YOUR NAME HERE: Olga Redko

ASSN 02: "Who Said It?"

PART B

1. My classifier's original accuracy scores based on my random data splits:

Accuracy score: 0.945

Accuracy score: 0.951

Accuracy score: 0.948

Accuracy score: 0.953

Accuracy score: 0.958

Average accuracy score from the five scores above: 0.951

My classifier's accuracy score after fixing the random seed:

Accuracy score: 0.951

The system's accuracy is higher than I had expected because Melville and Austen are novelists from a similar era, so I expected their writing to be more similar to each other.

2. a. Words are converted to lowercase and are added to featdict.

b. Many of the words from the list of the most informative features are names. Some entries that slightly surprise me are nobody, acquaintance, happiness, pleasure, and pain because based on my general personal experience, I feel like these are fairly common words that can be used in a broad range of situations.

3. b. The new classifier’s accuracy score is lower, at 0.938 with the random seed. The new classifier’s performance did not degrade as much as I had expected because I expected the names to be more prevalent throughout the texts, and I expected them to be a more defining feature of each writer’s texts. I suppose the degradation wasn’t severe because names weren’t mentioned as often as I had expected in each writers’ texts and/or many of the names weren’t repeated enough so that they became notable distinguishing features of the sentences they were found in. The top feature list has fewer names when the value of noCharNames is set to true.

4. a. The classifier gave the label “austen” to sent1 and “melville” to sent2. This matched my expectation; since Sent1 is from Austen herself, it makes sense for the sentence to likely be labeled “austen.” As for Sent2, it’s far harder to tell what the sentence should be labeled as, but because Caroll and Melville are male whereas Austen is female, I very slightly expected a Caroll sentence to more likely be identified as a Melville sentence.

5. a. P(austen|Sent1) = 0.9649058436514013

P(melville|Sent1) = 0.035094156348597375

b. P(austen|Sent2) = 0.4014435223953792

P(melville|Sent2) = 0.5985564776046165

c. My classifier is 96.49058436514013% confident that Sent1 is Austen, whereas it is 59.85564776046165% confident about Sent2 being Melville. The difference in confidence percentages between these two is about 36.6349366047%. Since the difference is not 0%, my classifier is not equally confident on Sent2 being Melville.

6. a. Sent3 was labeled as “melville” with P(melville|Sent3) = 0.5507885836016423 and P(austen|Sent3) = 0.44921141639835876.

Sent4 was labeled as “austen” with P(austen|Sent4) = 0.9314339848201395 and P(melville|Sent4) = 0.06856601517985973.

These results make me think that male characters appear slightly more in Melville’s work than in Austen’s work, but female characters appear significantly more in Austen’s work than in Melville’s work.

b. Sent5 was labeled as “melville” with P(melville|Sent5) = 0.5596581535009568 and P(austen|Sent5) = 0.44034184649904307. Perhaps the classifier made the prediction it did because maybe Melville’s work contains more words than Austen’s work, so if a foreign word was given, because Melville’s work tends to contain more words, it was assumed that the foreign word would more likely come from Melville’s work.

7. a. There are 15,152 sentences in train\_sents.

b. 6,672 are Austen.

c. 8,480 are Melville.

d. P(austen) = 0.440337909186906, and P(melville) = 0.5596620908130939.

e. P(austen) is virtually equal to P(austen|Sent5), and P(melville) is virtually equal to P(melville|Sent5). This suggests that when a sentence with only foreign words (i.e., words that are found in neither Melville’s nor Austen’s work) is given, it is assumed to come from Melville with a probability of about 0.560 because since Melville’s work seems to contain more sentences (sentences from his work comprise 56.0% of all sentences when Melville’s work and Austen’s work are combined), by random chance and without any data training, it would seem that an arbitrary sentence would have a 56.0% chance of being a Melville sentence given the choice of only two labels, Austen and Melville.

8. a. 927 Austen sentences contain “very”.

b. 272 Melville sentences contain “very”.

c. P(very|austen) = 0.1389388489208633

d. P(very|melville) = 0.03207547169811321

e. 0.1389388489208633 / 0.03207547169811321 ≈ 4.33162293694

9. a. {'melville': 0.0321306449711119, 'austen': 0.13899295669114342}

b. The weights virtually match up with what I calculated in 8.c and 8.d

10. a. The feature weights of ‘contains-whale’ is:

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

The feature weights of ‘contains-ahab’ is:

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

These weights for Austen are negligibly small; they are so small as to suggest that the tokens “whale” and “ahab” don’t ever actually appear in Austen’s work.

b. The feature weights of ‘contains-marriage’ is:

{'melville': 0.00029477655936799903, 'austen': 0.0038213696987861533}

The feature weights of ‘contains-emma’ is:

{'melville': 5.895531187359981e-05, 'austen': 0.10992057545331935}

The token “marriage” seems to appear in both Austen’s work and Melville’s work, but it is mentioned more often in Austen’s work. It seems like the token “emma” never appears in Melville’s work since its weight is negligibly small among Melville’s sentences, but it appears in Austen’s work.

c. The feature weights of ‘contains-bates’ is:

{'melville': 5.895531187359981e-05, 'austen': 0.018357560317698186}

The feature weights of ‘contains-groom’ is:

{'melville': 0.00017686593562079943, 'austen': 7.49288176232579e-05}

The token “bates” only occurs in Austen’s work, whereas the token “groom” only occurs in Melville’s work. It seems like even if a certain token doesn’t appear in one of the work’s sentences, that work doesn’t assign a weight of 0 to that token; instead, it assigns a very small number. Also, the greater the proportion of a certain token appears within a certain work, the greater the weight is for that token within that work.

d. The feature weights of ‘contains-cautiously’ is:

{'melville': 0.00017686593562079943, 'austen': 0.0002247864528697737}

These feature weights are further from 0 than the feature weights of words that never appeared in either work.

e. I get an error message:

*Traceback (most recent call last):*

*File "/home/oredko/Desktop/cs366/A2/A2.py", line 200, in <module>*

*print(whosaid.feature\_weights('contains-internet', 1))*

*File "/home/oredko/.local/lib/python3.6/site-packages/nltk/classify/naivebayes.py", line 253, in feature\_weights*

*wdict[l] = cpdist[l,fname].prob(fval)*

*KeyError: ('melville', 'contains-internet')*

f. The classifier treats features it has never encountered in the training data as if they are absent or have no effect on the classification; the token “internet” never appears in Austen’s work, yet the likelihoods of “She hates the internet” and “She hates the” being Austen sentences are equal.

11. a. P(austen) = 0.440337909186906

b. whosaid.feature\_weights('contains-he', 1) results in:

{'melville': 0.1554651574106827, 'austen': 0.16881462610520007}

Thus, P(he|austen) = 0.16881462610520007

Similarly,

P(knows|austen) = 0.003072081522553574

P(the|austen) = 0.37636745092162444

P(truth|austen) = 0.00502023078075828

c. P(Sent3, austen) = P(austen) \* P(he|austen) \* P(knows|austen) \* P(the|austen) \* P(truth|austen) = 0.440337909186906 \* 0.16881462610520007 \* 0.003072081522553574 \* 0.37636745092162444 \* 0.00502023078075828 = 4.31483928e-07

d. P(melville) = 0.5596620908130939

P(he|melville) = 0.1554651574106827

P(knows|melville) = 0.0026529890343119913

P(the|melville) = 0.5981016389576701

P(truth|melville) = 0.003832095271783988

P(Sent3, melville) = P(melville) \* P(he|melville) \* P(knows|melville) \* P(the|melville) \* P(truth|melville) = 0.5596620908130939 \* 0.1554651574106827 \* 0.0026529890343119913 \* 0.5981016389576701 \* 0.003832095271783988 = 5.29060949e-07

e. P(Sent3, austen) + P(Sent3, melville) = 4.31483928e-07 + 5.29060949e-07 = 9.60544877e-07

f. P(austen|Sent3) = P(Sent3, austen) / P(Sent3) = 4.31483928e-07 / 9.60544877e-07 = 0.449207463

g. The figure appears to match (minuscule differences are likely due to rounding).

12. a. 954 with the given random seed.

b. 0.954. The performance is quite close—the difference is 0.954 - 0.951 = 0.003

c. 46.5% of sentences are labeled as “austen” while 53.5% of sentences are labeled as “melville.” I think the results aren’t 50-50 maybe because more Melville tokens than Austen tokens are given, so when a new sentence is labeled, it’s slightly more likely for its tokens to be found and weighted more heavily among Melville’s sentences.

d. 94.6236559139785%

e. 96.07476635514018%

13. a. These sentences sound Melville-like to me largely because masculine pronouns/titles appear often among these sentences, and male characters or at least masculine pronouns seem to appear in Melville’s works more than in Austen’s works (based on what I found in 6a).

b. “Come , he knows himself there .” was judged as Melville with the lowest confidence, at 0.5578138418462254.

“It is a sort of prologue to the play , a motto to the chapter ; and will be soon followed by matter - of - fact prose ."” was judged as Melville with the highest confidence, at 0.9920235496289302.