Search Locality

# Problem Domain Analysis

As a staff researcher, you must repeatedly search through a set of plain text documents in order to do your work.

A common need is to find documents where two different topics or search terms are discussed in the same context, for example: oncology and morbidity, or virtualization and agility.

"Context" within a document can be ambiguous or ill-defined, but in this case, "context" is defined as the two search terms occurring within N words of each other irrespective of sentence, paragraph or page boundaries.

Using a programming language of your choice, develop a piece of software that can iterate over all the plain text documents in a directory and return the set of documents where two search terms occur within N words of each other.

Extra credit:

In addition to allowing two search terms, allow the user to specify phrases instead of terms, e.g. "Dr. Fowler" and "Continuous Delivery"

# SOLUTION DOMAIN ANALYSIS

## Solution flow

no

yes

Start

Open directory and get text documents

Set directory, first\_term, second\_term, word\_distance

Search next document

Found match

Save match info

Search for next match

Output match info

no

Any Documents left to search

yes

Done

## Test Cases

|  |  |  |  |
| --- | --- | --- | --- |
| # | Input | Output | Notes |
|  | No directory | Error Message “Directory invalid” |  |
|  | No files | List of files should be empty |  |
|  | One text file | List of files contains 1 entry |  |
|  | One text file and other non-text files | List of files contains 1 entry |  |
|  | Test string with no match | Empty list returned |  |
|  | Test string with one match and words are in range | Returns list with 1 match |  |
|  | Test string with one match and words are not in range | Empty list returned |  |
|  | No files contain match | Match count is zero |  |
|  | No files contain match | Return ‘No Matches found' for output |  |
|  | 1 File contains match | Match count is one |  |
|  | Multiple files contain multiple matches | Match count is correct |  |
|  |  |  |  |

# Assumptions

1. Only working with files in the named directory. Sub-directories will be ignored.
2. All files to search will end in “.txt”