**CCSW-438 Advanced Topics in Software Engineering**

**Mining Software Repositories –** Spring, 2022

Course Project (20 Marks) and Course Research Paper (20 Marks)

**Project Description**

In this course project, you need to work with your 2 to 3 members group from the research track selection to the writing-up phase. In addition to ***the empirical work***, you are also asked to submit a ***short research paper*** describing your project background/literature review, objectives, methodology, results, conclusions, ..etc (details below).

**Phase 1: Planning.**

You need to decide the (1) project track, (2) problem to investigate, and (3) assign a group leader on the following link ([Click here](https://docs.google.com/spreadsheets/d/1gTJHOiv6XoIsCU9YIyGY37meAK3gvu3gb1F_hE86bMI/edit#gid=0)).

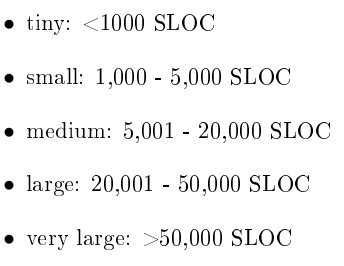
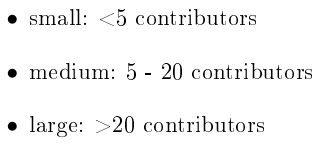
**Phase 2: Deciding and cleaning the dataset.**

1. Check the course datasets on Kaggle and decide the appropriate one to your problem definition and research track.
2. Take a look and explore the data.
3. Decide the languages you want to compare. Make sure the included languages can be categorised into at least 2 language classifications. If not, discuss and justify your language selection with me.

Languages : Java, python, C, Go, Swift, C++

1. Static vs dynamic
2. Memory and unmanaged
3. Select a *random* sample on *500-1000* projects.
4. Download the selected data, and clean them before analysis (remove duplicates, null and empty values).

**Phase 3: Analysis and Statistics**

1. Start with descriptive statistics for the dataset (N: sample size, mean, median, maximum, minimum, standard deviation).
2. Per project, specify the project size (small, medium, large.. etc.) in terms of SLOC and number of contributors. The scales we’re using here is:
3. List how many projects per language in your dataset.
4. How many projects per language classification (and per group).

**Note:** you can do this using statistical tools such as MS Excel, SPSS, Python language, R language.

1. Then, based on your selected track and research problem, investigate the dataset further.

Check the course slide and apply similar approaches.

For example, if your track is popularity and trends, draw a time series diagram for the number of projects per language individually and per language group over the years. Then, run a hypothesis U-test per language classification to check if statistical significant difference exist.

**Phase 4:** **Mid-project discussion.**

We will have a 10-minutes follow-up meeting for your projects. This is to get insights about your progress so far, and to give the help + approval to carry on working.

**Phase 5: Writing-up.**

You have to submit a short paper associated with this empirical investigation. You don’t have to wait until the end to write the paper up. *Start early*.

In the paper you need to have the following sections:

* Relevant Title.
* Introduction: problem definition, objectives, research questions to investigate.
* Background and related work.
* Methodology
* Results
* Discussion
* Conclusion

Paper format:

* Font type: Times New Roman
* Font size: 11
* 2 columns paper
* Number of pages: ~ 5-8 pages.

**Submission**

Submit the files of the empirical work in addition to the short paper on the course submission page on Blackboard. Deadlines will be announced later.

**Presentations**

You will have to present your work in 10 minutes. After that, you will be asked a couple questions to assess your knowledge. Deadlines will be announced later.

**All the best 😊**