Connor Hanle

LIN 313

Prof Venkat

August 1, 2022

Assignment 3

**Problem 1:**

|  |  |  |  |
| --- | --- | --- | --- |
| Description of measure | Value for  Beadle | Value for  Haldane | Value  for  Questioned |
| “Average Word Probability”, that is, the average of the all of the probabilities of every word to appear in each given text; i.e., how unique are the words. Naive assumption that the author will maintain idiosyncrasies in word use | 0.002873562 | 0.002968748 | 0.0029640908 |
| Average sentence length | 99 | 182 | 100 |
| Naive attempt at matching period of writing; comparing the probability of the word occurring within Darwin’s *Origin of Species* compared to *An Investigation into Free eBooks* (<https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2952>),left-value is the average probability that some text’s word will be written in Darwin, and the right-value is the probability the text’s word will appear in the newer text | (0.011803175, 0.010415092) | (0.00984144, 0.00849518) | (0.011920794, 0.010448548) |

Given these measurements, I think we are forced to conclude that the author of the questioned paragraph is John Beadle. The uniqueness of the words measured firstly is remarkably similar, in other words, the words used in the first text and the final text have a similar ratio to all other words in all three texts. Similarly, the final text and the first text both have ~100 character average word length. Finally, they both seem to have similar word occurrences in my attempt at historical analysis. Interestingly though, it doesn’t seem to go either way, i.e., neither of the two texts (or all three for that matter) have a rating that is most similar to the older text, or to the newer text.

For my “intuitive”, non-quantitative measurements, I think what would have been the first tell is the use of the first person compared to the other paragraphs. However, all three texts provided use the first-person. So, the next best method is to try and think about the way the text “communicates” its meaning. For example, the first and the final texts read more like reports, whereas the second text seems to try and convince us about the beauty of Chicago. The second qualitative measurement, I think, is the regions mentioned: the first and last list off many different areas in the Pacific Northwest and Texas, whereas the second text only mentions Chicago (and London in an off-hand reference). While, this might not give off any meaning to most people, and is completely arbitrary, I think we might be able to say that the regions mentioned in the first and final texts are more “rugged”, if that makes any sense. What I mean is I could imagine the names listed appearing in Little House on the Prairie in the first and final texts, whereas the second text seems through its mention of places to be like a Broadway show. Finally, and equally arbitrary, but still I find convincing, is the *way* these texts talk about the places. This importantly differs from the first qualitative measurement I think in that I’m focusing here on the adjectives used compared to the style the text is presented in. The first and final seem to relay how cool the places it is talking about by mentioning ranges, mountains, rivers, and pastures. Whereas in the second text it talks about buildings, commerce, and populations.

**Problem 2:**

*I. Distance Computation*:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | d1 | d2 | d3 | d4 | d5 | d6 |
| c1 | 1.259999999999998 | 53.09000000000002 | 9.800000000000006 | 1.140000000000005 | 8.180000000000005 | 96.52000000000001 |
| c2 | 10.339999999999995 | 23.25 | 0.7200000000000008 | 5.659999999999996 | 1.2599999999999991 | 65.68 |

I calculated this using Python :)

def distance(p: List[float], c: List[float]) -> float:

return pow((p[0] - c[0]),2.0) + pow((p[1] - c[1]),2.0) + pow((p[2] – c[2]),2.0)

*II. Group Membership*

d1 = c1, d2 = c2, d3 = c2, d4 = c1, d5 = c2, d6 = c2;

So, the new coordinates of c1 and c2 would be

c1 = [1.9, 10.05, 16.450000000000003]

c2 = [4.075, 11.9, 20.85]

Here’s the work for it :)

def update\_centroid(ds: List[List[float]]) -> List[float]:

sum\_x = 0.0

sum\_y = 0.0

sum\_z = 0.0

for d in ds:

sum\_x += d[0]

sum\_y += d[1]

sum\_z += d[2]

ds\_len = len(ds) # dynamic programming baby

return [sum\_x / ds\_len, sum\_y / ds\_len, sum\_z / ds\_len]

*III. Do It Again*

First round:

|  |  |
| --- | --- |
| c1 distance | c2 distance |
| 1.1850000000000023 | 30.40312500000001 |
| 51.75499999999996 | 5.2581249999999935 |
| 9.39499999999998 | 8.14812500000001 |
| 1.1850000000000005 | 26.993125000000006 |
| 7.904999999999982 | 12.383125000000009 |
| 94.23499999999997 | 27.448124999999997 |

This results in:

new c1

[1.9666666666666668, 9.866666666666667, 17.36666666666667]

new c2

[4.733333333333333, 12.700000000000001, 21.400000000000002]

Next round:

|  |  |
| --- | --- |
| c1 distance | c2 distance |
| 3.54000000000001 | 41.53444444444446 |
| 40.72333333333329 | 4.537777777777772 |
| 5.7399999999999824 | 15.13444444444446 |
| 0.5866666666666697 | 39.94111111111113 |
| 3.5133333333333163 | 22.014444444444464 |
| 86.11333333333329 | 17.054444444444435 |

With the final values of c1 and c2 being:

new c1

[1.9250000000000003, 10.25, 17.825]

new c2

[6.199999999999999, 13.350000000000001, 22.5]

*IV. New document*

The distances between the new values is: from c1, 60.98374999999999; from c2, 14.762499999999989. We can therefore conclude that d7 will be a member of c2.

Then, we can conclude, at least from this algorithm, that d1, d3, d4, d5 all have the same author; whereas, d2, d6, and d7 all have the same author.

**Problem 3:**

*I.* At least from what we have been provided, it seems that according to the independent study, the probability of hate speech just is 0.5%. Of course, this ignores the study’s precision or recall.

*II.* The probability of holocaust given hate speech, or P(“holocaust” | hate speech) = 1500/5000, or 30%.

*III.* The probability of hate speech given “holocaust”, or P(hate speech | “holocaust”) = (.5)\*.3 /(.153), or P(hs | “holocaust”) = P(“holocaust”)\*P(“holocaust” | hate speech) / P(“holocaust”), which is around 98%.

**Problem 4**:

*I.*

Total = 564 tweets

|  |  |  |
| --- | --- | --- |
| Description of Measure | Obj. Class | Subj. Class |
| Obj. Ann. | 112 | 41 |
| Subj. Ann. | 0? | 411? |

*II.1* The classifier produced 41 false objectives as a detector of objective tweets.

*II.2* The classifier had 73% precision

*II.3* I do not know how to classify recall, given that the information we have does not provide the amount of falsely classified subjective tweets, therefore making recall calculations impossible.

*III.1* The information given does not provide enough information about the subjective rating of the tweets, it only provides the false judgments regarding “objective” classifications. As “It labels 153 tweets as objective and the rest as subjective. Of the 153, only 112 were actually objective” is the only information given.

*III.2* I don’t know how to calculate precision either in this case, given that I need information about the true positives of in the subjective class.

*III.3* The same problem here, I am unable to accurately calculate the recall of the classifier given the information provided. For example, there could be as many as 411 falsely annotated as subjective tweets.

**Problem 5**.

1. Positive/negative tweet ratio over time might want to prefer recall in this measurement, given that even in the case where there are many false negatives, given the sheer quantity of tweets involved we should focus on the amount of positive predictions made. True information in this case would be more important for the “pulse” monitor, as one might be led to false judgments about people’s emotions about some subject if they ignored recall (maybe most tweets are positive, but you might also want to take into account the genuinely negative tweets).

2. Precision should be focused on in this situation, given that the researchers are interested in understanding *interesting* tweets, and not just sheer quantity of positive or negative value judgments. So, rather than attempting to monitor the pulse of a huge population, they would probably be more interested in having a sheer amount of data available to them to pick and choose fomr.

3. In this case, precision might be favored again, as for all intents and purposes of the company, they probably do not care if a specific ad gets shown to more people outside of their target audience. As long as they do hit their target audience to some extent, they ought to be happy.

*Here’s the code for the text analysis and the centroid stuff:*

*https://github.com/hanleyc01/LIN-313/blob/main/text-comp/src/main.rs*

*https://github.com/hanleyc01/LIN-313/blob/main/centroid/centroid.py*