TEXT PROCESSING AND RETRIEVAL

(CS5700)

EXECUTIVE M. TECH. IN DATA SCIENCES

IIT HYDERABAD

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

MAUNENDRA SANKAR DESARKAR

Text Processing and Retrieval

- The datasets for this assignment come from StackExchange, a community question answering (CQA) forum
- We have shared two datasets with the assignment
- For each of the datasets, each row corresponds to a stackexchange post.

Task 1

- Generate statistics of each dataset – about the documents and words in it
- You can put the information in graphs/tables

Task-2

- Create an application that can take a post as input, and output top posts that are similar to this post.
- Prepare a block diagram of your approach
- For each block, mention the solutions that you have considered
- Briefly outline your approach. Be precise and avoid unnecessary verbosity.

Task-2

- Do the following for each dataset:
  - For two sample posts for which you think the designed system is able to find good matches, show the top-5 posts output by your algorithm
  - For two sample posts from a dataset for which you think the designed system is not able to find good matches, show the top-5 posts output by your algorithm

Tasks

- Do you find any drawback/shortcomings of the used approach?
- If yes, how do you think the solution can be improved?
- If you have any insights to share, please include them in the report.

Deliverables

- A report. The report should contain your tables, graphs and your observations from the tables and graphs. If you want to include any interesting point/observation that you may have made by looking at the results, feel free to include that. Make sure to mark your name and roll number in the report.
- Zip together the code files and the report. Name the zip file as <your-roll-no>_A2.zip.
- Department plagiarism policy: https://cse.iith.ac.in/?q=node/254