Part A

In this project, I mainly worked on the data crawling, article preprocessing, and backend search algorithm. In my initial assessment, I mentioned the required skills for my work were Python, SQL, Elasticsearch, and text-retrieval algorithms. All of these skills directly applied to my contribution in this project. I crawled data from an online source using Python, extracted information and stored it in a MySQL database, applied some NLP techniques to preprocess the articles, and created an Elasticsearch index for the application to apply the search algorithm.

From my project work, I learned more about some NLP techniques and tools to help improve the text-retrieval algorithm. One particular tool was MetamapLite which helped me generate medical concept terms that added more search-relevant information to the titles and abstracts of the articles. I successfully completed the data part and created a reusable data pipeline for the website to occasionally update the database. One of the big obstacles lies in the processing time of the Metamap and Elasticsearch caching. It took 2-3 weeks to completely process 300,000 articles in our database. That will be a challenge whenever we decide to update our database in the future. If we continue to work on this project after this semester ends, I plan to develop an incremental update procedure that only processes new articles every time we do an update.

Part B

Our group completed the fully-functional final product. Our final product met our initial goals, except for some undeveloped extension features like user interaction collection. Teamwork played an important part in our success. Khanh and I both contributed our specialties to accomplish the goals. We succeeded in designing a teamwork model that focused on separation of concerns. All of our functions and classes were organized in a way that we could develop them in separation without blocking each other. That way, we utilized our time optimally without waiting for the other to complete his part.

One thing not quite successful in teamwork was that the division of work was not too balanced. I felt that Khanh worked much more than I did this semester. The reason for that was that there were not too many things to work on my data part this semester, since all of the pipelines were developed last semester. My only work was to run the pipeline on the new dataset. While I did not need much effort, the process actually took very long to complete (2-3 weeks). Without carefully planning in advance, we ran the data update too late and that part became a development blockage that delayed our completion timeline. I just had one teammate, Khanh, and he completely deserves special recognition in this project. He worked hard on the frontend development, and the end product looked professional and user friendly thanks to his immense efforts in both semesters.