0       2       2

Precision
        talk_politics_mideast:          0.7272727272727273
        comp_sys_mac_hardware:          0.75
        rec_sport_baseball:             0.6666666666666666
        rec_sport_hockey:               0.8
        talk_politics_misc:             0.6
        comp_windows_x:                 0.75
        comp_graphics:                  0.5
        comp_sys_ibm_pc_hardware:       0.75
        talk_politics_guns:             0.25
        talk_religion_misc:             0.2857142857142857

Recall
        talk_politics_mideast:          0.8
        comp_sys_mac_hardware:          0.6
        rec_sport_baseball:             0.6
        rec_sport_hockey:               0.8
        talk_politics_misc:             0.375
        comp_windows_x:                 0.6
        comp_graphics:                  0.6
        comp_sys_ibm_pc_hardware:       0.6
        talk_politics_guns:             0.4444444444444444
        talk_religion_misc:             0.2857142857142857

F1
        talk_politics_mideast:          0.761904761904762
        comp_sys_mac_hardware:          0.6666666666666665
        rec_sport_baseball:             0.631578947368421
        rec_sport_hockey:               0.8000000000000002
        talk_politics_misc:             0.4615384615384615
        comp_windows_x:                 0.6666666666666665
        comp_graphics:                  0.5454545454545454
        comp_sys_ibm_pc_hardware:       0.6666666666666665
        talk_politics_guns:             0.32
        talk_religion_misc:             0.2857142857142857

Accuracy:       0.5851063829787234
Precision:      0.607965367965368
Recall:         0.570515873015873
F1:                     0.5806191001980475
```

Rocchio – Half Set

```
D:\School\TAMU\CSCE_470\CSCE470_PA2_fall20\CSCE470_PA2>py src/rocchio.py train_half
-------------------- TRAIN --------------------
Confusion Matrix
18      0       0       1       0       0       0       1       0       0
0       17      0       0       0       0       0       0       0       0
0       0       18      0       0       0       0       0       0       0
0       0       0       17      0       0       0       0       0       0
0       0       0       0       15      0       0       1       0       0
0       0       0       0       0       18      0       0       0       0
0       0       0       0       0       0       18      0       0       0
0       1       0       1       0       0       1       16      0       0
0       1       1       0       0       1       0       1       18      1
0       0       0       0       0       0       0       0       0       11

Precision
        talk_politics_mideast:          0.9
        comp_sys_mac_hardware:          1.0
        rec_sport_baseball:             1.0
        rec_sport_hockey:               1.0
        talk_politics_misc:             0.9375
        comp_windows_x:                 1.0
        comp_graphics:                  1.0
        comp_sys_ibm_pc_hardware:       0.8421052631578947
        talk_politics_guns:             0.782608695652174
        talk_religion_misc:             1.0

Recall
        talk_politics_mideast:          1.0
        comp_sys_mac_hardware:          0.8947368421052632
        rec_sport_baseball:             0.9473684210526315
        rec_sport_hockey:               0.8947368421052632
        talk_politics_misc:             1.0
        comp_windows_x:                 0.9473684210526315
        comp_graphics:                  0.9473684210526315
        comp_sys_ibm_pc_hardware:       0.8421052631578947
        talk_politics_guns:             1.0
        talk_religion_misc:             0.9166666666666666

F1
        talk_politics_mideast:          0.9473684210526316
        comp_sys_mac_hardware:          0.9444444444444444
        rec_sport_baseball:             0.972972972972973
        rec_sport_hockey:               0.9444444444444444
        talk_politics_misc:             0.967741935483871
        comp_windows_x:                 0.972972972972973
        comp_graphics:                  0.972972972972973
        comp_sys_ibm_pc_hardware:       0.8421052631578947
        talk_politics_guns:             0.878048780487805
        talk_religion_misc:             0.9565217391304348

Accuracy:       0.9378531073446328
Precision:      0.946221395881007
Recall:         0.9390350877192981
F1:                     0.9399593947120446
```

Rocchio – Half Set

```
-------------------- VAL --------------------
Confusion Matrix
2       1       2       0       1       1       2       2       0       1
0       4       2       1       0       0       0       2       0       0
2       0       2       1       0       0       0       0       1       1
0       0       0       5       0       0       0       0       0       0
1       0       0       0       3       0       0       0       0       1
0       0       0       0       0       5       0       1       0       0
0       0       0       0       0       0       4       0       1       0
0       3       2       0       0       2       3       5       1       0
5       2       2       2       2       2       0       0       5       3
0       0       0       1       2       0       1       0       1       1

Precision
        talk_politics_mideast:          0.16666666666666666
        comp_sys_mac_hardware:          0.4444444444444444
        rec_sport_baseball:             0.2857142857142857
        rec_sport_hockey:               1.0
        talk_politics_misc:             0.6
        comp_windows_x:                 0.8333333333333334
        comp_graphics:                  0.8
        comp_sys_ibm_pc_hardware:       0.3125
        talk_politics_guns:             0.21739130434782608
        talk_religion_misc:             0.16666666666666666

Recall
        talk_politics_mideast:          0.2
        comp_sys_mac_hardware:          0.4
        rec_sport_baseball:             0.2
        rec_sport_hockey:               0.5
        talk_politics_misc:             0.375
        comp_windows_x:                 0.5
        comp_graphics:                  0.4
        comp_sys_ibm_pc_hardware:       0.5
        talk_politics_guns:             0.5555555555555556
        talk_religion_misc:             0.14285714285714285

F1
        talk_politics_mideast:          0.1818181818181818
        comp_sys_mac_hardware:          0.4210526315789474
        rec_sport_baseball:             0.23529411764705882
        rec_sport_hockey:               0.6666666666666666
        talk_politics_misc:             0.4615384615384615
        comp_windows_x:                 0.625
        comp_graphics:                  0.5333333333333333
        comp_sys_ibm_pc_hardware:       0.38461538461538464
        talk_politics_guns:             0.3125
        talk_religion_misc:             0.15384615384615383

Accuracy:       0.3829787234042553
Precision:      0.48267167011732237
Recall:         0.3773412698412698
F1:                     0.39756649310441877
```

Variance-Bias

Naïve-Bayes

This model was designed so that it considered only a few featueres, which provided a highly biased solution. Only term frequency was considered in the model, while laplace smoothing was used to make the predictions more uniform. To have a higher variance, more features would need to be considered, or increased weight could be given to certain docuemtns. The linearity of the naïve-bayes model also adds to its high bias. The difference in classes is more defined because of the classifiying of the documents. From the output statistics, a high bias is shown by the proportion between accuracy and precsision, where precsion is favored.

KNN

This model was designed using a single parameter, the K value which determined the number of neighboring documents to consider in the final prediction analysis. For my model I chose a high K value, 5. While having a higher K value offers greater bias in predictions, the KNN model still has a high variance due to its non-linearity. The difference in classes due to where each document is classified is a non-linear relationship. The normaliztion process aided to add some variance as is weighted the term frequecies. From the output statistics, a high variance is shown by the proportion between accuracy and precision, where accuracy is favored.

Rocchio

This model has high bias as it predicts based on linearly plotted vectors for each document. The boundary for a class is well defined because of the centriod which is based on a class's documents. When calculating the similarity for the prediction document, you want find the most similar centriod vector. The normaliztion process aided to add some variance as is weighted the term frequecies. From the output statistics, a high bias is shown by the proportion between accuracy and precsision, where precsion is favored.

Time Complexity

	T_START	T_END	T_TOTAL
Naïve-Bayes			
train()	1603078348	1603078355	6.38783002
predict()	1603078355	1603078383	27.88059
KNN			
train()	1603078622	1603078623	0.68500996
predict()	1603078623	1603078632	8.74106002
Rocchio			
train()	1603078723	1603078723	0.71581006
predict()	1603078723	1603078727	3.52181983

The Naïve-Bayes classifier is slower compared to the KNN and Rocchio classifiers when used on this data set. This is because the Naïve-Bayes classifier compares each document with the prediction document. The other two classifiers have less comparision to the prediction document because the KNN computes the cosine similarity, and the Rocchio computes the cosine similarity between the prediction document and the centriod vector for each class. If the data set size was much greater, it is possible that the Naïve-Bayes classifier would be the faster algorithm.