Homework 5

Jacob Sachs

20 May 2013

Approach

For this project, I chose to use the Maximum Likelihood Estimator (MLE) method. This meant picking an appropriate value of γ for psuedocounts. In order to accomplish this, I performed cross-validation on the training set, as shown in the table below. Interestingly, I found that most values of γ gave excellent accuracy. I'm somewhat dubious of this accuracy, since it is often perfect. Perhaps my algorithm was "too good", and was less naive than it was meant to be, but I believe my implementation is faithful to the specification.

Words were made all lower case. Hyphens were kept in place. Extra punctuation was stripped.

Dictionaries are built up as a preprocessing step. The documents are all stored in a Document class, meaning all information about them is determined before any training or classification is performed.

For extra credit, I decided to truncate the dictionary of the ten most common stopwords, using this wiki page as my source (http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/PG/2006/04/1-10000). The test data is then repeated again.

Tables for Basic Project

- **Values for γ**

For values of γ in the range [0, 30], all values yielded 100% accuracy. For this reason, I have not included a table, and am using $\gamma = 5$ (at first). I wish this could be less arbitrary, but I'm not sure why my cross-validation doesn't yield less accurate results (a strange problem, I guess). However, when running on the unknown documents, this made my code classify all of the unknowns as Madison (which seems unlikely). I tried many other values of gamma, but nothing seemed to change. I am not sure where the error is occurring (unless they really are Madison). At this point, I'm hoping the extra credit version yields better results.

Gamma: 1
Accuracy: 1.0
Gamma: 2
Accuracy: 1.0
Gamma: 3
Accuracy: 1.0
Gamma: 4
Accuracy: 1.0
Gamma: 5
Accuracy: 1.0
Gamma: 6
Accuracy: 1.0
Gamma: 7

Accuracy: 1.0
Gamma: 8
Accuracy: 1.0
Gamma: 9
Accuracy: 1.0
Gamma: 10
Accuracy: 1.0
Gamma: 11
Accuracy: 1.0
Gamma: 12
Accuracy: 1.0
Gamma: 13
Accuracy: 1.0
Gamma: 14
Accuracy: 1.0
Gamma: 15
Accuracy: 1.0
Gamma: 16
Accuracy: 1.0
Gamma: 17
Accuracy: 1.0
Gamma: 18
Accuracy: 1.0
Gamma: 19
Accuracy: 1.0
Gamma: 20
Accuracy: 1.0
Gamma: 21
Accuracy: 1.0
Gamma: 22
Accuracy: 1.0
Gamma: 23
Accuracy: 1.0
Gamma: 24
Accuracy: 1.0
Gamma: 25
Accuracy: 1.0
Gamma: 26
Accuracy: 1.0
Gamma: 27
Accuracy: 1.0
Gamma: 28
Accuracy: 1.0
Gamma: 29
Accuracy: 1.0

Training Documents

Document: hamilton1.txt
Hamilton: -6266.78681131 | Madison: -6709.21444924
Classified: Hamilton
Actual: Hamilton
Document: hamilton10.txt
Hamilton: -4934.64891644 | Madison: -5241.89752622

Classified: Hamilton
Actual: Hamilton
Document: hamilton11.txt
Hamilton: -6256.81049292 | Madison: -6666.46906558
Classified: Hamilton
Actual: Hamilton
Document: hamilton12.txt
Hamilton: -10995.2021699 | Madison: -11653.4402537
Classified: Hamilton
Actual: Hamilton
Document: hamilton13.txt
Hamilton: -5863.32603217 | Madison: -6183.27714259
Classified: Hamilton
Actual: Hamilton
Document: hamilton14.txt
Hamilton: -6879.47147666 | Madison: -7293.42452699
Classified: Hamilton
Actual: Hamilton
Document: hamilton15.txt
Hamilton: -4690.14762775 | Madison: -4956.99977694
Classified: Hamilton
Actual: Hamilton
Document: hamilton2.txt
Hamilton: -7150.3767776 | Madison: -7653.45662444
Classified: Hamilton
Actual: Hamilton
Document: hamilton3.txt
Hamilton: -6375.6401655 | Madison: -6805.78186599
Classified: Hamilton
Actual: Hamilton
Document: hamilton4.txt
Hamilton: -6270.25774498 | Madison: -6633.22294299
Classified: Hamilton
Actual: Hamilton
Document: hamilton5.txt
Hamilton: -7955.06637134 | Madison: -8537.20337811
Classified: Hamilton
Actual: Hamilton
Document: hamilton6.txt
Hamilton: -6851.4898456 | Madison: -7328.44977996
Classified: Hamilton
Actual: Hamilton
Document: hamilton7.txt
Hamilton: -3043.69970481 | Madison: -3252.22980501
Classified: Hamilton
Actual: Hamilton
Document: hamilton8.txt
Hamilton: -9787.35814061 | Madison: -10421.1568231
Classified: Hamilton
Actual: Hamilton
Document: hamilton9.txt
Hamilton: -6340.17427168 | Madison: -6763.0898085
Classified: Hamilton

Actual: Hamilton
Document: madison1.txt
Hamilton: -10173.5326535 | Madison: -9421.71041676
Classified: Madison
Actual: Madison
Document: madison10.txt
Hamilton: -9720.05134023 | Madison: -8889.39699656
Classified: Madison
Actual: Madison
Document: madison11.txt
Hamilton: -6927.40893351 | Madison: -6334.02779455
Classified: Madison
Actual: Madison
Document: madison12.txt
Hamilton: -8607.08053973 | Madison: -7923.31307823
Classified: Madison
Actual: Madison
Document: madison13.txt
Hamilton: -9339.41060897 | Madison: -8413.23394532
Classified: Madison
Actual: Madison
Document: madison14.txt
Hamilton: -6350.93843812 | Madison: -5739.99177336
Classified: Madison
Actual: Madison
Document: madison15.txt
Hamilton: -6958.00614876 | Madison: -6387.96323788
Classified: Madison
Actual: Madison
Document: madison2.txt
Hamilton: -7262.90219916 | Madison: -6718.77450046
Classified: Madison
Actual: Madison
Document: madison3.txt
Hamilton: -9256.96669973 | Madison: -8533.25654134
Classified: Madison
Actual: Madison
Document: madison4.txt
Hamilton: -11348.2478475 | Madison: -10437.6854803
Classified: Madison
Actual: Madison
Document: madison5.txt
Hamilton: -8514.8825254 | Madison: -7767.83492102
Classified: Madison
Actual: Madison
Document: madison6.txt
Hamilton: -10105.5320582 | Madison: -9231.60824953
Classified: Madison
Actual: Madison
Document: madison7.txt
Hamilton: -12090.3098272 | Madison: -11169.4504721
Classified: Madison
Actual: Madison

Document: madison8.txt
Hamilton: -9392.24014968 | Madison: -8647.02460159
Classified: Madison
Actual: Madison
Document: madison9.txt
Hamilton: -11488.8762092 | Madison: -10558.2583842
Classified: Madison
Actual: Madison
Gamma: 5
Accuracy: 1.0

Unknowns

Document: unknown1.txt
Hamilton: -5318.80320195 | Madison: -5195.35926576
Classified: Madison
Document: unknown10.txt
Hamilton: -7815.39794131 | Madison: -7670.21032635
Classified: Madison
Document: unknown11.txt
Hamilton: -9974.28968111 | Madison: -9769.52585399
Classified: Madison
Document: unknown2.txt
Hamilton: -3677.22534436 | Madison: -3598.40895739
Classified: Madison
Document: unknown3.txt
Hamilton: -6175.05384871 | Madison: -6009.16734871
Classified: Madison
Document: unknown4.txt
Hamilton: -6017.12733677 | Madison: -5889.58206703
Classified: Madison
Document: unknown5.txt
Hamilton: -7084.19802569 | Madison: -6955.52919742
Classified: Madison
Document: unknown6.txt
Hamilton: -6398.1744339 | Madison: -6278.41439088
Classified: Madison
Document: unknown7.txt
Hamilton: -6670.87698034 | Madison: -6553.09493354
Classified: Madison
Document: unknown8.txt
Hamilton: -5048.57363795 | Madison: -4942.99283393
Classified: Madison
Document: unknown9.txt
Hamilton: -7159.67336746 | Madison: -7038.89718496
Classified: Madison

Extra Credit

Here, I chose to truncate dictionaries by the ten most common stopwords. I then repeated the above tests.

Values for γ

As expected, γ values stayed the same, so I did not go past 10.

Gamma: 1
Accuracy: 1.0
Gamma: 2
Accuracy: 1.0
Gamma: 3
Accuracy: 1.0
Gamma: 4
Accuracy: 1.0
Gamma: 5
Accuracy: 1.0
Gamma: 6
Accuracy: 1.0
Gamma: 7
Accuracy: 1.0
Gamma: 8
Accuracy: 1.0
Gamma: 9
Accuracy: 1.0

■ Training Documents

Document: hamilton1.txt
Hamilton: -5194.40890089 | Madison: -5587.61898456
Classified: Hamilton
Actual: Hamilton
Document: hamilton10.txt
Hamilton: -4035.44225205 | Madison: -4305.53825111
Classified: Hamilton
Actual: Hamilton
Document: hamilton11.txt
Hamilton: -5241.11386654 | Madison: -5600.63772792
Classified: Hamilton
Actual: Hamilton
Document: hamilton12.txt
Hamilton: -9170.22703525 | Madison: -9742.09677769
Classified: Hamilton
Actual: Hamilton
Document: hamilton13.txt
Hamilton: -4878.26616682 | Madison: -5147.35077165
Classified: Hamilton
Actual: Hamilton
Document: hamilton14.txt
Hamilton: -5645.42613861 | Madison: -6000.6780733
Classified: Hamilton
Actual: Hamilton
Document: hamilton15.txt
Hamilton: -3839.07028015 | Madison: -4064.31637661
Classified: Hamilton
Actual: Hamilton
Document: hamilton2.txt
Hamilton: -6035.30604293 | Madison: -6482.95225288
Classified: Hamilton

Actual: Hamilton
Document: hamilton3.txt
Hamilton: -5424.19526975 | Madison: -5802.52877764
Classified: Hamilton
Actual: Hamilton
Document: hamilton4.txt
Hamilton: -5214.14280578 | Madison: -5525.11029978
Classified: Hamilton
Actual: Hamilton
Document: hamilton5.txt
Hamilton: -6723.53624533 | Madison: -7240.2710901
Classified: Hamilton
Actual: Hamilton
Document: hamilton6.txt
Hamilton: -5759.68567365 | Madison: -6180.93310577
Classified: Hamilton
Actual: Hamilton
Document: hamilton7.txt
Hamilton: -2592.82254344 | Madison: -2775.18220801
Classified: Hamilton
Actual: Hamilton
Document: hamilton8.txt
Hamilton: -8263.77400179 | Madison: -8820.39211676
Classified: Hamilton
Actual: Hamilton
Document: hamilton9.txt
Hamilton: -5323.80079498 | Madison: -5694.69875284
Classified: Hamilton
Actual: Hamilton
Document: madison1.txt
Hamilton: -8543.79535751 | Madison: -7925.09405456
Classified: Madison
Actual: Madison
Document: madison10.txt
Hamilton: -8130.30973653 | Madison: -7431.32535234
Classified: Madison
Actual: Madison
Document: madison11.txt
Hamilton: -5642.89433539 | Madison: -5160.20469299
Classified: Madison
Actual: Madison
Document: madison12.txt
Hamilton: -7089.60565512 | Madison: -6533.35828016
Classified: Madison
Actual: Madison
Document: madison13.txt
Hamilton: -7658.41235912 | Madison: -6872.91777545
Classified: Madison
Actual: Madison
Document: madison14.txt
Hamilton: -5341.82422982 | Madison: -4813.39735753
Classified: Madison
Actual: Madison

Document: madison15.txt
Hamilton: -5821.09847161 | Madison: -5343.01613866
Classified: Madison
Actual: Madison
Document: madison2.txt
Hamilton: -6124.21315831 | Madison: -5672.64738845
Classified: Madison
Actual: Madison
Document: madison3.txt
Hamilton: -7759.33372503 | Madison: -7156.74086462
Classified: Madison
Actual: Madison
Document: madison4.txt
Hamilton: -9499.40404222 | Madison: -8734.96578626
Classified: Madison
Actual: Madison
Document: madison5.txt
Hamilton: -7032.66554462 | Madison: -6405.72788216
Classified: Madison
Actual: Madison
Document: madison6.txt
Hamilton: -8312.67335925 | Madison: -7584.1897128
Classified: Madison
Actual: Madison
Document: madison7.txt
Hamilton: -10254.0232412 | Madison: -9478.61407432
Classified: Madison
Actual: Madison
Document: madison8.txt
Hamilton: -7843.604446 | Madison: -7222.65014955
Classified: Madison
Actual: Madison
Document: madison9.txt
Hamilton: -9633.78936054 | Madison: -8851.16266742
Classified: Madison
Actual: Madison
Gamma: 5
Accuracy: 1.0

Unknowns

Document: unknown1.txt
Hamilton: -6463.46942615 | Madison: -6383.74577157
Classified: Madison
Document: unknown10.txt
Hamilton: -9299.15518625 | Madison: -9208.00140693
Classified: Madison
Document: unknown11.txt
Hamilton: -11929.4336008 | Madison: -11797.0030532
Classified: Madison
Document: unknown2.txt
Hamilton: -4367.51543379 | Madison: -4315.07850704
Classified: Madison
Document: unknown3.txt

Hamilton: -7444.39258038 | Madison: -7324.28127014
Classified: Madison
Document: unknown4.txt
Hamilton: -7238.34591265 | Madison: -7158.22915553
Classified: Madison
Document: unknown5.txt
Hamilton: -8452.50379978 | Madison: -8374.18879834
Classified: Madison
Document: unknown6.txt
Hamilton: -7752.63039605 | Madison: -7683.31444726
Classified: Madison
Document: unknown7.txt
Hamilton: -7968.63386312 | Madison: -7896.88277537
Classified: Madison
Document: unknown8.txt
Hamilton: -6073.9719955 | Madison: -6005.80661542
Classified: Madison
Document: unknown9.txt
Hamilton: -8655.335212 | Madison: -8589.98670167
Classified: Madison

So for whatever reason, again, my model things every unknown document should be classified as Madison.