Information Retrieval: Text Processing

Group 13

Sequence of operations

1. Install Requirements
2. Add given wikipedia files to the corpus directory.
3. Download glove.6B.100d.txt dataset and place it in the project root directory.
4. Run construct_index.py
5. Run construct_index.py --zoned_index True
6. Run trim_embeddings.py
7. Run test_queries.py
8. Run test_queries.py --score_title True
9. Run test_queries.py --expand_query True

Installing Requirements:

```
pip install -r requirements.txt
```

corpus

Contains the files to be indexed. Add files directly to this directory. Do not create subdirectories.
For this assignment, we have used the following files present in the AA folder of wikipedia files.
wiki_00
wiki_01
wiki_05
wiki_06
wiki_10
wiki_11
wiki_15
wiki_16
wiki_20
wiki_21
wiki_25
wiki_26
wiki_30
wiki_31

index_files

Contains the inverted indices constructed using construct_index.py.

construct_index.py

Constructs the inverted indices used for query evaluation.
Command Line Arguments:
--zoned_index: True if zoned indexing must be used. Set to False by default.

trim_embeddings.py

Trims the GloVe embeddings to contain terms only from corpus. Download the glove.6B.100d.txt dataset before running this file.

test_queries.py

Evaluates queries and displays retrieved documents.
Command Line Arguments:
--score_title: True if zoned index considered for evaluation. Set to False by default.
--expand_query: True if query expansion must be used. Set to False by default.

helper_module.py

Contains helper functions used by other files. **Do not run this file**.

document_list.txt

Contains the document ids and names used for evaluation.