SI650 hw3

Yumin Zhang

My kaggle username: Aatroxz

Part2a

For Chatting Bot part, I use BM25 model to perform. I read in the *chatbot-replies.2.tsv* file as my database. In the preprocessing stage, I retrieve all the response message, and drop the duplicates and keep the unique value. Then the corpus is tokenized and input to the BM25 API class. Whenever a input message come, I will tokenize the input and make a query to find the top 1 related response from the corpus, which will be the response served to the user.

Part2b

I use *generateTable.py* file to preprocess and generate table *id\_text\_test.csv* , which keeps 2 columns, which are all the unique id in test table, as well as their corresponding message. In the main file, I read in the table, and build 2 dictionaries, allowing converting from id to text and vice visa, which transform the id for both message and response in test table to texts. Then, for each message, I select all the responses corresponding to it in the table as the database. Each message will make a query to their corresponding database by BM25 algorithms, and get top 10 responses as the best answers. If the database is smaller than 10, it will make a query to the whole response dataset to find others responses to fill up. In the end, the message with response table is built