## Software Mining & Analysis - Lab on Empirical Study for Software Engineering

### Objective

Software development and maintenance tasks often involve human, including developers, testers, managers, suppliers, users, etc. Their actions affect software quality, costs, productivity, etc. Analyzing logs that record their activities can provide useful insights for factors that can help to improve quality, productivity, and reduce costs. In this lab, you will replicate an empirical study of historical data collected from StackOverflow, and answer several research questions.

### Requirement

* Read one existing paper [1] and understand the study performed in the paper;
* With the provided dataset [2], perform experiments to answer the given research questions.
* 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] first to understand the problem that you need to solve.
* Look at the readme file describing the provided data format.
* Write code that can parse the given dataset.xml file and extract any needed attribute.
  + There are many different ways to parse XML files. Feel free to choose any way, or write your own simple parser specifically for this XML file. E.g., [3].
* Use the experimental details and results presented in the paper [1] to help you in understanding the following research questions, and then answer these questions using the dataset (coding may be needed):
  + a. What are the distributions of the contributors who post questions?
  + b. What are the distributions of the contributors who answer questions?
  + c. Do developers who ask questions answer questions too?
  + d. Do developers receiving help help others too?
* The answers for some of the questions may be better presented in the form of diagrams/plots/histograms/etc.
* Optionally, think of additional research questions that you can answer using the dataset. For each of these questions, please describe why it is interesting and provide an answer for it.
* Optionally, take a look at the latest data available from StackOverflow (links are given in the readme file [2], and propose additional research questions or hypotheses that may be interesting and you may answer based on the latest data.

### References

1. Shaowei Wang, David Lo, and Lingxiao Jiang: An Empirical Study on Developer Interactions in StackOverflow. In the proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC), pp 1019-1024, Coimbra, Portugal, 2013
2. A zip file on eLearn, containing the paper [1], a readme file, and a dataset.xml.
3. Oracle JavaTM tutorials: Java API for XML Processing (JAXP): <http://docs.oracle.com/javase/tutorial/jaxp/>

### Submission via eLearn Dropbox

* A compressed file containing all of your code for performing the above steps.
* A short report clearly describing what you have tried (including your pseudo-algorithms, if any, for generating answers for the research questions), and discuss possible implications/applications of your answers.
* Optionally, describe the additional research questions as mentioned in the "Tasks" section.