SENG404 Project - Proposal

Learning Outcomes

- Discover SE data problems
- Examine possible SE data sources

Submission details

Submit:
1. A PDF report (Brightspace)
2. A Markdown file (repo)
describing the study you wish to design with the following info:

- Team Members
- Motivation
- Research goal or question
- Expected data sources
- Research strategies
- Potential limitations
- Expected results
- Core references

Proposal should be considered a working document, ie will be updated once discussed in the upcoming meeting (Monday!)

Proposal Tips

Spend serious time on the proposal - it will guide project

Should answer the following questions:

1. What is the software problem you're focused on?

    - Even for a new study, you should have a clear sense of other approaches to this problem.
    - Your life will be easier if you have a clear guide.
    - Don't worry if it feels like "cheating" because the original paper was so clear and the data so accessible. We can easily add complexity; it is very hard to take it away.

2. The dataset you are using, and the pilot experiments you have done with it.

- Don't just trust that the paper URL is still there, or that the data is accessible or useful. Download it, load it into R/Jupyter, and do some simple experiments.

3. Any other tools you will need, how well you know them, and what they cost.

   - For example, you might require a .NET component that the original study used, but you do not have a Windows machine. Or you will use a DL approach that requires Google TPUs to train.

4. The research questions you are trying to answer.

   - What is the contribution your paper will make?
   - I suggest using the answers to #1 above, checking what those papers say was either hard, interesting new directions, or questions they didn't have time to answer.
   - You might also find gaps in the original analysis.

5. Rough sense of methodology you will follow and who will be doing each task.

   - At the very least there will be writing of the paper, creating the video, running the analysis, processing/preparing the data, writing analysis code, reading relevant background papers, … start outlining that.
   - Show some table or Gantt chart thing to codify this with your team.
   - Marks will be lost for superficial reports; you need to have a firm idea of how long each task will take.
   - I like MarkWhen for timelines.

6. Relative to the methodology, specify exactly what the workflow is:

   - Data sources, filtering criteria, data science algorithms to use, analysis validation etc.
   - Don't underestimate how long it takes to acquire data and write analysis code.

Project Ideas

- Should be recent (related papers in last 10 years)
- Research Q should be clear, either derived from an existing paper or from one of the following sources:
  - Software Engineering Research Ideas
  - Analyze This!

Data Sources

- Software Heritage Graph
- GHTorrent
- TravisTorrent
- GrimoireLab
- Datasets from MSR Data Challenges
- API's (ie github API)
- Awesome-MSR