IR2021, Assignment 2 - Group 29

Group Members:

Akshit Pal Singh -2017015
Rudraroop Ray - 2017311
Rahul Patwardhan - 2016075
Divyam Anshumaan - 2017147

To run the program -

1. To run Q1, run positional_index.py
2. To run Q2 a) (Jaccard index), run jaccard.py
3. To run Q2 b), c), run q2_b,_c_queries.py
4. To run Q3, run ir_ass_2_q3.py

Files and directories in the folder-

processed - folder containing the processed dataset (467 files)

file_map.pkl - maps files to doc IDs

Files for q1 -

1. Positional_index.py (main file)
2. Pos_index_final.pkl (pickle file containing index)

Files for q2 -

1. Jaccard.py (file used to run queries for part a)
2. Q2_b,_c_queries.py (file used to run queries for part b and c)
3. Q2_td_idf_matrices.py (file used to create TF-IDFf matrices)
4. Word_map.pkl (maps words to numerical IDs)
5. Inverse_word_map.pkl (inverse map of num. IDs to words)

The files for the TF-IDF matrices were too big to upload on Github, so they have been uploaded to the following drive folder. -
https://drive.google.com/folderview?id=1-Mz5PzQcLbFxOXk8yGt1wttrBYmxmtLD

Files for q3 -

1. Ir_ass_2_q3.py (main file for q3)
2. Ir_ass_2_q3.ipynb (Jupyter notebook for the same file)
3. Qid_4_max_dcg.txt - (output file for q3)