Keyword And Associated Synonyms Analyzer

1. Using decomposition, what are the primary sub-problems that need to be solved in solving the overall problem?
   Ans: The primary sub problems are as follows:
   a. How to use the thesaurus as part of the solution.
   b. Going over each file in the corpus.
   c. Counting the occurrences of the word.
   d. Counting synonyms of each word in every document by comparisons with the thesaurus.
2. Using pattern recognition, what patterns do you see in the solution, i.e., what processes need to be repeated?
   Ans: Some of the patterns are:
   a. Comparison with thesaurus for keyword's synonyms.
   b. Iterating over the entire corpus for finding keyword occurrences.
3. Using data abstraction and representation, how would you represent the thesaurus, the corpus, and each of the documents in the corpus?
   Ans: The data that will be required to build the solution is:
   a. Number of directories and subdirectories in the corpus.
   b. Number of files in each subdirectory.
   c. Type of files in the entire corpus.
   d. Synonyms from the thesaurus.
   e. Data about the thesaurus i.e. if it's an API then API Documentation to use it etc.

   Data not required is as follows:
   f. Information about the filesystem.
   g. Data about the user using the system.
4. Using the results of the first three pillars, what is the algorithm that you would use to solve this problem? Describe it in as much detail as possible.
   Ans: The algorithm is as follows:
   a. In the corpus, iterate over each directory and first check if it contains a subdirectory.
   b. If it contains a subdirectory, move to it and perform step a.
   c. If no subdirectory, start iterating over each file.
   d. Open the file and read its contents into memory.
   e. For robustness, lowercase the entire contents of the file.
   f. Tokenize or generate all words from the file.
   g. Iterate over the list of words and maintain a count of its exact occurrences.
   h. Get all synonyms of the keyword.
   i. In another loop, instead of the same one to reduce Big-O complexity, compare the synonyms to the list of words in step f.

    j. Increase count for each matched synonym.
    k. Repeat this entire process for all files.
    l. Another approach can be to use text similarity as well rather than this entire manual effort.
5. Describe a problem that you may face -- either in your career or in everyday life -- that involves determining the number of occurrences of a word and its synonyms in a corpus of documents.
   Ans: A problem would be to determine the overall sentiment of a given set of documents or text. Occurrences of a word and its synonyms directly impact the overall rating or sentiment of a document. An example would be to determine the overall score of a certain playlist of music based on the reviews given by people to the playlist.