BUS 243 Name:

Quiz 2.

Problem 1. What is the main advantage of using the TF-IDF representation for text data compared to other methods like bag-of-words?

A: The main advantage of using TF-IDF over other text representation methods is that it considers not only the frequency of a word in a document (as in bag-of-words) but also its importance in the entire corpus. TF-IDF weights words based on how often they appear in the document but also down-weights words common across the entire corpus, which may not be as useful for discriminating between different documents. This helps to highlight words that are more unique and important to a particular document, and can lead to better results in tasks like text classification or information retrieval.

Problem 2. In a corpus of text where the frequency distribution of words follows Zipf's law, why does a log-log plot of word frequency against rank produce a straight line?

A: The log-log transformation of the power law is just a straight line.