CECS 429/529 — Sec. 01

—

Milestone 3: The Federalist Papers

—

Harold Agnote

Braulio Flores

David Van

December 15, 2017

1 Terms

1.1 The first 30 terms in the vocabulary of the ENTIRE CORPUS, including disputed documents, sorted in alphabetical order.

```
Current Working Directory: /home/harold/Fall_2017_Work/CECS-429-Group-Project/
                   ↳ search_engine/assets/disputed

Select Mode:
1. Build Index
2. Query Index
3. Classify

4. Quit
3
Classifier: bayesian
Input Command: :vocab

1. 1
2. 10
3. 11
4. 128
5. 13
6. 136
7. 13th
8. 14
9. 15
10. 1685
11. 1688
12. 1706
13. 1774
14. 1783
15. 1784
16. 1786
17. 1787
18. 1788
19. 18
20. 1808
21. 181
22. 186
23. 195
24. 1abb
25. 1st
26. 2
27. 20
28. 21
29. 22
30. 23
Vocabulary Size: 4971 Terms
```

2 Bayesian Classifier

2.1 The top 10 discriminating terms and their I (C,T) score.

```
Time taken to build discriminating vocab: 0 seconds. Total number of things in
              ↳ priority_queue: 8682
Discriminating Vocab 1: Term: upon, Score: 0.5664605541299845
Discriminating Vocab 2: Term: although, Score: 0.27577723284878464
Discriminating Vocab 3: Term: wish, Score: 0.2573596818868562
Discriminating Vocab 4: Term: whilst, Score: 0.23596596277142606
Discriminating Vocab 5: Term: few, Score: 0.22592745532481184
Discriminating Vocab 6: Term: lie, Score: 0.19458127684985455
Discriminating Vocab 7: Term: intend, Score: 0.18168304357019316
Discriminating Vocab 8: Term: kind, Score: 0.1752235272028166
Discriminating Vocab 9: Term: readili, Score: 0.17397551301134465
Discriminating Vocab 10: Term: constitut, Score: 0.16569621635200926
```

2.2 For the document, paper_49.txt

2.2.1 The exact Bayesian classification score for each of the three classes on this paper, using 10 discriminating terms

```
Classifier: bayesian
Input Command: :classify paper_49.txt

Using k=10
Time taken to build discriminating vocab: 0 seconds. Total number of things in
          ↳ priority_queue: 8682
Hamilton Classification Score: -21.935259363249624
Madison Classification Score: -15.28486523292957
Jay Classification Score: -17.044615396460532

paper_49.txt was written by Madison
```

2.2.2 The same data, except using 50 discriminating terms

```
Classifier: bayesian
Input Command: :classify paper_49.txt

Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
          ↳ priority_queue: 8682
Hamilton Classification Score: -78.52855021314149
Madison Classification Score: -73.0309789702613
Jay Classification Score: -83.73277456428272

paper_49.txt was written by Madison
```

2.3 The classification decision for all disputed documents using 50 discriminating terms

```
Classifier: bayesian
Input Command: :classify all

Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
            ↳ priority_queue: 8682
Hamilton Classification Score: -75.7591487108412
Madison Classification Score: -65.50637576540075
Jay Classification Score: -81.49708532033281

paper_52.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
            ↳ priority_queue: 8682
Hamilton Classification Score: -46.27237186505575
Madison Classification Score: -46.071721505461
Jay Classification Score: -43.708631116342325

paper_50.txt was written by Jay
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
            ↳ priority_queue: 8682
Hamilton Classification Score: -115.25498938854301
Madison Classification Score: -96.55783654739776
Jay Classification Score: -113.39047436411155

paper_63.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
            ↳ priority_queue: 8682
Hamilton Classification Score: -91.60316030247
Madison Classification Score: -87.22298400862651
Jay Classification Score: -96.26914321383654

paper_51.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
            ↳ priority_queue: 8682
Hamilton Classification Score: -83.12253250311922
Madison Classification Score: -72.7567012789703
Jay Classification Score: -85.90269956572502

paper_57.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
            ↳ priority_queue: 8682
Hamilton Classification Score: -85.19793372709266
Madison Classification Score: -76.04618753047892
Jay Classification Score: -86.69360396700205

paper_62.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
```

```
              ↳ priority_queue: 8682
Hamilton Classification Score: -100.51785062970565
Madison Classification Score: -90.91942302928226
Jay Classification Score: -101.49300702238969

paper_56.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
              ↳ priority_queue: 8682
Hamilton Classification Score: -78.52855021314149
Madison Classification Score: -73.03097897026132
Jay Classification Score: -83.73277456428272

paper_49.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
              ↳ priority_queue: 8682
Hamilton Classification Score: -61.487680225517224
Madison Classification Score: -55.44654069638152
Jay Classification Score: -62.59814659139426

paper_55.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
              ↳ priority_queue: 8682
Hamilton Classification Score: -80.82518471659084
Madison Classification Score: -76.10182967887961
Jay Classification Score: -78.54955274022694

paper_54.txt was written by Madison
Using k=50
Time taken to build discriminating vocab: 0 seconds. Total number of things in
              ↳ priority_queue: 8682
Hamilton Classification Score: -107.21407396386248
Madison Classification Score: -96.43444140965786
Jay Classification Score: -103.49565660080361

paper_53.txt was written by Madison
Classifier: bayesian
```

3 Rocchio Classification

3.1 The first 30 components of the normalized centroid vectors for the three classes

```
Classifier: rocchio
Input Command: :centroid_vectors

Hamilton
1. Term: a, Score: 0.03411514665108945
2. Term: abil, Score: 0.02902122350801603
3. Term: abl, Score: 0.05363112785583057
4. Term: abridg, Score: 0.03974648725240956
5. Term: absorb, Score: 0.03842986294007918
6. Term: abund, Score: 0.03137119468110281
7. Term: acquaint, Score: 0.02617062737183156
8. Term: acquir, Score: 0.027074493222453224
9. Term: activ, Score: 0.030280900880415674
10. Term: address, Score: 0.02810510103668744
11. Term: administ, Score: 0.029297683085508897
12. Term: administr, Score: 0.028849622184405153
13. Term: admit, Score: 0.03047874340434609
14. Term: advantag, Score: 0.029336859809063393
15. Term: affair, Score: 0.027311950415478802
16. Term: affect, Score: 0.032395182234266205
17. Term: affirm, Score: 0.032406683308601865
18. Term: afford, Score: 0.031802287225572996
19. Term: against, Score: 0.03139978113934042
20. Term: agricultur, Score: 0.031588921509724401
21. Term: all, Score: 0.03276036527695492
22. Term: allot, Score: 0.03874152404281772
23. Term: allow, Score: 0.03728883888724523
24. Term: allur, Score: 0.033072678408899554
25. Term: almost, Score: 0.036867231204956025
26. Term: alon, Score: 0.03629319373696218
27. Term: alreadi, Score: 0.029291199975202836
28. Term: alway, Score: 0.03700319078384223
29. Term: am, Score: 0.03996934288245812
30. Term: ambit, Score: 0.04027446238800081

Jay
1. Term: 1788, Score: 0.0338119330781322
2. Term: 7, Score: 0.03351520790727475
3. Term: a, Score: 0.050580605109160126
4. Term: abil, Score: 0.023715290531242165
5. Term: abl, Score: 0.030475693896191806
6. Term: abound, Score: 0.019992971069720762
7. Term: absolut, Score: 0.0230197583073803
8. Term: account, Score: 0.019992971069720762
9. Term: accumul, Score: 0.019992971069720762
10. Term: accur, Score: 0.023647195396430468
11. Term: achiev, Score: 0.019992971069720762
12. Term: acknowledg, Score: 0.019992971069720762
13. Term: acquaint, Score: 0.019992971069720762
14. Term: act, Score: 0.03371683933894803
15. Term: activ, Score: 0.03209066970767617
```

```
16. Term: actuat, Score: 0.02344771157182362
17. Term: admit, Score: 0.02344771157182362
18. Term: advanc, Score: 0.019992971069720762
19. Term: advantag, Score: 0.03349118446808264
20. Term: advic, Score: 0.03372068998673097
21. Term: affair, Score: 0.043768509366003995
22. Term: affect, Score: 0.03209876399249191
23. Term: afford, Score: 0.034304443327651805
24. Term: afterward, Score: 0.02690245207392648
25. Term: aggreg, Score: 0.019992971069720762
26. Term: all, Score: 0.04847948007407419
27. Term: also, Score: 0.0230197583073803
28. Term: alter, Score: 0.036256205757481
29. Term: although, Score: 0.054081921758680415
30. Term: alway, Score: 0.0362279437230157

Madison
1. Term: 1, Score: 0.030797233545304503
2. Term: 2, Score: 0.021385639187361718
3. Term: 3, Score: 0.05511376085628728
4. Term: 4, Score: 0.03678910186635688
5. Term: 5, Score: 0.02357542794001587
6. Term: 6, Score: 0.02684254660615941
7. Term: a, Score: 0.035829708495521174
8. Term: abl, Score: 0.025268932738535917
9. Term: abroad, Score: 0.030222153935469072
10. Term: absurd, Score: 0.0265788339735214
11. Term: abus, Score: 0.024895400604871387
12. Term: accommod, Score: 0.026728680569450044
13. Term: act, Score: 0.022343924529005615
14. Term: ad, Score: 0.02752346080545629
15. Term: addit, Score: 0.02835945318749984
16. Term: address, Score: 0.024290113568947107
17. Term: adequ, Score: 0.029924870133227324
18. Term: admiss, Score: 0.024636498702884143
19. Term: admit, Score: 0.023360527221066053
20. Term: advanc, Score: 0.03287009665150468
21. Term: advantag, Score: 0.027630844910368213
22. Term: adventur, Score: 0.034043891341806395
23. Term: affair, Score: 0.032416767494722595
24. Term: affirm, Score: 0.030913351324053394
25. Term: again, Score: 0.023432913862012964
26. Term: against, Score: 0.03079864295813389
27. Term: aggreg, Score: 0.0307968979332397
28. Term: ago, Score: 0.02441942423237494
29. Term: agricultur, Score: 0.027692270906955408
30. Term: all, Score: 0.040946581248306364
```

3.2 For the Document, paper_52.txt

3.2.1 The first 30 components of the normalized vector for the document — using terms in disputed index only

```
Classifier: rocchio
Input Command: :classify paper_52.txt

1. Term: 1788, Score: 0.024060341553311018
2. Term: 8, Score: 0.024060341553311018
3. Term: a, Score: 0.10960323025364851
4. Term: abil, Score: 0.024060341553311018
5. Term: abl, Score: 0.024060341553311018
6. Term: about, Score: 0.024060341553311018
7. Term: abov, Score: 0.024060341553311018
8. Term: abridg, Score: 0.024060341553311018
9. Term: absolut, Score: 0.024060341553311018
10. Term: abus, Score: 0.024060341553311018
11. Term: access, Score: 0.040737699464297845
12. Term: accident, Score: 0.024060341553311018
13. Term: act, Score: 0.024060341553311018
14. Term: addit, Score: 0.024060341553311018
15. Term: adopt, Score: 0.024060341553311018
16. Term: advantag, Score: 0.040737699464297845
17. Term: affect, Score: 0.024060341553311018
18. Term: after, Score: 0.024060341553311018
19. Term: age, Score: 0.024060341553311018
20. Term: alarm, Score: 0.024060341553311018
21. Term: all, Score: 0.05049332845333046
22. Term: allud, Score: 0.024060341553311018
23. Term: alon, Score: 0.024060341553311018
24. Term: alreadi, Score: 0.024060341553311018
25. Term: also, Score: 0.040737699464297845
26. Term: alter, Score: 0.040737699464297845
27. Term: alway, Score: 0.024060341553311018
28. Term: ambit, Score: 0.024060341553311018
29. Term: among, Score: 0.040737699464297845
30. Term: an, Score: 0.05049332845333046
```

3.2.2 The first 30 components of the normalized vector for the document — using terms in entire corpus

```
Classifier: rocchio
Input Command: :classify paper_52.txt

1. Term: 10, Score: 0
2. Term: 11, Score: 0
3. Term: 128, Score: 0
4. Term: 13, Score: 0
5. Term: 136, Score: 0
6. Term: 13th, Score: 0
7. Term: 14, Score: 0
8. Term: 1685, Score: 0
9. Term: 1688, Score: 0
10. Term: 1706, Score: 0
11. Term: 1774, Score: 0
12. Term: 1786, Score: 0
```

```
13. Term: 1787, Score: 0
14. Term: 1788, Score: 0.024060341553311014
15. Term: 18, Score: 0
16. Term: 1808, Score: 0
17. Term: 181, Score: 0
18. Term: 186, Score: 0
19. Term: 195, Score: 0
20. Term: 1abb, Score: 0
21. Term: 1st, Score: 0
22. Term: 2, Score: 0
23. Term: 20, Score: 0
24. Term: 21, Score: 0
25. Term: 22, Score: 0
26. Term: 23, Score: 0
27. Term: 25, Score: 0
28. Term: 257, Score: 0
29. Term: 26, Score: 0
30. Term: 262, Score: 0
```

3.2.3 The exact Euclidians distance between the normalized vector for the document and each of the 3 class centroids

```
Classifier: rocchio
Input Command: :classify paper_52.txt

Hamilton Euclidian Distance: 0.5690125222964909

Jay Euclidian Distance: 0.8639851005362514

Madison Euclidian Distance: 0.6647660668790742

paper_52.txt was written by Hamilton
```

3.3 The classification decision for all disputed documents

```
Classifier: rocchio
Input Command: :classify all

paper_62.txt was written by Hamilton
paper_49.txt was written by Hamilton
paper_54.txt was written by Hamilton
paper_56.txt was written by Hamilton
paper_57.txt was written by Hamilton
paper_55.txt was written by Hamilton
paper_50.txt was written by Hamilton
paper_52.txt was written by Hamilton
paper_63.txt was written by Hamilton
paper_53.txt was written by Hamilton
paper_51.txt was written by Hamilton
```