

Part 1:

To find the invertedIndex, run the following command on the command line

```
python indexer.py tccorpus.txt index.out
```

Here,

indexer.py is the python script for finding inverted index

tccorpus.txt is the corpus

index.out file where the output is written

Output format

The output in index.out is in the following format:

```
# word1
docId count docId count .....
# word2
docId count docId count .....
```

The output in tokens.txt is in the format

```
docID noOfWordsinDocument
docID noOfWordsinDocument
.
.
.
```

Part 2:

To find the BM25 Score, run the following command on the command line

```
python bm25.py index.out queries.txt 100 results.eval
```

Here,

bm25.py is the python script for finding bm25 score

index.out is the output from part 1

queries.txt are the queries

100 is the maximum number of documents fetched for a query

results.eval output for part 2 in the following format

The output in results.eval is in the following format:

```
query_id Q0 doc_id rank BM25_score system_name
```