Matthew Lowery
Assignment 2: Part B: Who Said it? Analysis & Write-up
Due 11:59pm 02/26/20

1. Classifier Accuracy:

The system's accuracy is .951 or 95.1 % accurate relative to guessing whether a sentence was written by Jane Austen or Herman Melville. If it would have been the case that we were testing the classifier with a text that was different than that of the training set, I would have said the accuracy would not have been nearly as high; but, because we used the same test data as that of the novel we trained the classifier with, it seems somewhat sensible that the accuracy was quite high. Otherwise, it definitely is the case that author's writing styles and writing dynamics change with each book they write so it might've given the classifier a little bit of trouble.

2. Features:

a. The gen_feats function seems to be getting rid of the all the proper nouns so they do not skew the classifier in any particular way; this, by changing any of the 'Top 35 proper nouns' to 'Monty Python' and making the classifier not bias to either of the author's writing for any of the such names.

b. For the list of the top 40 most informative features, I noticed obviously that some of the main objects/characters were very relevant in classifying the text. Things like the 'whale' let the classifier know pretty much without a doubt that the sentence would be Melville's. Also main characters like emma, starbuck, queequeg, starbuck help out the classifier. It seems to help that the setting for Moby Dick was on a boat as that language is quite specific and would rarely be used in other contexts; so that words like 'deck,' 'boat,' 'ship,' 'boats,' 'crew,' 'mast,' 'whaling,' 'sail,' 'voyage,' and 'ships' would clarify that some particular sentence was Melville's. I notice that the classifier picked up an apostrophe as entailing the text is that of Austen's which is peculiar to me. And further, it is interesting that 'thee' was entailing that the text was Melville's; it seems that this would be picked up as it's just a colloquialism of the time period relative to Austen's novel.

```
('contains-`', 1)        austen : melvil =    166.5 : 1.0
('contains-thee', 1)        melvil : austen =    162.9 : 1.0
```

3. Main Character names:

The classifier's script to neutralize the top 35 most common names did not affect the script as much as I would have guessed considering that a lot of the most informative features from before were based off of the same such names. The accuracy only went from 95.1% to 93.8%. Barely more than a 1% change in correct guesses. The top feature list remains to be mostly words having to do with the oceanic/ship/sailing setting of Moby Dick, but also adds more of such words to replace the missing main character names. Again such words like ‘deck,’ ‘boat,’ ‘ship,’ ‘boats,’ ‘crew,’ ‘mast,’ ‘whaling,’ ‘sail,’ ‘voyage,’ ‘cabin,’ ‘ocean,’ ‘mate,’ ‘harpooneer,’ voyage,’ and ‘ships’ take over the list as they are so particular to the context that Moby Dick’s story takes place in.

4. Trying out sentences:

The labeler guessed that the first sentence was written by Austen, while guessing that Melville wrote the second one. I think that both of these responses matched my expectations. For the first sentence, the classifier can account for the author’s style and it makes sense when only having to guess between two authors that the classifier would choose the right one. For the second sentence, I would say Melville only because of the use of the word White as it encompasses a major theme of the novel with the whale frequently being referred to as the ‘white whale.’ With further checking the word white is used about 300 times in Moby Dick relative to 5 times in Emma. It is definitely less likely that the classifier can be sure about this sentence.

5. Label probabilities for a sentence:

a. The classifier is 96.49% sure that the first sentence was written by Austen, while being 3.51% sure that the same such sentence was written by Melville.

b. The classifier is 40.14% sure that the second sentence was written by Austen, while being 59.86% sure that the same such sentence was written by Melville.

c. The classifier is very confident the the first sentence was written by Austen, while being a lot less sure if the second sentence was written by Melville. This makes sense considering the first sentence was written by Austen and the second sentence was not written by either Melville or Austen.

6. Trying out made-up sentences:

a. The classifier gave sentence 3 the label Melville with a 55.08% likelihood and gave sentence 4 the label Austen with a 93.14% likelihood. I would say that it could be a gender thing for both cases. As Melville’s text is almost entirely encompassed by men while Austen’s has a lot more female inclusion. I think that the classifier is more sure about sentence four because there are almost no women discussed in Moby Dick so with the ‘she’ it can be somewhat certain that it isn’t by Melville. And because both texts incorporate men, the classifier isn’t as sure about sentence three but still guesses the text that is predominantly about man.

b. I would say that the classifier saw more fragments of the word “blah” in Melville’s writing as compared to Austen’s. It guessed with a 55.97% likelihood that the sentence “blahblahblah blahblah” was written by Melville. In searching the text a good many more “bla” fragments come up in Moby Dick than in Emma. But the classifier isn’t incredibly certain, which makes sense as this is some random onomatopoeia sentence, not the words of either author. Although, I don’t believe word fragments are even considered by the classifier.

7. Base Probabilities (=priors):

a. Total number of training sentences 15152
b. There are 6672 Austen sentences
c. There are 8480 Melville sentences
d. P(austen) = 6672/15152 = .4403 = 44.03%
P(melville) = 8480/15152 = .5597 = 55.97%
e. The probabilities match the classifier’s prediction on sent5 exactly.

8. Calculating Odds Ratio:

a. 927 Austen sentences contain the word ‘very’ or ‘Very’
b. 272 Melville sentences contain the word ‘very’ or ‘Very’
c. P(very | austen) = 927/6672 = .1389 = 13.89%
d. P(very | melville)= 272/8480 = .0321 = 3.21%
e. The Austen-to-Melville odds ratio of very is 927 to 272 or 3.41 to 1

9. Feature weights in model:

a.

the weights of ‘very’ are:

```
{'melville': 0.0321306449711119, 'austen': 0.13899295669114342}
```

b.

They do match, with only very very minor differences (the hundred-thousandths place)

10. Zero counts and feature weights:

a.

feature weight for ‘whale'

```
{'melville': 0.11407852847541564, 'austen': 7.49288176232579e-05}
```

feature weight for ‘ahab’

```
{'melville': 0.04993514915693904, 'austen': 7.49288176232579e-05}
```

I notice that there is an essentially non-existent use of the word ‘whale’ for Austen and a pretty good chance for use of the word for Melville. The same is true for the word ‘ahab.’

b.

```
feature weight for 'marriage'
{'melville': 0.00029477655936799903, 'austen': 0.0038213696987861533}
feature weight for 'Emma'
{'melville': 5.895531187359981e-05, 'austen': 0.10992057545331935}
```

In this case it ‘marriage’ isn’t used much in either text, but it is still an order of magnitude more likely in Austen’s text. For the word ‘Emma,’ there is essentially a non-existent use of the word in Melville’s text, while there is a pretty good chance for use of the word in Emma’s text. This all seems to be pretty intuitive to me.

c.

A word that occurs in Austen’s work only would be ‘woodhouse’:

```
feature weight for 'woodhouse' (only in austen's text):
{'melville': 5.895531187359981e-05, 'austen': 0.03843848344073131}
```

A word that occurs in Melville’s work only would be ‘boat’:

```
feature weight for 'boat' (only in melville's text):
{'melville': 0.031776913099870296, 'austen': 7.49288176232579e-05}
```

The general idea is that if the word does not appear in the text, its feature weight will be quite minuscule. (i.e. .000075 = .0075%)

d.

```
feature weight for 'cautiously' (1 time in both texts)
{'melville': 0.00017686593562079943, 'austen': 0.00022478645286977737}
```

Its feature weights are a few orders of magnitude more likely than that of the words that were never used in the texts. So definitely a noticeable difference from something like .0075% to .022% and .018%.

e.

For internet we get a “KeyError” because it never shows up in either texts (the word didn’t exist obviously when the texts were written)

```
Traceback (most recent call last):
  File "who-said-it-naive-bayes.py", line 230, in <module>
    print(whosaid.feature_weights('contains-internet', 1))
  File "/Library/Frameworks/Python.framework/Versions/3.8/lib/
python3.8/site-packages/nltk/classify/naivebayes.py", line 253, in
feature_weights
    wdict[l] = cpdist[l,fname].prob(fval)
KeyError: ('melville', ‘contains-internet’)
```

f.

The probability for ‘she hates the internet’ being an austen sentence is quite high, being 0.899546491362117. While the probability of ‘she hates the’ being an austen sentence (again quite high) is 0.899546491362117. So it seems that the classier just ignores features it hasn’t encountered before as the probability is the same with or without ‘internet.’

11. Combining feature weights:

a. P(austen) = 6672/15152 = 0.440337909186906

b.

```
{'melville': 0.1554651574106827, 'austen': 0.16881462610520007}
{'melville': 0.0026529890343119913, 'austen': 0.003072081522553574}
{'melville': 0.5981016389576701, 'austen': 0.37636745092162444}
{'melville': 0.003832095271783988, 'austen': 0.00502023078075828}
```

c.

P(Sent3, Austen) = 0.440337909186906 ∗ 0.16881462610520007 ∗ 0.003072081522553574 ∗ 0.37636745092162444 ∗ 0.00502023078075828 = 4.314839275539011e-07

d. P(Melville) = 8480/15152 = 0.5596620908130939

P(Sent3, Melville) = 0.5596620908130939 ∗ 0.1554651574106827 ∗ 0.0026529890343119913 ∗ 0.5981016389576701 ∗ 0.003832095271783988 = 5.290609490934476e-07

e.

P(Sent3) = P(Sent3, austen) + P(Sent3, melville) = 9.605448766473486e-07

f.

P(austen|sent3) = P(sent3,austen) / P(sent3)
= 4.314839275539011e-07 / 9.605448766473486e-07
= 0.44920746343464674

g.

The probability I got from #6 was 0.44921141639835876 so this is almost identical to the probability we just calculated.

12. Performance on the development-test data:

a. The classifier correctly labeled 514 + 440 / 1000 = .954 sentences of the development test set

b. Whosaid’s accuracy on the development test data is .954 or 95.4% where the test data performance is .951 or 95.1% which is quite close.

c. The classifier labeled 25 + 440 = 465 sentences as Austen and 21 + 514 = 535 as Melville.

d. 440 / 465 = .9462 = 94.62 %
e. 514 / 535 = .9607 = 96.07 %

13. Error Analysis:

a. These do sound like Melville only in that I see masculine-gendered terms like he, fellow, himself, kings, lords. It seems like the classifier is definitely somewhat reliant on guessing Melville for such masculine-gendered terms. Otherwise, it’s really hard to say whether they were written by either author. I highly doubt I would be able to tell the difference any better in these cases.

b.

0. 0.8275435597564982 P(M)
1. **0.5578138418462254 P(M) Lowest Confidence**

sentence was – Come , he knows himself there .

2. 0.6008078675177236 P(M)
3. 0.9705122073394814 P(M)
4. 0.595692880216195 P(M)
5. 0.8095811018801172 P(M)
6. 0.5998700771690774 P(M)
7. 0.6556408967482348 P(M)
8. **0.9920235496289302 P(M) Highest Confidence**

sentence was – It is a sort of prologue to the play , a motto to the chapter ; and will be soon followed by matter - of - fact prose ."

9. 0.7587000028053951 P(M)
10. 0.9021175690663855 P(M)
11. 0.7309451259887633 P(M)
12. 0.6734505415265042 P(M)
13. 0.9879944457649772 P(M)
14. 0.6977307000923476 P(M)
15. 0.7371769422074301 P(M)
16. 0.9883894699701071 P(M)
17. 0.684360668747964 P(M)
18. 0.9113312747088789 P(M)
19. 0.8835679123758482 P(M)
20. 0.8283091713061425 P(M)