Revanth Chowdary Ganga (rg361)
Divya Sharma (ds655)

NLP Assignment 4: Document Vectors

Group Members:

1. Revanth Chowdary Ganga (rg361)
2. Divya Sharma (ds655)

The Documents used in this exercise are from the Gutenberg Dataset:
H0: austen-sense
H1: carroll-alice

The Following 4 encoding methods were tested out:

S.No	Method	Accuracy Achieved	
		Train	Test
1	Token Counts (with One-Hot encoding)	98.87%	96.86%
2	TF-IDF (Term Frequency-Inverse Document Frequency)	99.93%	97.16%
3	LSA (Latent Semantic Analysis)	96.63%	94.92%
4	Word2Vec	93.89%	90.86%

Out of the 4 tested methods, TF-IDF had the best performance, and Word2Vec had the worst performance.

The 2 “simpler” methods (Token Counts, TF-IDF) had better accuracy than the more advanced and complex methods (LSA, Word2Vec), this may be because of the following reasons:

1. Word2Vec:
    a. might not be performing well since the pre-trained model we used (Google News Data) might have different semantic data as compared to our input documents (novels).
    b. Additionally, there may be words in our novels that are out-of-vocab for the pre-trained model.
2. LSA:
    a. Since we are restricting the dimensionality (to 300 dimensions only), it might be restricting the granularity and the amount of information the model can have to distinguish between the documents.
3. General:
    a. Data Size: The simpler models might be performing well given the comparatively smaller size of training data as the complex models generally tend to perform better with larger training data in which case the simpler models start to fail or are not as efficient.
    b. Preprocessing: The data we used has been tokenized but other operations like removal of stop words, stemming, lemmatization, etc. haven’t been performed which may improve the performance of the complex models
    c. The complex models while capable of capturing more semantic information, are also more prone to overfitting, esp. with smaller datasets like in the current case.