INFO 300

Term Project

2022.11.27

Project Name	Chinese Game Job Application Data Retrieval System		
Data Source	https://www.shixiseng.com		
Course Name	INFO 300		
Instructor	Su Wei & An Yuan		
Team Leader	Jiahe Xie		
Team Member	Name	Student ID	E-mail
	Jiahe Xie	320200945941	xiejh20@lzu.edu.cn
	Aoyu Liu	320200931091	liuay20@lzu.edu.cn
	Yunyi Wang	320200931251	wangyunyi20@lzu.edu.cn
	Yuanzhe Yang	320200931441	Yangyzh2020@lzu.edu,cn

1. Abstract

Interested in the gameplay development, our team choose the game job data from Chinese website ShiXiSeng as our test collection to learn about the information about the related work in the industry. ShiXiSeng is a Job recruiting website. Using Game as keywords, over 250 search results are crawled for our project. On this foundation, 3 different systems are constructed based on 3 distinct similarity score algorithms including BM25, IB and DFR. After the development finishes, these 3 system are evaluated by their MAP index. The result show that the DFR algorithm fits the datasets best. i.e. System3.

2. Duties

Team work is carried out in the form of file transmission. Every one does their parts and transfer their results and dependencies to the next one.

Jiahe Xie is responsible for the website selection, data crawl and analysis (Step 1).

Aoyu Liu is responsible for the baseline IR system construction. His works focus on Creating UI of the search engine using ElasticSearch Python API (Step 2).

Yunyi Wang is responsible for the Extension of the IR system using distinct similarity score algorithm to create new systems (Step 3).

Yuanzhe Yang is responsible for the MAP calculation so that we can know about pros and cons of different systems. (Step 4)

3. Content

3.1 Step 1

3.1.1 Preparation

After deciding search dataset of game job index, the next problem which is needed to solved is to select a job website to crawl. The best choice has to be structured web page and weak anti-spider. Then the Shixiseng attracts our team attention. When we finish checking the information completeness and anti-spider mechanism, it is chosen

3.1.2 Scrapy API Using

Instead of using CSS selector, xpath is called to match the html text. It is better for the crawl part. Also, callbacks are used for solve pages division logic.

3.1.3 Challenge

The largest difficult in this part of work is to solve the anti-spider problem about the website. The website uses custom encoding method differing from usual encoding method like utf-8 to encrypt their text information. Then we found the rules about some definite chars and decode them in Python to achieve the goal of data cleaning.

3.1.4 Dataset

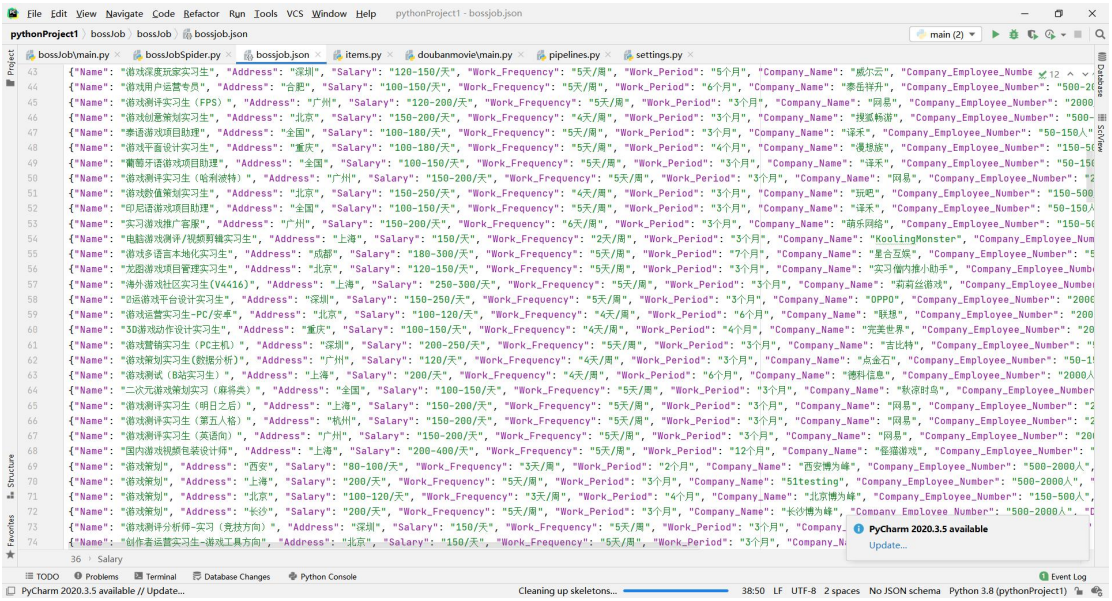


Figure 3.1.4 dataset

3.2 Step 2

3.2.1 Upload the dataset

At first, we thought this would be easy because we have learned it from the previous class, however, when we began our project, the server has been closed, so we can't upload the dataset through the website. What we do to address this problem is that we use the ElasticSearch Python API. We access the ElasticSearch with Aoyu Liu's account information like below:

```
context = create_ssl_context()
context.check_hostname = False
context.verify_mode = ssl.CERT_NONE

es = Elasticsearch(
    ['121.43.234.160'],
    http_auth=('liuay20', 'liuay20'),
    scheme="http",
    port=9200,
    ssl_context=context,
)
```

Figure 1 Connecting to ElasticSearch

Then we upload the dataset line by line with the index function.

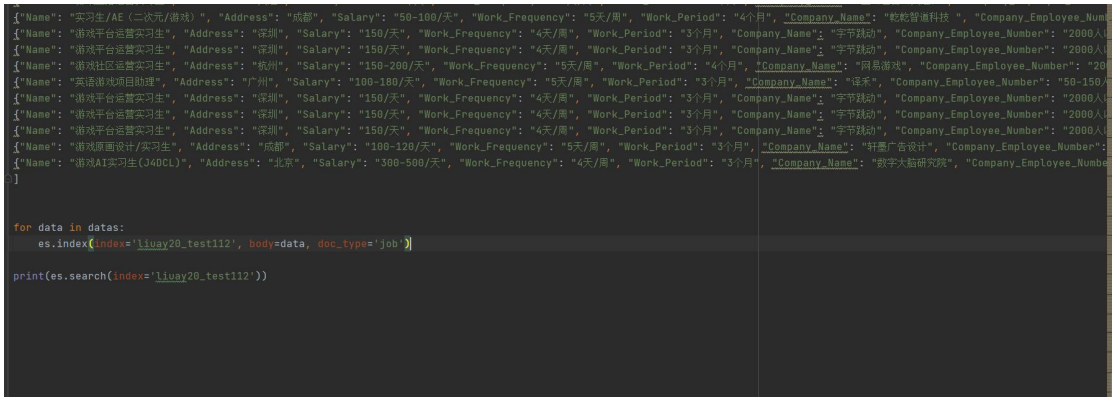


Figure 3.2.1 Upload the dataset

3.2.2 Establish the engine with Flask

Once the dataset is uploaded, we began to create the engine. Similarly, we connect to the ElasticSearch first, then with Flask, we define the homepage as home.html. In the homepage we can select different ranking methods and input the keywords you want to search, once they are given, with the get method, we can get a specific query and with the query to find the result, and turn to the result page which is result.html. In the result page you can see that how many records have been retrieved.

3.2.3 Applying the CSS style

With the intention of beautifying the page, we add some CSS style to the pages. For the framework, we choose the Bootstrap and use the online link instead of the local files.

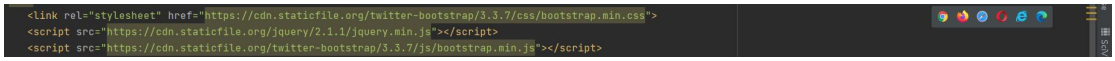


Figure 3.2.3.1 Applying Bootstrap

For the homepage, we generally use the jumbotron component, just to highlight the search engine and make it in the center of the page.

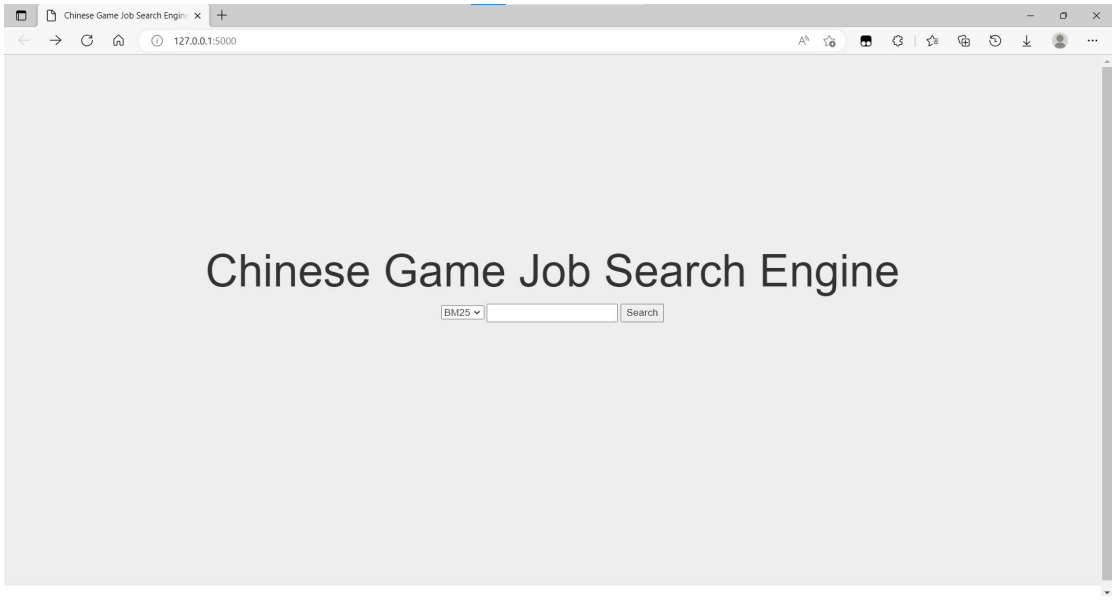
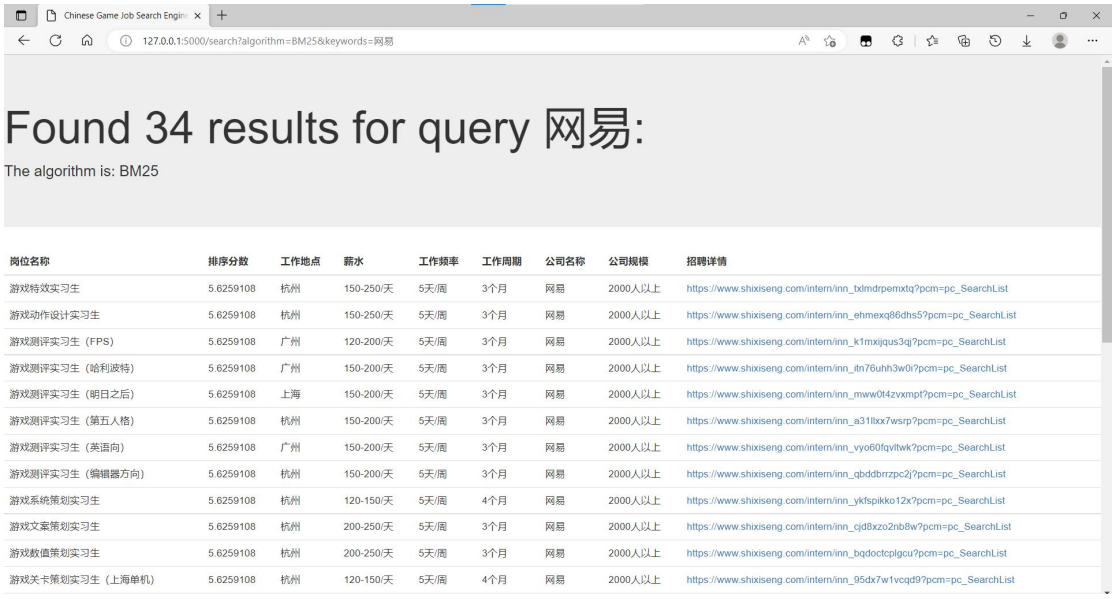


Figure 3.2.3.2 The homepage after applying the CSS style

For the result page, we also apply the jumbotron to emphasize how many results have been retrieved. For the results that are represented, we use the table component to make them uniformed, and we can clearly see all the columns.



岗位名称	排序分数	工作地点	薪水	工作频率	工作周期	公司名称	公司规模	招聘详情
游戏特效实习生	5.6259108	杭州	150-250/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_txlmdrpemxtq?pcm=pc_SearchList
游戏动作设计实习生	5.6259108	杭州	150-250/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_ehmexq86dhs5?pcm=pc_SearchList
游戏测评实习生（FPS）	5.6259108	广州	120-200/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_k1mxijqus3qj?pcm=pc_SearchList
游戏测评实习生（哈利波特）	5.6259108	广州	150-200/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_itn76uhh3w0i?pcm=pc_SearchList
游戏测评实习生（明日之后）	5.6259108	上海	150-200/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_mww0t4zvxmpt?pcm=pc_SearchList
游戏测评实习生（第五人格）	5.6259108	杭州	150-200/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_a31lbx7wsrp?pcm=pc_SearchList
游戏测评实习生（英语向）	5.6259108	广州	150-200/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_vyo60fqvltwk?pcm=pc_SearchList
游戏测评实习生（编辑器方向）	5.6259108	杭州	150-200/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_qbddbrrzpc2j?pcm=pc_SearchList
游戏系统策划实习生	5.6259108	杭州	120-150/天	5天/周	4个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_ykfspikko12x?pcm=pc_SearchList
游戏文案策划实习生	5.6259108	杭州	200-250/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_cjd8xzo2nb8w?pcm=pc_SearchList
游戏数值策划实习生	5.6259108	杭州	200-250/天	5天/周	3个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_bqdoctcplgcu?pcm=pc_SearchList
游戏关卡策划实习生（上海单机）	5.6259108	杭州	120-150/天	5天/周	4个月	网易	2000人以上	https://www.shixiseng.com/intern/inn_95dx7w1vcqd9?pcm=pc_SearchList

Figure 3.2.3.3 The result page after applying the CSS

At first, we only represent the top 30 results for the page limitation. However, in the following learning we leaned about how to paging, so we can represent all the results that have been retrieved with each page represent 20 records, and the users can select the results according to their own requirements.

Figure 3.2.3.4 The paging function

3.3 Step 3

3.3.1 System 2&3

There's already an index which has been uploaded the data we collect, and for the system 2 and 3, we use different similarity algorithm to distinguish them. After reindex, we upload the new settings for each, which contain different default algorithms, one is IB algorithm and another one is DFR algorithm.

3.3.2 making query

After collecting the keywords and algorithm selection from the webpage, the system will make a judge, if the algorithm is "DFR", system will send search query to the index for DFR algorithm in elasticsearch. if algorithm is "IB", system will choose the index for IB. And if it's "BM25", system will choose the first index because the default algorithm of it is BM25.

3.3.3Challenge

We find that straightly giving the new settings will return error message, so we close the index first, then sending the new settings, which will make it success.

3.4 Step 4

3.4.1 Calculation and Comparation

Query 1：游戏营销实习生

Retrived	System1	System2	System3
Item1	1	1	1
Item2	1	0	1
Item3	1	1	1
Item4	0	1	1
Item5	1	1	1
Item6	0	1	1
Item7	1	1	1
Item8	1	1	1
Item9	1	1	1
Item10	0	0	1

ap1 = (1/1+2/2+3/3+4/5+5/7+6/8+7/9)/7 = .86

ap2 = (1/1+2/3+3/4+4/5+5/6+6/7+7/8+8/9)/8 = .83

ap3 = 1*10/10 = 1

Query 2：上海

Retrived	System1	System2	System3
Item1	1	1	1
Item2	1	1	1
Item3	1	1	1
Item4	1	1	0
Item5	0	1	0
Item6	1	0	1
Item7	0	1	1
Item8	1	1	1
Item9	1	1	1
Item10	1	1	1

ap1 = (1/1+2/2+3/3+4/4+5/6+6/8+7/9+8/10)/8 = .895

ap2 = (1/1+2/2+3/3+4/4+5/5+6/7+7/8+8/9+9/10)/9 = .95

ap3 = (1/1+2/2+3/3+4/6+5/7+6/8+7/9+8/10)/8 = .84

Query 3：游戏原画实习生 长沙

Retrived	System1	System2	System3
Item1	1	1	1
Item2	0	1	1
Item3	1	1	1
Item4	1	1	1
Item5	1	1	0
Item6	1	1	1
Item7	1	0	1
Item8	1	0	1
Item9	0	1	0
Item10	0	1	0

ap1 = (1/1+2/3+3/4+4/5+5/6+6/7+7/8)/7 = .83

ap2 = (1/1+2/2+3/3+4/4+5/5+6/6+7/9+8/10)/8 = .95

ap3 = (1/1+2/2+3/3+4/4+5/6+6/7+7/8)/7 = .94

Query 4：4 天/周 北京

Retrived	System1	System2	System3
Item1	1	1	1
Item2	1	1	1
Item3	1	1	1
Item4	1	1	1
Item5	0	0	1
Item6	1	1	1
Item7	1	1	1
Item8	1	1	1
Item9	1	0	0
Item10	1	1	0

ap1 = (1/1+2/2+3/3+4/4+5/6+6/7+7/8+8/9+9/10)/9 = 0.93

ap2 = (1/1+2/2+3/3+4/4+5/6+6/7+7/8+8/10)/8 = .92

ap3 = (1/1+2/2+3/3+4/4+5/5+6/6+7/7+8/8)/8 = 1

Query 5：北京 成都

Retrived	System1	System2	System3
Item1	1	1	1
Item2	1	1	1
Item3	1	0	1
Item4	1	0	1
Item5	0	0	1
Item6	1	1	1
Item7	0	1	1
Item8	0	1	1
Item9	1	1	1
Item10	1	1	1

ap1 = (1/1+2/2+3/3+4/4+5/6+6/9+7/10)/7 = .89

ap2 = (1/1+2/2+3/6+4/7+5/8+6/9+7/10)/7 = .72

ap3 = (1*10/10 = 1

Query 6：150-200/天 游戏测试工程师

Retrived	System1	System2	System3
Item1	1	1	1
Item2	1	1	1
Item3	1	1	1
Item4	1	1	1
Item5	1	0	0
Item6	0	0	0
Item7	0	0	1
Item8	0	1	0
Item9	1	1	0
Item10	1	1	0

ap1 = (1/1+2/2+3/3+4/4+5/5+6/9+7/10)/7 = .91

ap2 = (1/1+2/2+3/3+4/4+5/8+6/9+7/10)/7 = .86

ap3 = (1/1+2/2+3/3+4/4+5/7)/5 = .942

Map

Query--ap	Ap1	Ap2	Ap3
1	.86	.83	1
2	.895	.95	.84
3	.83	.95	.94
4	.93	.92	1
5	.89	.72	1
6	.91	.86	.942
map	.886	.872	.954

Map(System 3) > Map(System 1) > Map(System 2)

So system3 is better, and system3 and system1.

3.4.2 Evaluation

In this session, we run 6 different queries, and the details are presented in the report.

The way to show the search results of 6 queries on 3 different systems come from the module provided on blackboard except we use 1 for relevant and 0 for not revelant.

Then after the 6 queries, there's a table showing the average precision of each query on each system. The last row of the table is the mean average precision of each system.

4. Code

Github: https://github.com/rufuzhanshi/INFO_300_Team_Project

5. Conclusion

After the comparasion of the mean average precision, we came to the conclusion that system 1 is the best, system 3 is the second and system 2 is the worst.