Apply state-of-the-art text retrieval methods on electronic medical records

Yeyu Ma, Yuanpei Ma, Hanyin Wang

October 23, 2022

Apply state-of-the-art text retrieval methods on electronic medical records

1. **What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.**
    - Hanyin Wang (hanyinw2)
    - Yuanpei Ma (yuanpei3)
    - Yeyu Ma (yeyuma2)

    Yuanpei will be the captain to lead the whole team and in charge of the administrative duties. The group name is “MMW”.

2. **What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?**
    - Apply state-of-the-art text retrieval methods on electronic medical records (EMR).
    - To be specific, our project aims to build a text retrieval engine on a database of hospital discharge summaries to identify patients treated with certain diagnoses during hospital stay. For example, we may wish to find all patients treated with antibiotics of ceftriaxone for pneumonia during hospital stay by searching “pneumonia ceftriaxone”.
    - Such a patient screening system based on text retrieval could have wide implications in the field of healthcare, for example identifying candidates for enrollment into clinical trials.
    - Our data will be based on MIMIC-IV (https://physionet.org/content/mimiciv/0.4/), a well-established real world medical database with rich content of text data (e.g., hospital discharge summaries).
    - We will utilize existing software pipelines (e.g., MeTA) to apply text retrieval methods (e.g., BM25) to build a search engine for the discharge summary database.
    - We will apply domain knowledge to refine and improve performance of the search engine. For example, to address numerous abbreviations in medical records.
    - We will measure the performance of our search engine (e.g., mean average accuracy). A collection of testing data will be annotated by one of our group members with domain knowledge.
    - As the outcome, we anticipate building a high-performing search engine to identify patients treated with certain diagnosis based on text query of discharge summaries. Searching results returned by the search engine should be within a reasonable time range and having relative relevance with the query keywords.

3. **Which programming language do you plan to use?**
    - Python

4. **Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.**

   Note: There are 3 students in the team, the total hours of workload is 20*3 (60 hours)

Task	Person in charge	Total hours
• Complete the required CITI research training to get access of MIMIC-IV data (https://physionet.org/content/mimiciv/view-required-trainings/0.4/#1)	Hanyin	5
• Curation, clean and transform of the hospital summarizes data from MIMIC-IV into appropriate format/file.	Hanyin	7
• Apply software pipelines (e.g., MeTA) to build the search engine using different retrieval methods (e.g., BM25).	Yuanpei	10
• Manually annotate a collection of discharge summaries as testing data with domain knowledge	Hanyin	8
• Evaluation performance of the search engine based on testing data	Yeyu	8
• Refine search engine to improve performance. For example, try to fix the abbreviation problem.	Yeyu	8
• Source code and main results documenting and project demo preparation	Yuanpei	6
• Self-evaluation performance again based on testing data. May identify defects and possible solutions for future improvement	Yeyu/Yuanpei	4*2