CS 410 Project Progress Report
2023-11-17

Predictive Maintenance Scheduling Using Text Analysis (Free Topic)

Team Members

Captain: Gabe West (NetID: gawest2)
Team Member 1: Anastasia Serebryakova (NetID: as124)

Completed Tasks

- Cleaned training data set
  - removed references to individual's names
  - swapped proper names with unique general name references
  - removed some incomplete data from the schedules
  - removed some unused fields of ICS Events in the json data
- Wrote scripts to parse training data and generate intermediate representation of data set
  - automate pre-processing of data in terms of tokens, counts, and hour ranges when events occurred
- Implemented general EM algorithm using Numpy
  - using numpy, wrote a small library of functions to do Expectation Maximization for specific data set of calendar events based on ICS SUMMARY field terms
- Implemented command-line tool for EM topic modeling based on event subject query string and specifying number of topics to try
- Wrote documentation for all implemented library functions via docstrings
- Added python typing to all implemented library functions.

Pending Tasks

- Implement LDA and LSA topic modeling for data set
  - Need to write library functions akin to the general EM functions that implement LDA and LSA for the ICS data set. Implementation will be around using the Gensim topic modeling library
- Implement evaluation / comparison framework of general EM vs. LDA vs. LSA
  - initially we plan on trying different automated topic metrics for evaluation as used in Gensim topic coherence
  - implement novel / naive approach to evaluation using scoring suggested hour range based on term counts in the query at suggested hour ranges normalized across total term counts in the query
- Testing Testing Testing
- Documentation of project and how to use

Challenges

- Main challenge is determining which metric is useful for evaluation of suggestions made by the system.