Documentation of the COMP_9900 website flask and jinjia2 part

Tree structure of the project can be find at the end of this file.

run.py

This script is the entrance of the whole project. In this python file, we set the listening IP is 0.0.0.0 in google cloud and listening port is 8080.

Moreover, in line 7, we set the python hash seed to 0. So that we can get the same doc vector and lsh value every time.

testing (folder)

Keeps all test scripts and some result files generated by the scripts.

user_log(folder)

Contains 4 user logs.

webapp

fast_query

This folder contains all fast_query part. You can check the details in ../webapp/fast_query/lsh.html

Static

This folder keeps all necessary data/ models/ articles.

The css, fonts, images and js folder contains the materials which are needed in the frontend part.

The models folder contains NSWSC folder and all models

NSWSC contains all legal articles in html format.

Others are models which will be used in vectorisation and fast query part. We read and store them in the ram, so that we can give the result of each time query fast enough.

Templates

This folder contains all html files of the search engine.

vectorisation

This folder contains all python script about the doc2vec part. You can find more details in the folder.

__init__.py

Initialize the flask app. It is used in run.py.

config.py

This file stores all necessary directories. All directories can be used after import this file.

load_models.py

This script is used to load all models before the web server runs.

So that all models can be loaded once at the beginning instead of loaded every time. This can short the search time evidently.

search_function.py

There is only a function called get_results(input_string, model, dataset) in the script.

The function has 3 input variables:

input_string contains the query string which gets it from the frontend part.

model is used in vectorisation part.

Dataset is used in fast query part.

After using this function and input all 3 variables, it can return a list of results.

Each element in this list is a dictionary which contains the file name, article title, rank number, and similarity value.

views.py

This is the main part of the website part. It is used to generate different html files to users according to their actions and past the query string or query file to vectorisation and fast query part.

There are many functions in it:

def allowed_file (filename):

This function can check if the upload file is txt or html

def index():

it can give user welcome page of the website if user first time enters the website or click relative buttons.

def results():

it can catch the search request and query string from the website and past the query string to get_results(input_string, model, dataset). After get_results() function return the result list, give the result list to results.html

def article_details():

This function will run if user clicks the link to an article. It will read the article filename and return the article to the user.

def upload():

This function can let user upload file and give the search result to the user based on the file content. When a user tries to upload a file. This function will try to get the name of the upload file, if the format of the upload file is not *.txt or *.html, nothing happened. If the format is txt or html, it will read the content in the file and give this information to get_results(input_string, model, dataset) function. After this function gets the search result, it will give the result to results.html.

```
|-- README
|-- run.py
|-- testing
|   |-- change_files.py
|   |-- evn.txt
|   |-- lsh_result.csv
|   |-- lsh_results.py
|   |-- lsh_results_summary.txt
|   |-- lsh_test.py
|   |-- test_count_results.py
|   |-- test_rank2.py
|   |-- test_result.csv
|   |-- tree.py
|-- user_logs
|   |-- z3413954.txt
|   |-- z5042181.txt
|   |-- z5083021.txt
|   |-- z5124787.txt
|-- webapp
|   |-- config.py
|   |-- load_models.py
|   |-- search_function.py
|   |-- views.py
|   |-- __init__.py
|   |-- fast_query
|   |   |-- convert.py
|   |   |-- lsh.bak
|   |   |-- lsh.html
|   |   |-- lsh.py
|   |   |-- lsh_script.py
|   |   |-- README.md
|   |   |-- test.py
|   |   |-- __pycache__
|   |   |   |-- lsh.cpython-36.pyc
|   |-- static
|   |   |-- css
|   |   |   |-- bootstrap-theme.css
|   |   |   |-- bootstrap-theme.css.map
|   |   |   |-- bootstrap-theme.min.css
|   |   |   |-- bootstrap-theme.min.css.map
|   |   |   |-- bootstrap.css
|   |   |   |-- bootstrap.css.map
|   |   |   |-- bootstrap.min.css
|   |   |   |-- bootstrap.min.css.map
|   |   |   |-- style.css
|   |   |-- fonts
|   |   |   |-- glyphicons-halflings-regular.eot
|   |   |   |-- glyphicons-halflings-regular.svg
|   |   |   |-- glyphicons-halflings-regular.ttf
|   |   |   |-- glyphicons-halflings-regular.woff
|   |   |   |-- glyphicons-halflings-regular.woff2
|   |   |-- images
|   |   |   |-- background.png
|   |   |   |-- lg_logo.jpg
|   |   |   |-- Logo.png
|   |   |   |-- sm_logo.jpg
|   |   |   |-- test.gif
|   |   |-- js
|   |   |   |-- bootstrap.js
|   |   |   |-- bootstrap.min.js
|   |   |   |-- npm.js
|   |   |-- models
|   |   |   |-- count_files.py
|   |   |   |-- dictionary.pkl
|   |   |   |-- doc2vec.model
|   |   |   |-- doc2vec.model.docvecs.vectors_docs.npy
|   |   |   |-- doc2vec.model.trainables.syn1neg.npy
|   |   |   |-- doc2vec.model.wv.vectors.npy
|   |   |   |-- filename_to_casename.pkl
|   |   |   |-- reverse_dictionary.pkl
|   |   |   |-- similarity_matrix.npy
|   |   |   |-- NSWSC
|   |   |   |   |-- articles.html
|   |   |-- upload
|   |   |   |-- NSWSC_1993_1.txt
|   |-- templates
|   |   |-- article_details.html
|   |   |-- results.html
|   |   |-- upload.html
|   |   |-- welcome.html
|   |   |-- common
|   |   |   |-- base.html
|   |   |   |-- footer.html
|   |   |   |-- header.html
|   |-- vectorisation
|   |   |-- doc2vec.py
|   |   |-- DocIterator.py
|   |   |-- html_scraper.py
|   |   |-- Instructions for sheng.txt
|   |   |-- NameEntity.py
|   |   |-- parsers.py
|   |   |-- preprocess.py
|   |   |-- query.py
|   |   |-- README.md
|   |   |-- scraper.py
|   |   |-- test.py
```