ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Problem: when implementing text search over a corpus (a collection of documents), systems often rely in part on the term frequency (TF) of words in a document. A TF score is computed for each word wi in each document djby computing the frequency of that word in that document.

$$TF(w_i, d_j) = \frac{\textit{the number of times word } w_i \textit{ occurs in document } d_j}{\textit{the count of words in document } d_j}$$

Write a program that takes as input a set of documents (sample link in email) and a list of words, and returns the document with the highest TF score for each word and the TF score for that word in that document. To break the document into words, you can strip out punctuation, split by whitespace, and convert everything to lowercase. Please include documentation for running your program as well as the the output for the words “queequeg”, “whale”, and “sea”. Try to make it easy-to-use and efficiently implemented.