Methodology

I used a.parquet file and decided to put the code to prepare data inside of the index.sh, because I was unable to succeed with your suggestion (for some reason nothing appeared in hadoop). I just do the same logic and put data in hadoop after python script that prepares data for hadoop.

In app.sh I initialize cassandra keyspace and tables inside that to store the data and call other scripts (prepare_data, index.sh, search.sh)

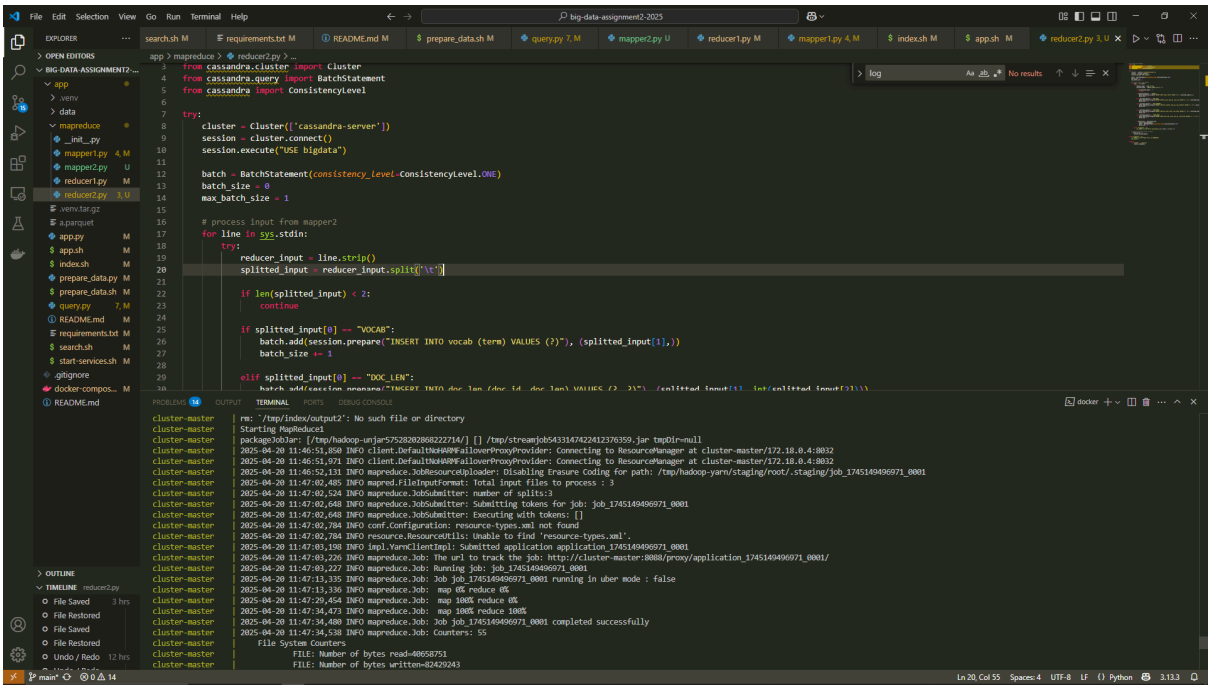
In the query.py I just read data from cassandra, calculate necessary for bm25 calculation values, call functions to calculate it and produce top documents.

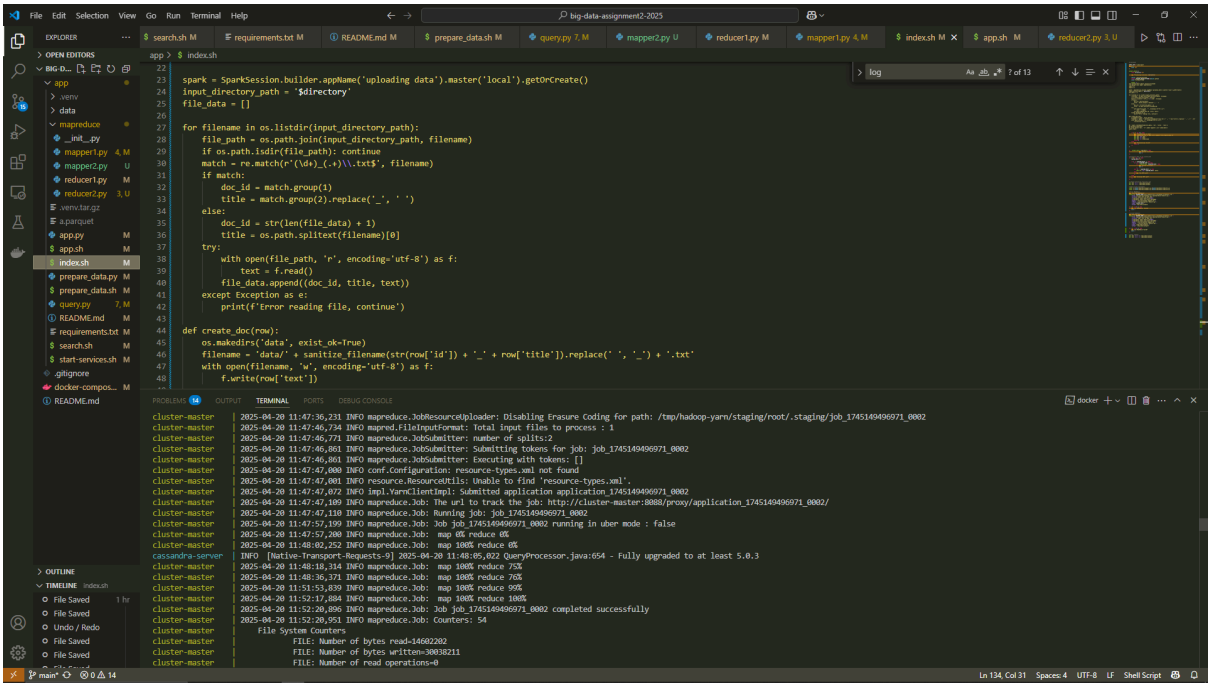
mapper and reducer were done using reference from lab5 and youtube videos. I pass output from mapper1 to reducer1 (structured info about doc len, term freq, and document content). In reducer 1 it's combined together and passed further. In reducer2 I used batches and process records to avoid overloading of the system and be sure that the system will work properly.

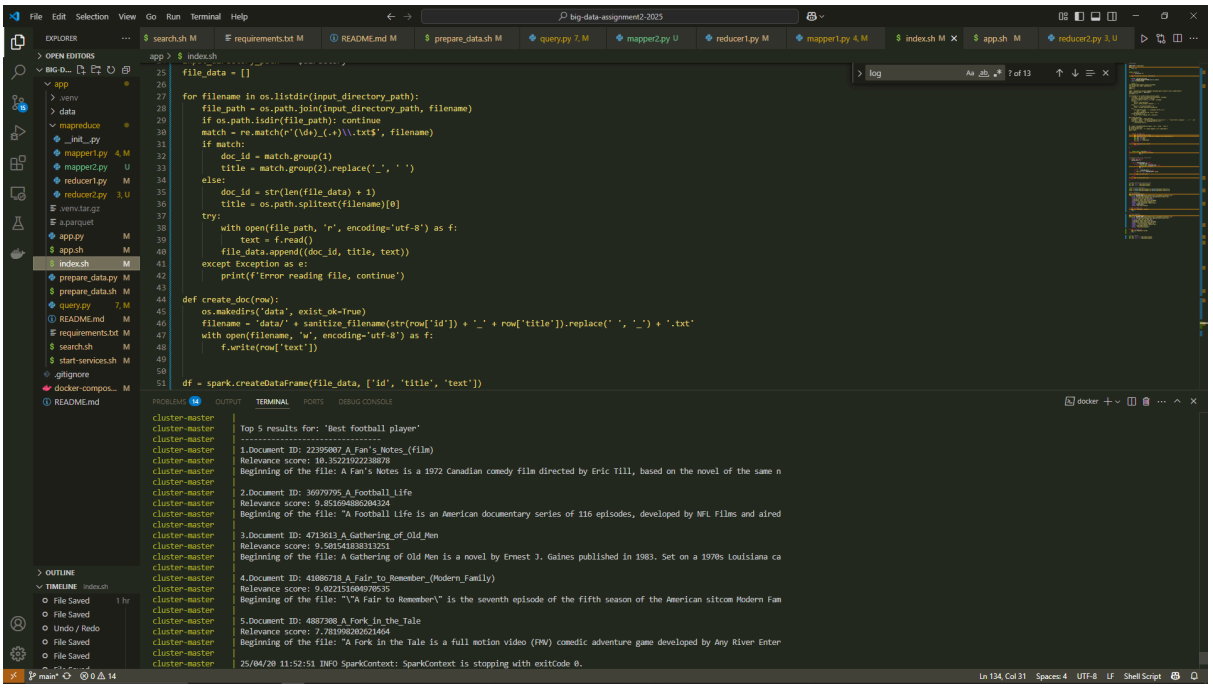
Also, I modify docker-compose file and provide port 9042 to connect to it

Demonstration

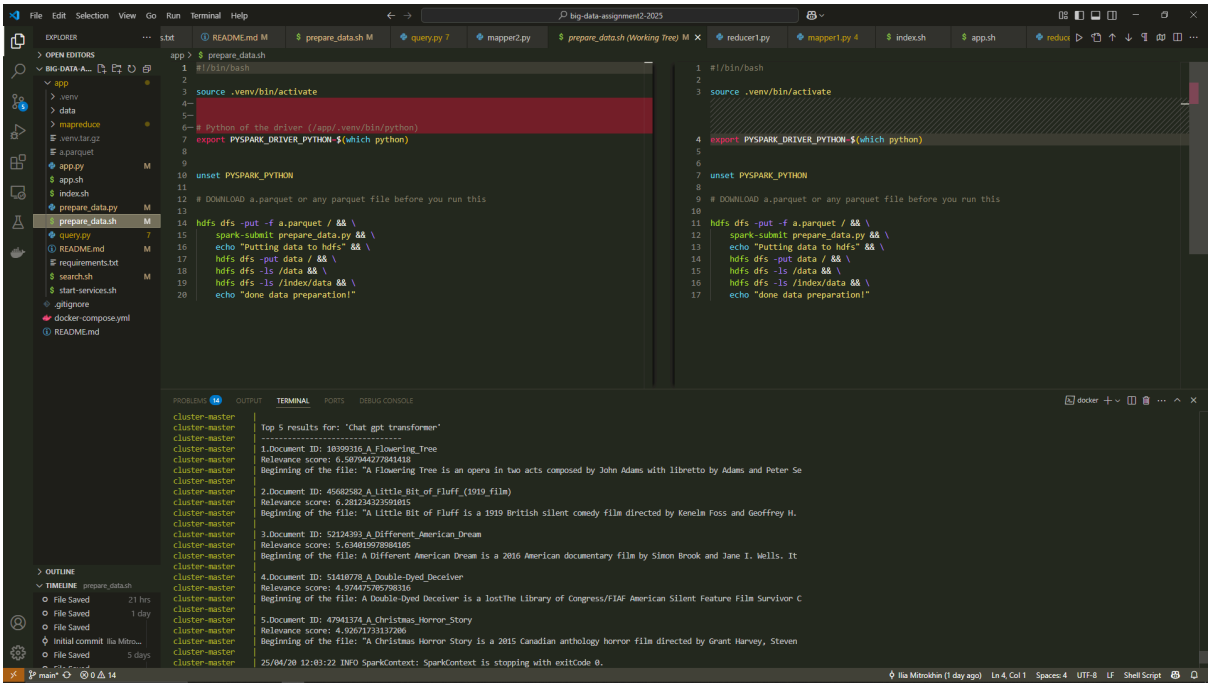
to run the system just run docker compose up –build
1st query: best football player







2nd query: Chat gpt transformer



We can see that it is the case that nothing related to this is found in documents. but. at least we obtain the output somehow related to the query

3rd query: Interesting producer movie