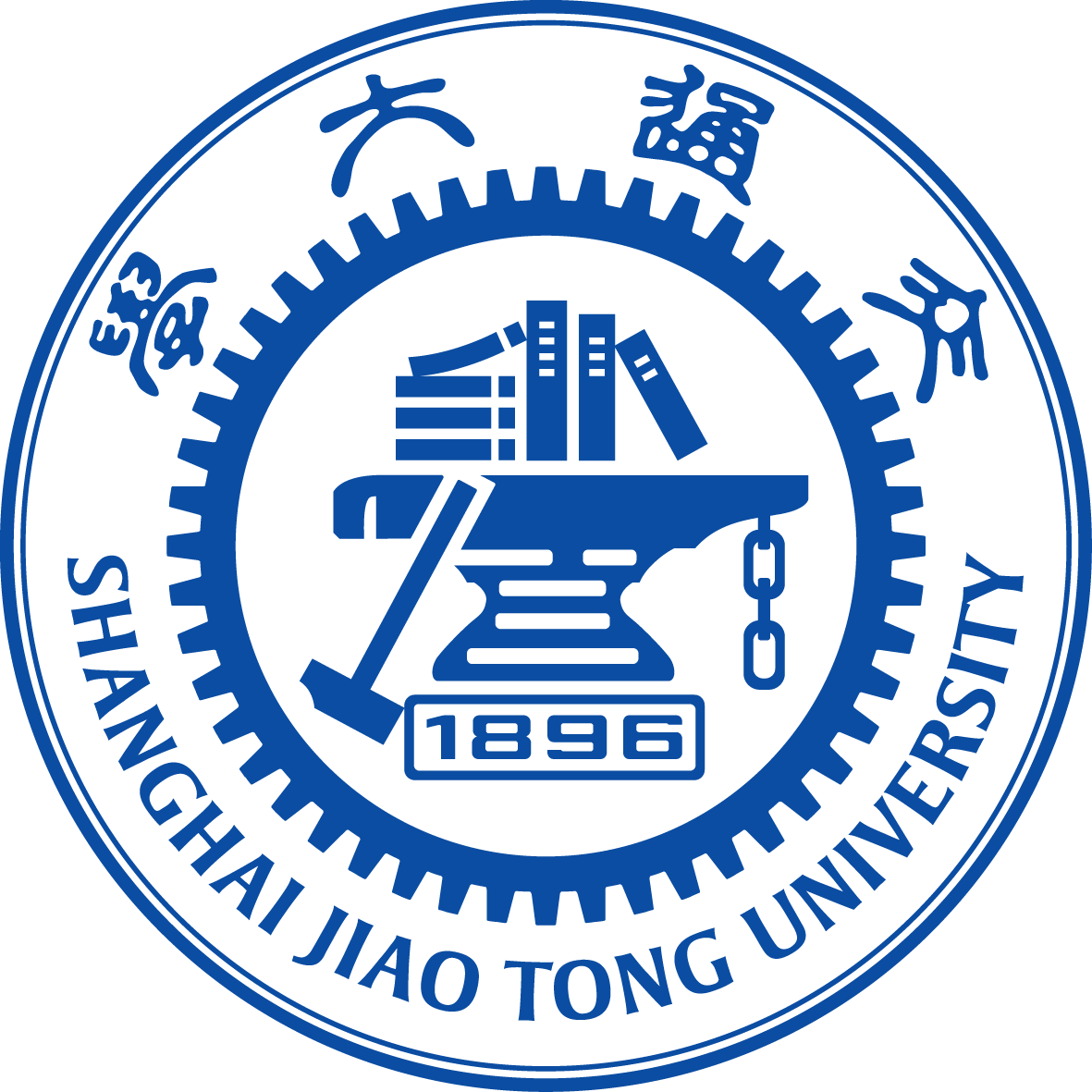
Web Search and Mining (F033583)



The Search System

Charlie Kruger J11603099010

Daniel Aldana 116030990081

Introduction

The search system consists basically of the following items listed below:

1. **The Web Crawler** which crawls the follow site (<https://www.realself.com/>) and obtain the frequently asked question (FAQ) and the doctor list and these are stored in a local directory.

2. **The Parser** will obtain the specific information of the crawled files and will make a first term list.

3. **The Intermediate** creates second list terms based on the parsed files.

4. **Query Parser** realizes spell corrections and utilize refinement techniques.

5. **Scoring table** constructs and build a full table of the scores and ranking the terms getting from the index.

6. **TF-IDF** the matrix for tf-idf scoring will sort all the documents by the respective relevance.

7. **The Intermediate** construct a term scoring by the numbers of reviews.

8. **Return Results** The results will be retrieved for the wait of the graphical interface.

9. **GUI** The Graphical User Interface will consist of a search button and a textbox for the input of the keyword.

10. **Returned Results** Number of item that can be returned and shown by page.

Detailed Description of the Search System Elements

**Web Crawler:**

The .html files are stored in the user’s directory under the folder tsg/raw once there the folders will be created per the crawled files downloaded, those categories are listed as following

* Frequently Asked Question (FAQ)
* Doctors

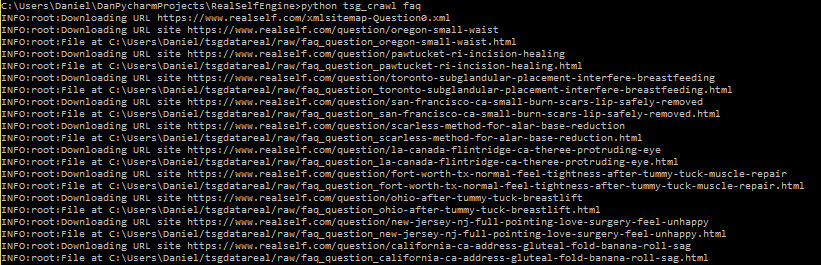
**Item Names**

**Crawled Site:** SITE = '*https://www.realself.com*'

**FAQ items:**

FAQ Site: */xmlsitemap-Question{}.xml*

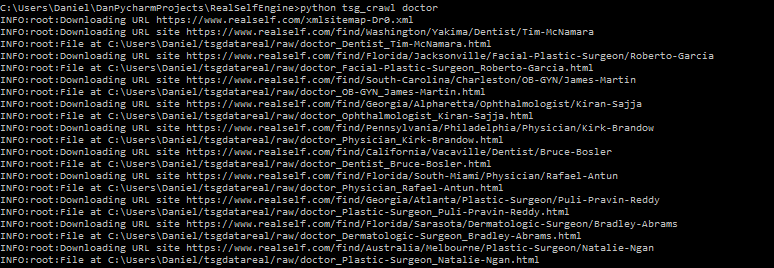
FAQ Files*: faq\_question\_oregon-small-waist.html*



**Doctor items:**

Doctor Site: */xmlsitemap-Dr{}.xml*

Doctor Files *doctor\_Anesthesiologist\_Alex-Roher.html*



**Parser:**

The parser read all the crawled html files and converts into a json file with a more structured information with parsed contents.

All the output files are listed bellowed:

* type (whether the file is faq o doctor)
* title (faq’s title or doctor name)
* content (parsed content)
* uuid (the identification of the file, how to find unique IDs by hash or content)
* link (the entire link of the html file on the realself.com site)

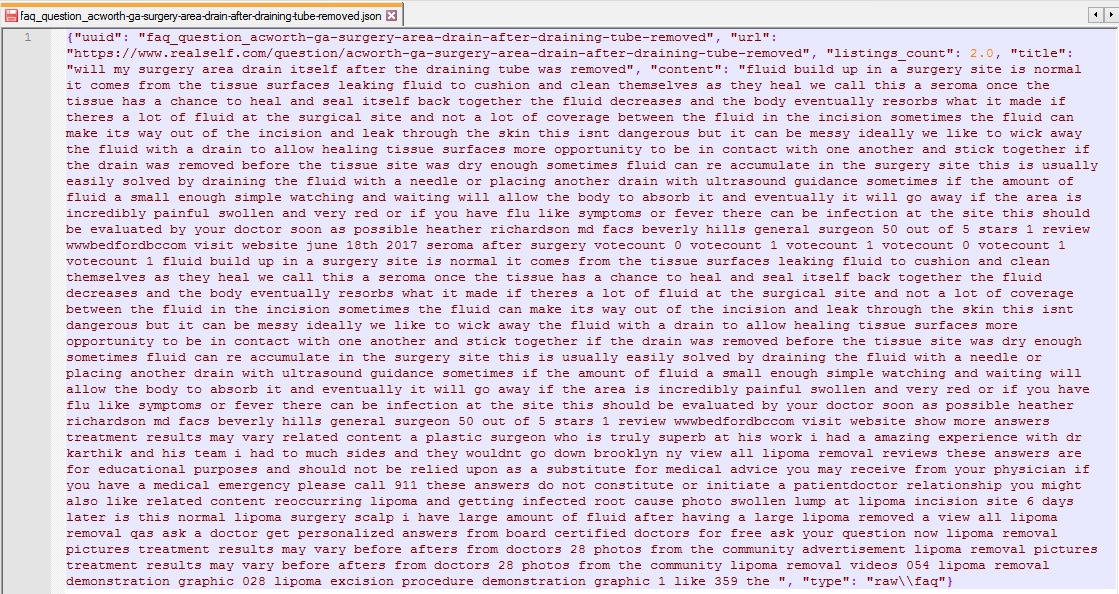
The content section is a parsed string list of the content separated by spaces and converted into lowercases.

**OUTPUT**

The folder parsed will be created inside the raw directory, the names of the files will be UUID.json

**FAQ Files**



**\*** Parts of the content section were removed, otherwise the follow image could not be displayed correctly.

**Doctor Files**

