Assignment 2

Design

The main idea is that `DocIndex.py` is a module that handles anything modifying the index directly. `test.py` is just an interface to interact with `DocIndex` module. The main way for handling the index is to use the `*-index.txt` files as a pseudo json database.

High Level Overview of Implmentation

```
Queries are loaded.
Output file results_file.txt cleared/made
Index cleared using DocIndex.py
Trec fil es added to index using DocIndex.py

For each query in queries
    For each term in query
        find term using DocIndex.py
        save findings
    perform calculations for TF/IDF/Cosine Similarity based on findings
    rank findings
    print findings
    clear variables to prepare for next loop
```

The core loop logic shown above can be found in the function `iterateQueries` in `trecTest.py`

Requirements

- Python Version 3.7.2 used to develop

Basic Usage

To run with respect the assignment's requirements,

```
python3 src/trecTest.py
```

If needed, there is the ability to start with fresh psuedo-database files.

```
python3 ./DocIndex.py --clear
```

Or simply delete the `./output` directory.

Perl Script

```
    ./data/trec_eval.pl -q ./data/qrels.txt ./output/results_file.txt >
output.txt
```

DocIndex only options

All DocIndex options

`--clear`: clears the index files

`--find <arg>`: finds the `<arg>` in the index and prints out if found. No output if not found.

`--dir <directory>`: searches the `<directroy>` and adds any `*.txt` file to the index

`-f <filename>, --file <filename>`: adds `<filename>` to the index; must be a .txt file.

`--trec <filename>`: adds the specified trec file to the index

Files

Generated

All Files listed below are generated in a `./output` directory. `./output` is made if it does not already exist.

- document-index.txt
- document-index-backup.txt
- posting-index.txt
- posting-index-backup.txt
- results_file.txt

Source

- DocIndex.py
- trecTest.py