Information Retrieval Project

Team Members:

1. K.Sai Sri Thanya (S20180010083)
2. E.Sumasree (S20180010052)

Project idea : IR System for News Articles

Tasks Performed:

1. Crawling
2. Scraping
3. Creating Inverted Index
4. Scoring and Ranking

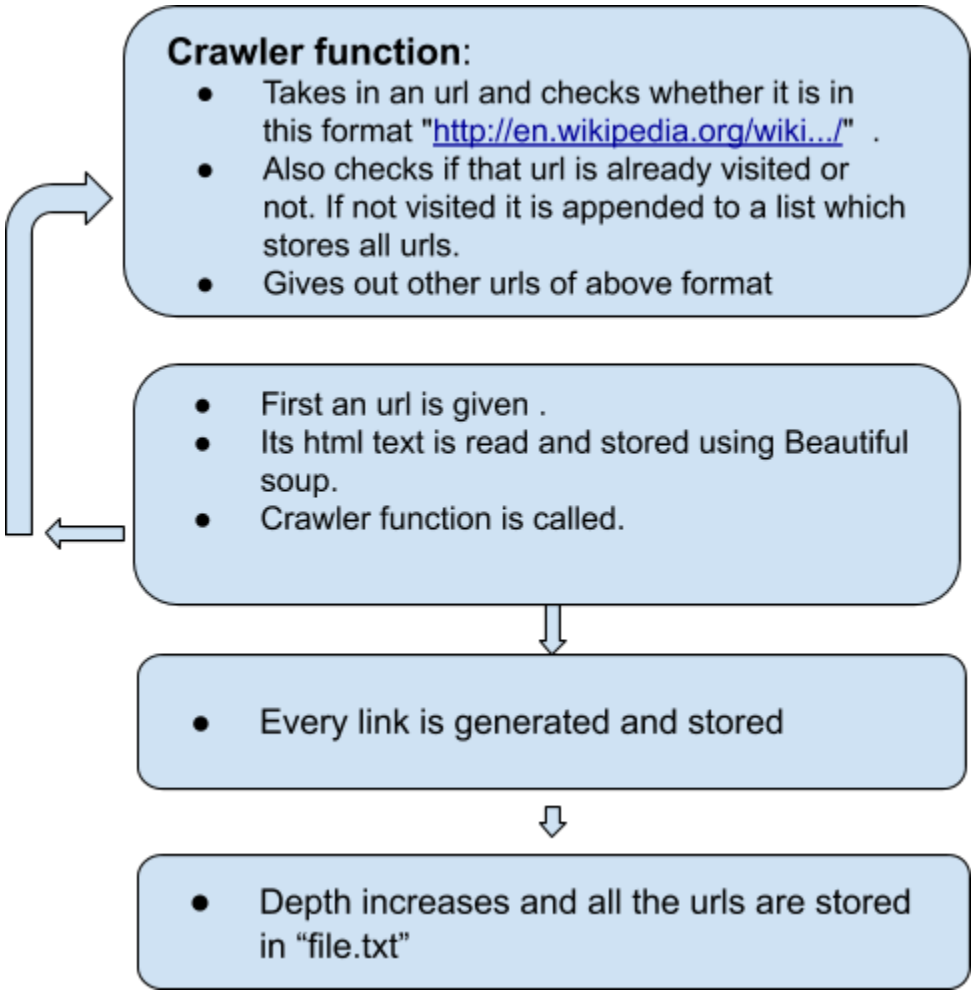
Detailed View:

Crawling :

Libraries used:

1. urllib : To request urls and also parse them
2. From bs4 : BeautifulSoup : To read html files
3. Time
4. re : for comparing regular expressions
5. Sys

Algorithm or Code Flow:

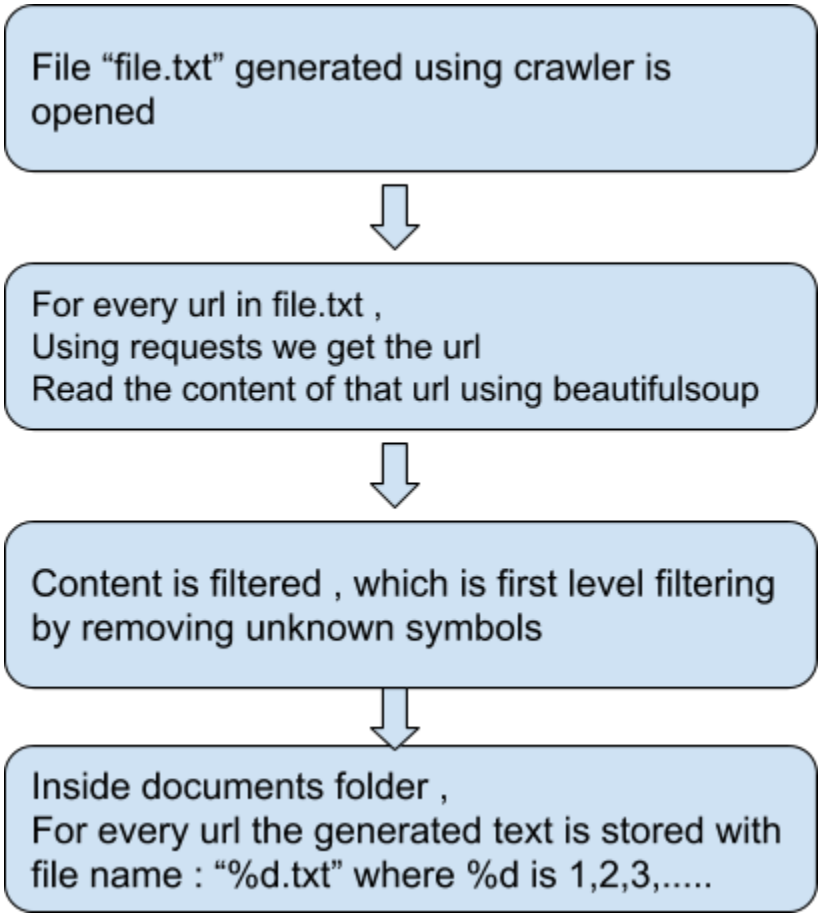


Scraping:

Libraries used:

1. Requests
2. From bs4 : Beautiful Soup: To read html files

Algorithm or Code Flow:

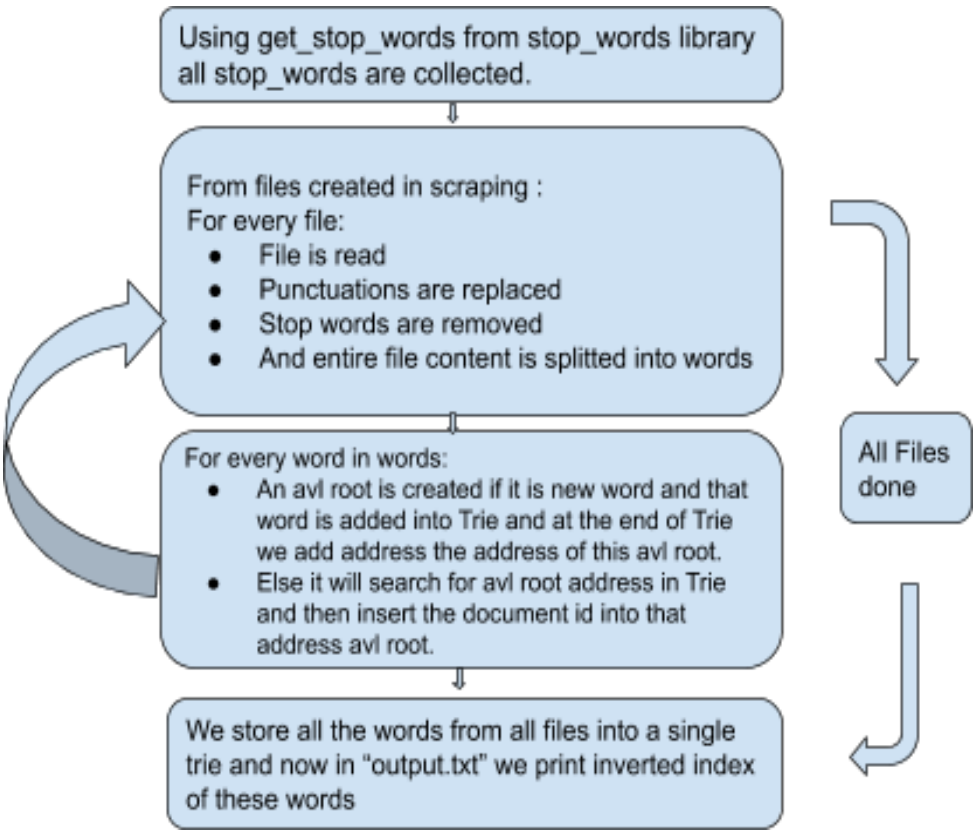


Inverted Index :

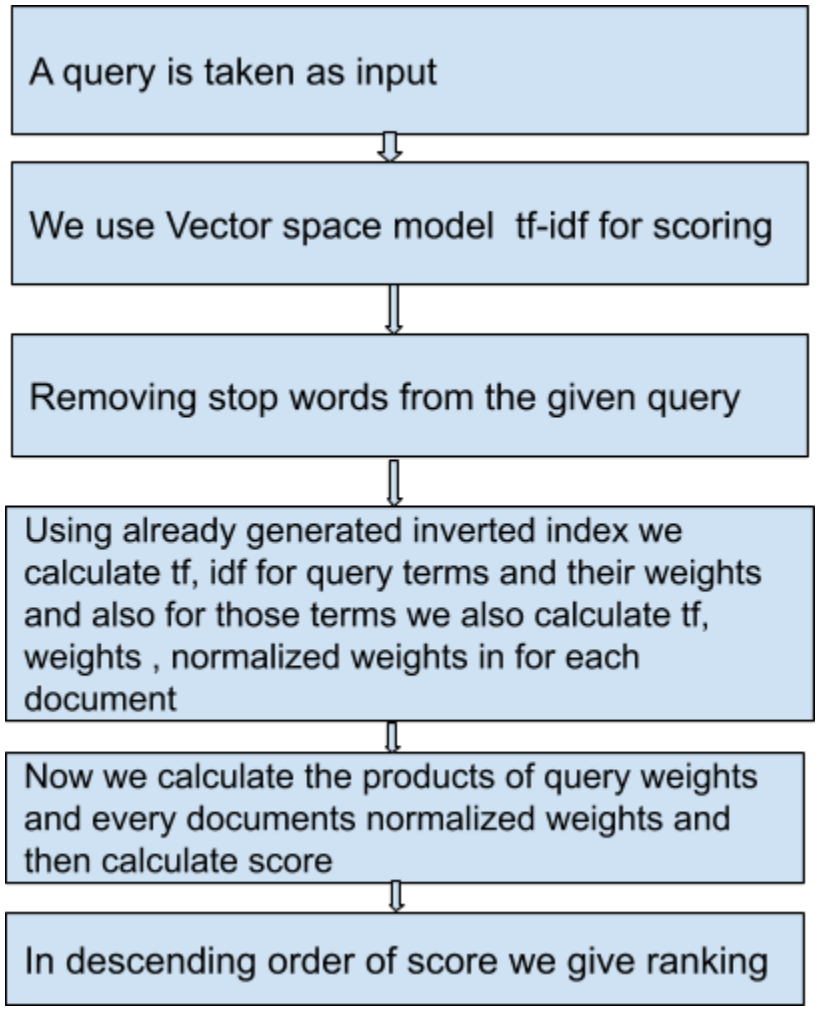
Libraries used:

1. Glob : The `glob` module finds all the pathnames matching a specified pattern according to the rules
2. String
3. From collections : defaultdict : Defaultdict is a sub-class of the dict class that returns a dictionary-like object.
4. Math : To use functions like square root , log etc.

Algorithm or code flow:



Scoring and Ranking :



Output :

```
sree@SUMASREE:~/Education/ir$ gedit index_rank.py
sree@SUMASREE:~/Education/ir$ python3 index_rank.py
enter query:
transport developement meaning made record
303
281
123
256
56
122
6
75
285
76
sree@SUMASREE:~/Education/ir$
```