IR : Information Retreival System (Text Processing)

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Group 5

Installation and Run IR_TP

1. Extract the files

```
$ tar -xvf Group_5.tar.gz
```

*** You now have the following files:**

```
$ ls
```

```
indexing.py     requirements.txt     test_queries.py     Group_5_Assignment_Report.pdf
readme.pdf   readme.md      AA
```

2. Install the dependencies and devDependencies and start the server.

```
$ python3 -m venv envIR
```
```
$ source envIR/bin/activate
```
```
$ pip install -r requirements.txt
```

3. Run indexing.py

```
$ python indexing.py
```

> *Prerequisite: pip install nltk pip install bs4 pip install numpy pip install spacy*

```
Input file:  AA/wiki_00
Output folder:  ./indexFiles
...created and Stored inverted_index_dict.json ...
...created and Stored freq_list.json ...
...created and Stored title_list_file.json ...
Completed!
```

OR OPTIONAL:

provide input as cmd line args for:

- filepath of wiki_00 from current project folder (default: AA/wiki_00)

- the folder name which will store built index (default: indexFiles) following is optional : `$ python indexing.py AA/wiki_00 indexFiles`

  ```
  Input file:  AA/wiki_00
  Output folder:  ./indexFiles
  ...created and Stored inverted_index_dict.json ...
  ...created and Stored freq_list.json ...
  ...created and Stored title_list_file.json ...
  Completed!
  ```

  #### Extra folder is created (indexFiles) which houses the following index files:

  `$ ls indexFiles/`

  ```
  freq_list.json      inverted_index_dict.json      relatedWords.pickle
  title_list_file.json
  ```

  ## 4. Run test_queries.py

  > *Prerequisite: for using spacy vectors(GloVe) we need to have the corpus. There are two choices, we are currently using the better corpus of around 900 MBs Download it using `python -m spacy download en_core_web_lg` The smaller corpus can be downloaded using: [But need to comment line 258 and uncomment 259 in test_queries.py] `python -m spacy download en_core_web_sm`*

`$ python test_queries.py`

```
<Enter folder storing the index files (ex- indexFiles)>:
indexFiles
<Enter your query:>
Main cause of poverty
<Enter Option:-
    1:Normal Part1 retreival,
    2:Improvement1,
    3:Improvement2,
    4:All three,
    5:All three but Lengthy
    0:exit>
1
```

This will output-

The top 10 retreived documents corresponding to options 1,2 or 3 Options 4 and 5 are fancy to print all at once. Option 0

```
--------------------------------------------------
Query Terms:  Counter({'main': 1, 'cause': 1, 'of': 1, 'poverty': 1})

PART1: The top 10 documents matching with the query ' Main cause of poverty ' are:

0. DocumentID: 322  , Score: 0.064, Title: Extreme poverty
```

```
1. DocumentID: 127  , Score: 0.05 , Title: Aegean Sea
2. DocumentID: 55   , Score: 0.049, Title: A Modest Proposal
3. DocumentID: 254  , Score: 0.046, Title: Economy of American Samoa
4. DocumentID: 60   , Score: 0.039, Title: Affirming the consequent
5. DocumentID: 439  , Score: 0.037, Title: Albert, Duke of Prussia
6. DocumentID: 146  , Score: 0.035, Title: Motor neuron disease
7. DocumentID: 224  , Score: 0.035, Title: Abscess
8. DocumentID: 97   , Score: 0.034, Title: Abortion
9. DocumentID: 247  , Score: 0.034, Title: Economy of Armenia
--------------------------------------------------


Time Taken= 0.9344477653503418 seconds
<Enter your query:>
```

- **And waits for another query input**
- **Option 0 will exit you**

6. Deactivate virtual env

```
$ deactivate
```