11/28/2020
CS 410 - Text Information System
Difan Gu, Yanyue Wang, Wen Long

Team APlus Progress Report

1) Which tasks have been completed?

- Researched pattern mining and generated ideas on how to demonstrate the significance of analyzing frequent patterns with context units.
- Understood the main steps and algorithms (hierarchical and one-step microclustering, etc.) mentioned in the paper.
- Loaded and cleaned source data from https://dblp.uni-trier.de/xml
- Implemented vector space modeling on the DBLP dataset mentioned in the paper, constructed a set of frequent models.
- Selected context units as patterns (minimal units that carry semantic information in a dataset).
- In summary, we completed about 30% of the paper.

2) Which tasks are pending?

- Redundancy removing
- Strength weighting for context units
- Extracting strongest context indicator
- Extracting representative transactions
- Extracting semantically similar patterns

3) Are you facing any challenges?

- Some parts of algorithms are too complicated to implement within a short period of time. We may need simplification in order to deliver a reasonable outcome.
- dataset is large and we may need some time to perform more detailed data cleaning and manipulation
- collaboration of the project development