## Software Mining & Analysis - Lab on Information Retrieval for Software Engineering

### Objective

Information retrieval (including natural language processing techniques for texts, multimedia processing techniques for image/audio/video) may be applied to various software engineering data and help solve many software engineering tasks. In this lab, you will learn to use vector space model (VSM) to solve a retrieval problem, which is termed as "concern localization" or "concept/feature location" or "documentation-to-source-code traceability" in software engineering.

### Requirement

* Read one existing paper [1] and understand the problem addressed in the paper;
* With the provided dataset, perform experiments to retrieve top-k documents similar to the given queries.
* Coding would be needed for this exercise.

### Tasks

* Download the resource file [2] on eLearn.
* Read at least the abstract and introduction of the paper [1] to understand the problem that you need to solve. The readme file briefly describes the provided dataset.
* Task 1: Write code to
  + Compute the term frequency and the inverse document frequency of every term contained in the documents
  + Create vector space representations of all documents
    - Use the logarithm TF-IDF weighting scheme
  + Create vector space representations of all queries
    - Use the logarithm TF-IDF weighting scheme
  + Compute cosine similarity of each pair of query and document
  + For each query, find the top-10 most similar documents
* Task 2 (Optional): Write code to
  + Create inverted index for every term in the documents
  + Use invert index to speed up the computation of cosine similarity among queries and documents
* Task 3 (Optional): try with different natural language processing and information retrieval techniques mentioned in the paper [1] to retrieve top-10 most similar documents for the queries, and compare and contrast the output lists from different techniques.

### References

1. Shaowei Wang, David Lo, Zhenchang Xing, Lingxiao Jiang: Concern Localization using Information Retrieval: An Empirical Study on Linux Kernel. In the proceedings of the 18th Working Conference on Reverse Engineering (WCRE), pp 92-96, 2011.
2. A zip file on eLearn, containing the paper [1], a readme file, a set of documents, and a set of queries.

### Submission via eLearn Dropbox

* A compressed file containing all of your code for performing the above steps.
* A report clearly describing what you have tried (including your pseudo-algorithms for calculating TF-IDF and similarities, even if you use third-party libraries), and list all the top-10 most similar documents for each query.