Group number 16
Group Members:
1) Rahul Aditya (18CS30032)
2) Aryaman Jain (18CS30007)
3) Maitrey Govind Ranade (18CS30026)
4) Abhishek Srivastava (18CS10068)

Task 1A - Building the index

Algo :

- We find all the non-empty directories.
- We crawl through them and open each file one by one.
- We read the text and obtain the text under ‘DOCNO’ tag and ‘TEXT’ tag using string searching in python.
- We process the text under the ‘TEXT’ tag as follows:
    - Convert to lower case.
    - Remove punctuations using regex.
    - Break the string in word tokens.
    - Remove stopwords using nltk.
    - Lemmatize the tokens using nltk.
- For each token in a document we added the document in the postings list of the token along with its frequency.
- We sorted each postings list to have proper ordering.
- We saved the inverted index using pickle.

Assumptions :

- We saved doc id and frequency of a token in the doc. This will be useful in part 2 when we try to create the term frequency vector of each document. We won’t have to parse all the documents again.

Libraries and versions:

- Python 3.9.7
- We used nltk, os, pickle, re, sys, and Counter (from collections)

Task 1B - Query Preprocessing

Algo:

- Opened “raw_query.txt”.
- Kept on reading it until we got ‘num’ and ‘title’ tags. Collected the text under these tags.
- Preprocessed the text under the title tag as we did in the previous task.
- Saved query num and query text tokens in the output file.

Assumptions:

- Saved the preprocessed queries using json for easier and faster parsing later on.

Libraries and versions:

- Python 3.9.7
- We used nltk, re, json, sys.

Task 1C - Boolean Retrieval

Algo:

- Loaded the saved inverted index and preprocessed queries.
- For each query, obtained the postings list for each term and kept on finding the intersection in linear time using a simple list merging algorithm taught in class.
- Save the resultant list after intersection in the output file.

Assumption:

- If an OOV term is encountered, the result of the query will become an empty list. This is because that term has no relevant documents, and hence AND query should give an empty doc list.

Libraries and version:

- Python 3.9.7
- We used pickle, json and sys.