Counting the number of occurrences of a word and its synonyms in a corpus of text documents

Decomposition

- First, we need to get key inputs, which includes the keyword, its synonyms, and the corpus to search through
- Second, collect data / documents (determine the medium that you want to extract the data from – email / online reviews / textbook etc.
- Third, we need to parse the corpus for and count how many times the keyword and its synonyms appear

Pattern recognition

- Getting key inputs
    - Step 1, determine the keyword
    - Step 2, retrieve the associated synonyms of the keyword
- Collecting data / documents
    - Step 3, identify all the documents that we want to analyse and add them to the corpus
    - Step 4, clean data from documents
    - Step 5, repeat step 4 until all the data are cleaned
- Parsing the collection of documents
    - Step 6, parse each document and check each word in the document against our collection of keywords
    - Step 7, repeat step 6 until all the documents have been parsed

Data representation and abstraction

- Thesaurus: a collection of keywords
- Corpus: a collection of documents
- Each document: a file / a string

Algorithm

- **Start:** User inputs keyword →
- User inputs close synonyms of keyword from Thesaurus and create a collection of keywords to check against →
- User inputs a set of documents that they want to add into the corpus →
- Check the data structure of all the documents to ensure that the data is readable by the computer →
- If the data is not readable, either find another data source or convert the data into a readable format →
- Clean up the data by extracting essential data identified under Data Representation →
- Repeat the previous step until all the data is cleaned →
- Once data is cleaned, begin parsing through each document within the corpus and matching against our collection of keywords →
- Repeat previous step until all the documents have been parsed

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.

One problem in my career relating to the above problem is that of analysing financial market sentiment. By analysing comments on social media websites, to news articles, to market commentaries, we can determine if the overall market sentiment is trending positively or negatively.

This can be used in conjunction with real time market data and algorithmic trading applications to take advantage of momentum trades, and potentially make a profit.