Assignment for application Data Science track in Information Studies Master’s programme

It is often that data available in databases, or flat files of a company/organization need to be pre-processed, and stored in the appropriate data structures so that they can be easily used by data mining or machine learning algorithms. This data may come in a variety of forms and modalities (e.g. structured records, unstructured, or semi-structured text, images, etc.). In this assignment we will consider a sample dataset (collection.txt) that contains three articles from LA Times in a semi-structured format. The tags in the collection dictate the beginning and the end of an article (<doc> and </doc>), the article id, the headline of the article and the main text (<text> and </text>), along with several other information on the article.

PART I: Design and code up a class that can pre-process and store the LA Times articles. Specifically, the methods of the class should take as an input the LA Times articles collection, extract each article in the collection, and construct a hash table, the key of which is a word (in the collection) and the value a linked list of all the document that contain this word, and the count of the word in each document. For example, if the word “the” appears in all three articles, 20 times in the first, 34 times in the second, and 12 times in the third, while the word “author” appears 7 times in the first, 3 times in the second and does not appear at all in the third, the hash table should look as follows:

[the] -> [1, 20] -> [2, 34] -> [3, 12]

[author] -> [1, 7] -> [2,3]

Create an object of the type of your class and use the data collection to initiate it. Think about how could you handle different forms of the same word, e.g. "author", "Author", "authors". Turn your code in a pdf document.

PART II: Generate a plot (histogram should be good enough) of the count distribution of the words in all documents (that is, the x-axis is the number of times a word appears in the entire collection - total count -, and the y-axis the frequency of that count). Characterize this distribution.

Have a look at the example solution to compare.