Please do your homework using an R script. Homework is NOT collected in this class. However, on the day it is due, you may be asked to share your screen and run some of your code when we discuss this homework in class. This will count towards your participation grade.

1. We are going to create a function that will take as input a corpus of documents and will return the most informative of those documents.  
   1. Create a function called bestDocs with 3 arguments; co, which should contain a corpus, num, which should contain an integer equal to 1 or greater, and n, which should also contain an integer equal to 1 or greater. The argument called num should by default be equal to 3, while the argument n should by default be 10. Even though all functions should have error checking, for now we will not do error checking in this function; we will assume that the correct information will be passed to it.   
        
      The function should select the num most informative documents from co and return a corpus with just those num documents. The criteria to use to determine whether a document is informative would be the number of high frequency phrases it contains; it is considered more informative if it contains more high frequency phrases. The set of high frequency phrases equals the n (argument) number of phrases with the highest frequency over the entire corpus.  
        
      Note that you will need to create a phraseDoc object for the input corpus; when doing this, I suggest you select only principal phrases with a minimum frequency of 3. Also select the silent option so you do not get progress information when running the function (check ?phraseDoc).  
        
      In addition, the meta data for each of the num most informative documents should be updated with the index of the document in the original corpus as oldID, and the number of high frequency phrases it contains in hfPhrases.
   2. Create a corpus in R from the PubMed file Covid19March1.txt. Create a phraseDoc object on it called pd, and a term-document matrix called pdm.
   3. Find the 10 most informative documents using the function you created in a), based on the default number of high frequency phrases. Store the corpus with those 10 documents in the variable bd. Show the old indices and the number of high frequency phrases for all 10.
   4. What is the index (in the original corpus) for the document with the highest number of high-frequency phrases? Store this index in the variable id. How many high frequency phrases does it contain? You can access this document in the original corpus co as well as in the corpus returned by the function (bd); inspect it both ways. Show all principal phrases in this document and compare it to the top 10 high frequency phrases in co.
   5. Do the same as in c) but this time only use the top FIVE high frequency phrases. Does the same document “win” this? How about if you use the top 20? If you find a difference, inspect the document in question.