Homework 5

In this project, I define Knowledge Base as a SQL that keys are the term-frequency & inverse-document-frequency scores of all tokens; and entries are the index of each document/sentence. Given the input sentence/query, the query need to be tokenized and preprocessed before performing look-up. Say the query sample is “Trump and Pence will have a debate next week”. This sentence will be tokenized and preprocessed to count the word frequency. Then, the calculated word frequency of each token is to look for the sentence entry/index that has the closet tf-idf score for that token.

This is the screenshot of the SQL knowledge base:

Text

Description automatically generated

Sample chatbot dialog

Command: “cnn news”

* Parsed into tokens: [cnn, news] -> cnn: 1, news : 1 (term frequency)
* Then, the output could be the first entry in the query:

Text

Description automatically generated