

Image Search IR System

WS2020-21 Information Retrieval Project

Claudio Maggioni

December 7, 2020

Contents

1	Introduction	2
2	Installation instructions	2
2.1	Project repository	2
2.2	Solr installation	2
2.3	UI installation	2
2.4	Run the website scrapers	2
3	Scraping	3
4	Indexing and Solr configuration	3
5	User interface	3
6	User evaluation	3

1 Introduction

This report is a summary of the work I have done to create the “Image Search IR system”, a proof-of-concept IR system implementation implementing the “Image Search Engine” project (project #13).

The project is built on a simple *Scrapy-Solr-HTML5+CSS+JS* stack. Installation instructions, an in-depth look to the project components for scraping, indexing, and displaying the results, and finally the user evaluation report, can all be found in the following sections.

2 Installation instructions

2.1 Project repository

The project Git repository is located here: <https://git.maggioni.xyz/maggicl/IRProject>.

2.2 Solr installation

The installation of the project and population of the test collection with the scraped documents is automated by a single script. The script requires you have downloaded *Solr* version 8.6.2. as a ZIP file, i.e. the same *Solr* ZIP we had to download during lab lectures. Should you need to download a copy of the ZIP file, you can find it here (on USI’s onedrive hosting): <http://to-do.com/file>.

Clone the project’s git repository and position yourself with a shell on the project’s root directory. Then execute this command:

```
./solr_install.sh {ZIP path}
```

where `<ZIP path>` is the path of the ZIP file mentioned earlier. This will install, start, and update *Solr* with the test collection.

2.3 UI installation

In order to start the UI, open with your browser of choice the file `ui/index.html`. In order to use the UI, it is necessary to bypass **Cross Origin Resource Sharing** security checks by downloading and enabling a “CORS everywhere” extension. I suggest this one for Mozilla Firefox and derivatives.

2.4 Run the website scrapers

A prerequisite to run the Flickr crawler is to have a working Scrapy Splash instance listening on port `localhost:8050`. This can be achieved by executing this Docker command, should a Docker installation be available:

```
docker run -p 8050:8050 scrapinghub/scrapy
```

In order to all the website scrapers, run the script `./scrape.sh` with no arguments.

3 Scraping

4 Indexing and Solr configuration

5 User interface

6 User evaluation