In this project, you will use computational thinking to develop and then implement an algorithm to solve the problem of counting the number of occurrences of a word 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?

Sub-problems:

- Locating the corpus of documents (or provide an input for an exact location)
- Aggregate list of words from all documents
- Counting words that are identical from each document in the corpus
- Comparing words within the document to match synonyms
- Aggregating identical word count with synonyms and providing a new count
- Return the answer in some format

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

   Patterns:

   - Scan a document and add the words to an aggregated list
   - Scanning and generating word count for a single word
   - Locating synonyms for a single word

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

   Data (corpus):

   - Include filename (or at least a unique location/identification for each file)
   - Include location of the corpus
   - Exclude file size
   - Exclude file count

   Data (document):

   - Include words
   - Exclude punctuation (maybe some exceptions like hyphenated words)

   Data (thesaurus):

   - Include source word
   - Include synonyms of source word
   - Exclude synonyms of synonyms

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.

   Algorithm:

   1) Locate a corpus (or be given a location)
   2) Scan the location for unique documents
   3) Extract words from each document into a super document which is an aggregate of all words from the corpus using the following pattern
      a. Take file in the corpus list that has not been added
      b. Append all words from this file into aggregated file (create this file if this is the first loop)

   4) Scan the aggregated document with the first word in the document – check if it or any of it's synonyms have been counted and if not, count its occurrences using the following pattern
      a. Count the number of uses of the exact word and add to the total
      b. Check for thesaurus matches
      c. Count the number of exact matches for each thesaurus match and add to the total
      d. Return total count

   5) Repeat 4 above starting with the second word in the document
   6) Return final list of words and their count

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.

   Our problem:
   Passwords in plain text in source control. We need to scan our repositories for "password" or other synonyms and generate issues if any are found.

   It is a little bit different than this example, as the count is not the most important part – but still is a similar problem. We may also need to expand this to find generated keys in plaintext – and alert the team to encrypt those keys