Title: COSI 132a Assignment 5 Elasticsearch

Author: Yulun Wu yulunwu@brandeis.edu
Date: Apri. 13th 2019

Description:

Implemented a information retrieval system using Elasticsearch and apply it to your wikipedia movie corpus.

Dependencies:

Python 3.7.2

In Addition to Packages required in Assignment

Elasticsearch 6.3.1

http://www.nltk.org/install.html

Elasticsearch DSL 6.3.1

https://elasticsearch-dsl.readthedocs.io/en/latest/

Flask 1.0.2

http://flask.pocoo.org/

JDK 8

https://www.java.com/en/download/manual.jsp

Virtualenv 16.3 (not required)

https://virtualenv.pypa.io/en/latest/

Running time

building index in about 6s on Windows 10 x64 1903 build 18362.30

Building Instructions:

(Optional) First run the format_corpus.py to pre-process the film corpus if you have not done that

```
python3 format_corpus.py
```

Go to the Elasticsearch directory and run the Elasticsearch service
(on windows run)

\bin\elasticsearch.bat

then run the vs_index.py to construct the index

```
python3 index.py
```

Run Instructions:

To start the Flask service, run

```
python3 query.py
```

and access the web interface at 127.0.0.1:5000

Modules:

|- index.py
| (used to construct the document index for the movie data json)
|- qeury.py
| (used to start the flask web service and display the query webpage)
|- format_corpus.py
| (pre-processing tool for the corpus to fix unusable string/data field)
|- films_corpus.json
| (the films corpus used)
|- test_corpus.json
| (the minimizied corpus for test purpose, containing 20 entries)
|--- templates
    | the directory containing web page templates
    |--- page_query.html
        the main query page, showing the query results, modified from the original
    |--- page_SERP.html
        showing the query results, modified from the original version
    |--- page_targetArticle.html
        the page showing single film information

Testing:

Example query for testing:

Indian (valid token)
2 (valid token)
Indian 2018 (valid conjunctive query)
indian hadeel (valid disjunctive query)
“Indian film” (explicit query)
“Indian film” x (explicit query with non-matching term)
“Indian film” “eee xsx” (explicit query with non-matching term)
meiyou (invalid token)