IST736 Text Mining
HW2

Vectorization

In this assignment, you will vectorize the data that you collected in HW1. Because the goal is to identify the public sentiment toward AI on social media, you need to think about what vectorization options, regarding both what to count and how to count, would be the best for this goal. Make sure to explain the decisions you made during the vectorization process, e.g., if you removed stopwords and why.

Write a report to include the following information:

(1) Briefly recap how you collected the data.
(2) Describe your vectorization choices and corresponding result. For example, if you chose to do stemming, how did the vocabulary size change after stemming? Did the stemming eliminate important linguistic information that you'd rather keep, or not?
(3) Conclude with the best vectorization option(s).

Your report should provide sufficient information for others to replicate what you did. Submit your report with your original data file and the vectors from your best vectorization options.

Follow the HW1 requirement on formatting and grading rubrics.