AUTHORS:

Aliya Valieva
Haoyuan Bai

LANGUAGE:

Python 3.9

DESCRIPTION:

The project consists of two parts, Indexer.py and SearchEngine.py.

The Indexer.py includes lemmatizing words, calculating the TF-IDF score and handling HTML parsers. SearchEngine.py is the main module to run, it connects to the querying index database and retrieve 20 URLS for each query

Collecting statistics through tools including sql and sqlite.

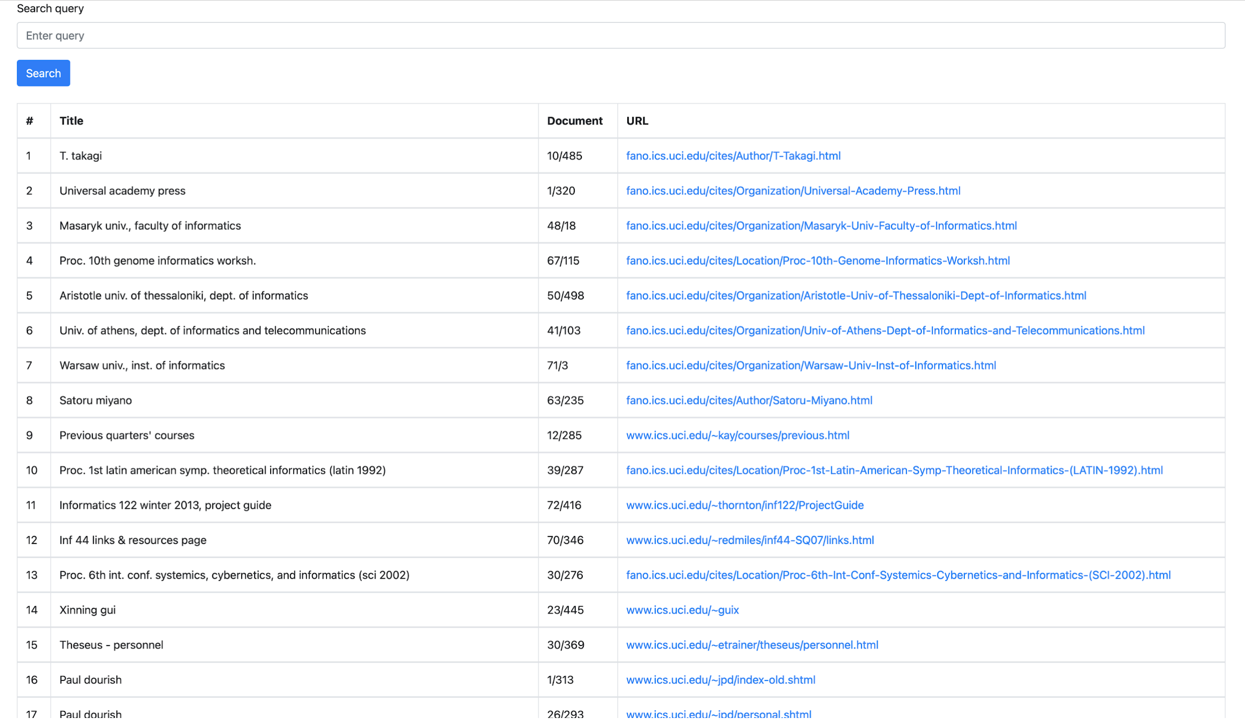
Libraries:

```
pip install spacy, parser, bs4, lemmatizer, Flask, sqlite3
python -m spacy download en_core_web_sm
```

Please download DB Browser for SQLite to have a better user friendly experience for viewing DOCINDEX.db

Running on: http://127.0.0.1:5000/

(sample interface with 20 fetched URLS)



#	Title	Document	URL
1	T. takagi	10/485	fano.ics.uci.edu/cites/Author/T-Takagi.html
2	Universal academy press	1/320	fano.ics.uci.edu/cites/Organization/Universal-Academy-Press.html
3	Masaryk univ., faculty of informatics	48/18	fano.ics.uci.edu/cites/Organization/Masaryk-Univ-Faculty-of-Informatics.html
4	Proc. 10th genome informatics worksh.	67/115	fano.ics.uci.edu/cites/Location/Proc-10th-Genome-Informatics-Worksh.html
5	Aristotle univ. of thessaloniki, dept. of informatics	50/498	fano.ics.uci.edu/cites/Organization/Aristotle-Univ-of-Thessaloniki-Dept-of-Informatics.html
6	Univ. of athens, dept. of informatics and telecommunications	41/103	fano.ics.uci.edu/cites/Organization/Univ-of-Athens-Dept-of-Informatics-and-Telecommunications.html
7	Warsaw univ., inst. of informatics	71/3	fano.ics.uci.edu/cites/Organization/Warsaw-Univ-Inst-of-Informatics.html
8	Satoru miyano	63/235	fano.ics.uci.edu/cites/Author/Satoru-Miyano.html
9	Previous quarters' courses	12/285	www.ics.uci.edu/~kay/courses/previous.html
10	Proc. 1st latin american symp. theoretical informatics (latin 1992)	39/287	fano.ics.uci.edu/cites/Location/Proc-1st-Latin-American-Symp-Theoretical-Informatics-(LATIN-1992).html
11	Informatics 122 winter 2013, project guide	72/416	www.ics.uci.edu/~thornton/inf122/ProjectGuide
12	Inf 44 links & resources page	70/346	www.ics.uci.edu/~redmiles/inf44-SQ07/links.html
13	Proc. 6th int. conf. systemics, cybernetics, and informatics (sci 2002)	30/276	fano.ics.uci.edu/cites/Location/Proc-6th-Int-Conf-Systemics-Cybernetics-and-Informatics-(SCI-2002).html
14	Xinning gui	23/445	www.ics.uci.edu/~guix
15	Theseus - personnel	30/369	www.ics.uci.edu/~etrainer/theseus/personnel.html
16	Paul dourish	1/313	www.ics.uci.edu/~jpd/index-old.shtml
17	Paul dourish	26/293	www.ics.uci.edu/~jpd/personal.shtml