IDS 703 HW Assignment V
Submitted by Jiahui (Heather) Qiu & Song Young Oh
October 18th, 2022

Assignment Objectives: You should turn in a document describing the third method that you used and discussing all of the results.

The third method is a TF-IDF variant. For term frequency (TF), we divided raw counts of terms in each document by the raw counts of the most frequently mentioned term of the document. The approach enables us to calculate the augmented frequency, which mitigate biases on longer documents. Then, we multiplied it by 0.5 and added another 0.5 for normalization. For inverse document frequency (IDF), we added a value of 1 to the existing denominator (i.e., the number of documents in which a word occurs) before scaling logarithmically. The normalization prevents a division-by-zero when a term does not appear in any documents. In addition, we added a value of 1 for smoothing again after log calculation so that the value of IDF will not become zero if the ratio inside the log function is equal to 1.

The performance of the three classification methods are recorded in the following table:

Method	Accuracy Percentage
1	78.46 %
2	78.46 %
3	83.08 %

The third method, a variant of TF-IDF weighting, produced a better outcome when compared to the first two methods: an accuracy of 83.08 percent. We achieved better results because (1) the augmented term frequency avoids bias towards longer documents by normalizing the term frequency, and (2) the smoothing in the inverse document frequency prevents the results from becoming zero.