Please do this quiz using an R script and submit it via Moodle. All questions should be answered using R. Make sure your code is well-organized by each question, commenting the question numbers. Avoid hardcoding. **NOTE**: You may use any functions I have provided in the lectures.

1. The tm package comes with a second set of Reuters files under the acq directory instead of the crude directory. Find where this is on your file system using system.file("texts", "acq", package = "tm"), then load the contents into a VCorpus in R. Make sure to use the correct functions to load it (the ones specifically for Reuters .xml files). Call your corpus acq.
   1. How many documents are there in this corpus?
   2. Create a corpus meta field called Description and set it to “All about companies”. View the corpus meta field afterwards.
   3. Create a TermDocumentMatrix object, removing punctuation, stop words, and numbers. Also remove words from a document if they occur less than 3 times in the entire corpus, or if they contain less than 4 characters. Call the object tdm. Find out its sparsity (it is a percentage), and how many words it contains.
   4. Create a term-document matrix from tdm; call it tdM. Find the document with the least number of words in the term-document matrix, then display those words with their frequencies as a matrix (so not as a vector). Do NOT show words with frequency zero. **Hint:** The function which.min might be useful here. Also remember the drop argument when subsetting.
   5. Find the frequencies of all the words in the term-document matrix, then show the 10 with the largest frequency. What is the frequency of the word “company”?
   6. Store the indices for all 3 documents that contain the word “products” in the variable idx. Show the content of each of those 3 documents, then store the first of those three documents in the variable called ptd. **Hint:** When you subset a corpus (instead of using the double square brackets), you obtain another corpus.
   7. Change the content of ptd so that each occurrence of “\n” is replaced with a space. Then inspect ptd. Also find the date the text in this document was produced.