Group No: 10

Members:

- 18CS10037 Nikhil Popli
- 18CS10048 Sahil Jindal
- 18CS10052 Somay Chopra
- 18CS30009 Ashutosh Varshney

Task-1A

- First read all the files in en_BDNews24.
- We maintain a dict inverted_index to get the posting lists. For each file:
    - First we got the string between <TEXT> </TEXT> using a regex.
    - Then replaced all the punctuation marks by a “ “ using regex.
    - From the resulting string, we got the tokens using word_tokenize() and converted to lowercase. Removed stop words from these tokens. These tokens are then lemmatized using WordNetLemmatizer
    - For the resulting tokens, we update the inverted_index. If the token is already there we append the (doc_id, token frequency) otherwise we add the token to the dict and add (doc_id,token frequency)
- We sort the resulting inverted_index to make the posting list sorted.
- Finally saved the sorted inverted_index using pickle

```
CMD: python PAT1_10_indexer.py [path to en_BDNews24]
Eg: python PAT1_10_indexer.py /home/sahil/IR/Data/en_BDNews2
```

Task-1B

- Read the title and number using regex.
- Removed the punctuations, got the tokens, removed stopwords and lemmatized final tokens similar to Task1A.
- Saved the queries file using pickle

```
CMD: python PAT1_10_parser.py [path to raw_query.txt]
Eg: python PAT1_10_indexer.py /home/sahil/IR/Data/raw_query.txt
```

Task-1C

- For each modified query, scanned it’s tokens. Extracted the posting list of that token from model_queries.pth
- Merged the posting lists using the method discussed in class by taking the smaller two lists and merging them and then repeating until exhausted
- Stored the final merged posting list as the result

```
CMD: python PAT1_10_bool.py [path to model_queries_10] [path to queries_10]
Eg: python PAT1_10_bool.py /home/sahil/IR/Data/model_queries_10.pth
/home/sahil/IR/Data/queries_10
```