PEER-GRADED ASSIGNMENT: APPLYING COMPUTATIONAL THINKING TO KEYWORD DETECTION

Problem: count the number of occurrences of a keyword and its synonyms in a corpus of text documents.

1. Using decomposition, what are the primary sub-problems that need to be solved in solving the overall problem?

* Select or determine the keyword
* Interrogate the thesaurus to establish appropriate synonyms for the identfied keyword
* Select or determine the corpus of text documents to search
* Undertake a search methodology to quantify the number of occurences of the keyword and cognate words in the corpus

2. Using pattern recognition, what patterns do you see in the solution, i.e., what processes need to be repeated?

* For each keyword synonym in the thesaurus, review the appropriateness of the word to the intended contextual meaning of the keyword
* For each document in a broader corpus of documents, review the appropriateness of the document for inclusion in the final corpus of documents for quanitative assay
* For each word in the selected list of keyword and synonyms, apply a search algorighm to the corpus to quantify it occurence

3. Using data abstraction and representation, how would you represent the thesaurus, the corpus, and each of the documents in the corpus?

* Thesaurus: this can be represented as a dictionary of key-value pairs, where each key will represent a single word with a matching value which will be a tuple of single or multiple synonyms
* Documents of the corpus: these can be represented as a continuous string of text, of varying lengths
* Corpus: this can be represented as an array of strings, with each string representing a document of the corpus

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.

a. Select a keyword
b. Hash the thesaurus dictionary for the keyword, and pull the matching synonyms
c. For each synonym returned, either retain or reject, based upon appropriateness to the context of the principal keyword
d. For the keyword and each selected synonym, iterate over the array of strings constituting the corpus, searching for frequency of the word
e. Add the frequency of each word in the corpus to a dictionary, with key-value pairs representing each word iterated, and the frequency of occurence within the corpus

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. The problem you face may be much bigger than that and require that calculation as only a small part of the solution, but should involve looking through some collection of text and looking for certain words.

* A problem I typically face in my job is to find word grams that have specific patterns. To this effect, I need to create regular expressions which capture specific patterns of the gram, for which I am searching databases. I need to be mindful of the format of the fields that are being parsed for this data representation, so additional steps around pre-processing are critical.