Part A - Task 1

Building an inverted index and Boolean retrieval

Group No.- 14

Group Members	**Roll No.**
Ajay Kumar Meena	18CS30006
Kushal Natani	18CS30025
Yash Fulzele	18CS10058
Mudit Gupta	18CY20019

TASK 1A

In this part the below steps were followed to make the inverted index:

- The input is taken from the folder “en_BDNews24”.
- From each document i.e. a file from the folder, the text was extracted using normal string operations, and the doc-id is kept as a “<DOCNO>” tag in it.
- The text preprocessing was also done to remove any unwanted characters and lemmatization used as mentioned in the task.
- Finally, all this is stored in an invertedIndex dictionary in which key is term and value is a list of (docno,frequency) pairs and it is saved in a pickled-file.

TASK 1B

In building the parser file the below steps were followed:

- The file “raw_query.txt” was parsed using the XML module in python which generates a tree-like structure for tags and we can access each tag independently.
- The required text fields were extracted using the appropriate tags and then done preprocessing over them as well.
- Lemmatization also performed over all the texts.
- The final results are dumped into a text file “queries_14.txt” which contains query-id and associated tokens.

TASK 1C

In building the results the below steps were followed:

- The input is taken as the results from the above two sections.
- The lists corresponding to a particular query is extracted from the inverted index, we build in the first part.
- The lists are sorted with respect to their lengths and converted to sets.
- The intersection is taken later for the generated sets.
- The results are stored in a text file as {query id : [final doc ids]} for every query.

Python version: Python 3.8.10

Libraries Required:
1. nltk: pip3 install nltk
2. xml module of python was used for parsing the “raw_query.txt” file.

Running Steps:
$>> python3 PAT1_14_indexer.py <path_to_the_en_BDNews24 folder>
$>> python3 PAT1_14_parser.py <path_to_the_query_file>
$>> python3 PAT1_14_bool.py <path_to_model_queries_14.pth> <path_to_query14.txt>