CS 410 Final Project Progress Report: Apply state-of-the-art text retrieval methods on electronic medical records

Prepared by:

- Hanyin Wang (hanyinw2)
- Yuanpei Ma (yuanpei3)
- Yeyu Ma (yeyuma2)

1. The progress made

Task
● Completed the required CITI research training (Data and Clinical Sample) to get access of MIMIC-IV data
● Experimenting possible software pipelines (e.g., MeTA) to build the search engine using different retrieval methods
● Generated testing queries
● Developed rules to judge relevance of text retrieval results of queries based on domain knowledge
● The team decided to self-generate input EMR data for the search engine with customized format

2. Remaining tasks

● Curation, clean and transform of the hospital summarizes data from MIMIC-IV into appropriate format/file.
● Manually annotate a collection of discharge summaries as testing data with domain knowledge
● Evaluation performance of the search engine based on testing data
● Refine search engine to improve performance. For example, try to fix the abbreviation problem.
● Source code and main results documenting and project demo preparation

- Self-evaluation performance again based on testing data. May identify defects and possible solutions for future improvement

3. Any challenges/issues being faced

Currently, the team is still waiting for the approval of access to the EMR database. The approval request processing time may take up to 4 weeks. Without the actual data, the task of cleaning and curation data cannot be started and no formatted data could be used as input source for our search engine. To avoid this blocker temporarily, the team decides to self-generate input EMR data using the customized format so that the rest of the tasks could be carried on as planned. The team will feed the search engine with the self-generated data and continue to work on the evaluating and refining tasks in such a way.