

REPORT OF INFO 300 TERM PROJECT

Team Members:

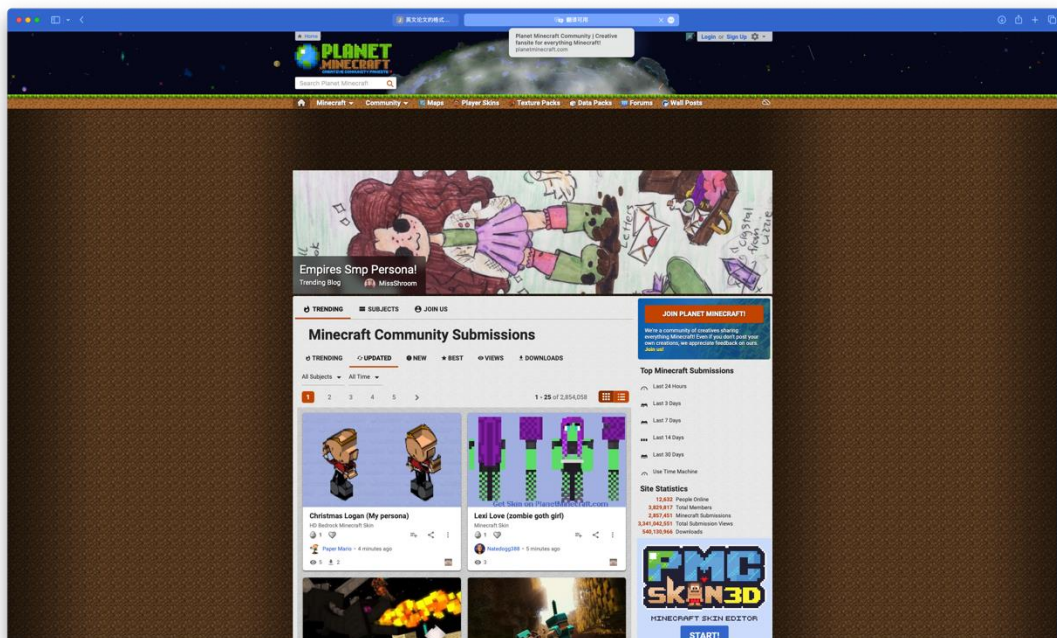
Name	Email	Student ID
Wenbo Liu	wbliu20@lzu.edu.cn	320200931121
Xuda Han	hanxd20@lzu.edu.cn	320200945781
Chengkai Lin	linchk@lzu.edu.cn	320200945841
Yifei Guo	guoyf@lzu.edu.cn	320200930901

Step 1. Prepare a Test Collection:

- This part is completed by Chengkai Lin.

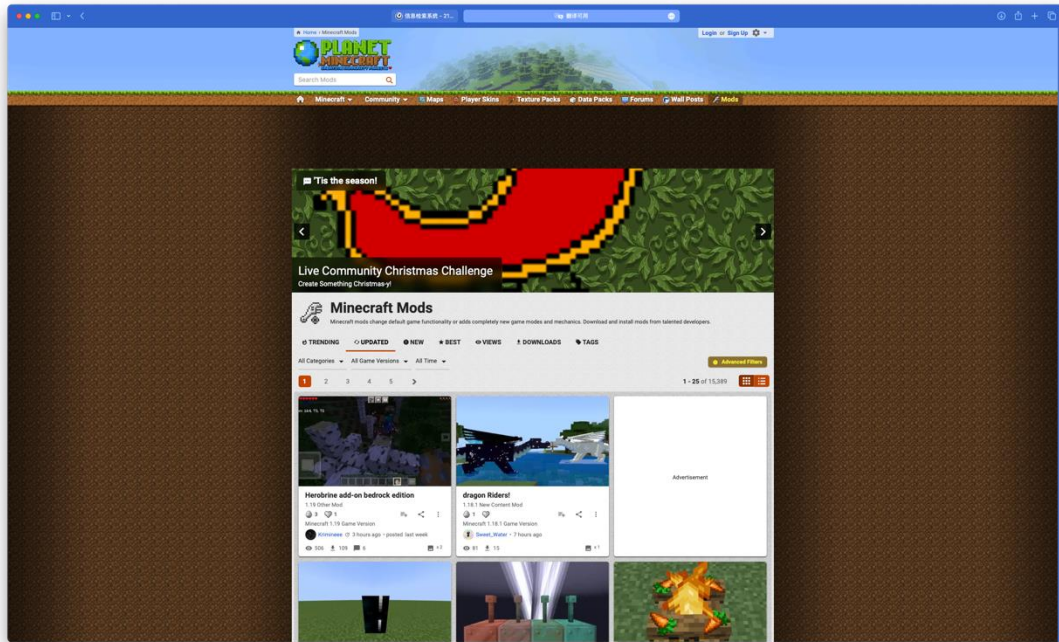
Minecraft is a sandbox video game developed by Mojang Studios. In Minecraft, players explore a blocky, procedurally generated 3D world with virtually infinite terrain and may discover and extract raw materials, craft tools and items, and build structures, earthworks, and simple machines. Depending on their chosen game mode, players can fight hostile mobs, as well as cooperate with or compete against other players in the same world. Game modes include a survival mode (in which players must acquire resources to build in the world and maintain health) and a creative mode (where players have unlimited resources and access to flight). There is also a wide variety of user-generated content, such as modifications, servers, skins, texture packs, and custom maps, which add new game mechanics and possibilities. That user-generated content is what we're interested in.

We can easily get a variety of user-generated content from the Minecraft community which is supported by gamers from all over the world. The image (P-1.1) below shows one of the Minecraft communities.



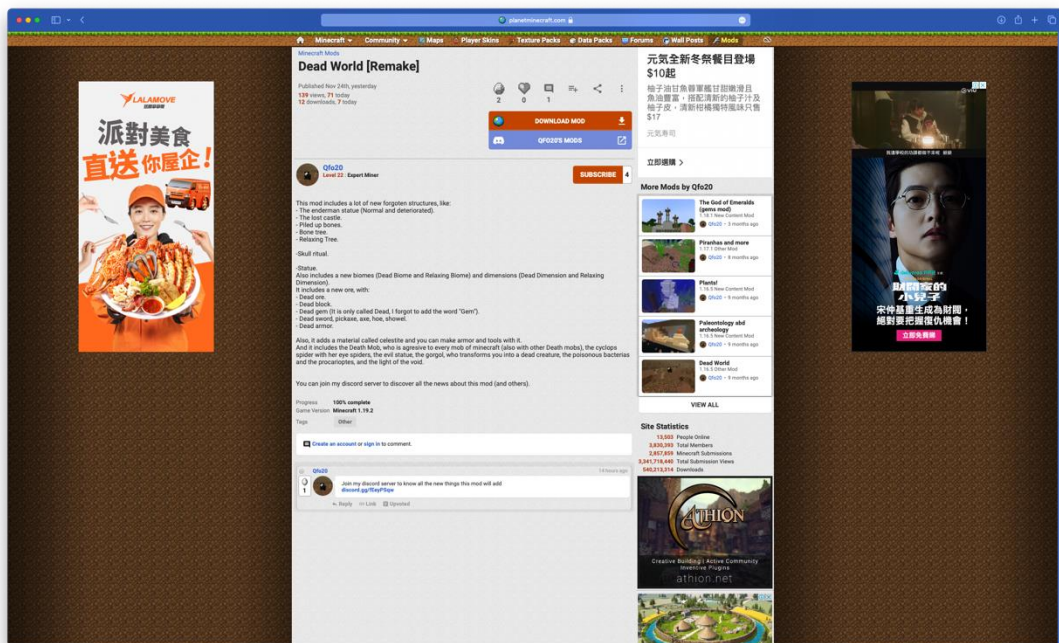
P-1. 1

What we want to crawl are game mods uploaded on the community. The image (P-1.2) behind shows one mods page.



P-1.2

Through the analysis of the web pages, we find that different mods pages have similar URLs. For example, URL of the first mods page is <https://www.planetminecraft.com/mods/?p=1>, and the second's is <https://www.planetminecraft.com/mods/?p=2>. Only the value of p varies in different URLs. Because of this, we can crawl all mods pages by changing the value of p. After getting all the mods pages, we need to further analyze the pages to get the address of each mod. We extract the URLs from the pages through CSS. The final step is to extract the data contained in each mod page. As you can see from the image (P-1.3), each mod page contains information such as title, description, published time, views, downloads, credit, progress, game version, tags, and URL. We extract the data from the page through XPath.



P-1.3

The image (P-1.4) below shows our spider script (**minecraft.py**). The function 'start_request' crawls the first 100 pages containing mod addresses and sends them to the parser 'parse' to crawl each mod page. After that, each mod page is handed over to the parser 'parse_detail' to parse the detailed data of each mod.

```
import scrapy
import time

class MinicraftSpider(scrapy.Spider):
    """
    scrapy spider
    """
    name = 'minicraft'

    def start_requests(self):
        url = f'https://www.minicraft.com/zh-cn/?p={id}'
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)

    def parse(self, response, request, meta):
        for url in response.css('div.r-previse a::attr(href)').getall():
            time.sleep(1)
            yield scrapy.Request(url=url, callback=self.parse_detail)

    def parse_detail(self, response):
        yield {
            'title': response.xpath('//div[@id="resource-title-text"]//div/text()').extract_first(),
            'description': response.xpath('//div[@id="r-text-block"]//div/text()').extract(),
            'published_time': response.xpath('//div[@id="resource-info"]//div[@id="time"]//div/text()').extract_first(),
            'views': response.xpath('//div[@id="resource-info"]//div[@id="views"]//div/text()').extract_first(),
            'downloads': response.xpath('//div[@id="resource-info"]//div[@id="downloads"]//div/text()').extract_first(),
            'tags': response.xpath('//div[@id="resource-info"]//div[@id="tags"]//div/text()').extract_first(),
            'name': response.xpath('//div[@id="resource-info"]//div[@id="name"]//div/text()').extract_first(),
            'version': response.xpath('//div[@id="resource-info"]//div[@id="version"]//div/text()').extract_first(),
            'tags': response.xpath('//div[@id="resource-info"]//div[@id="tags"]//div/text()').extract_first(),
            'url': response.url
        }
```

P-1.4

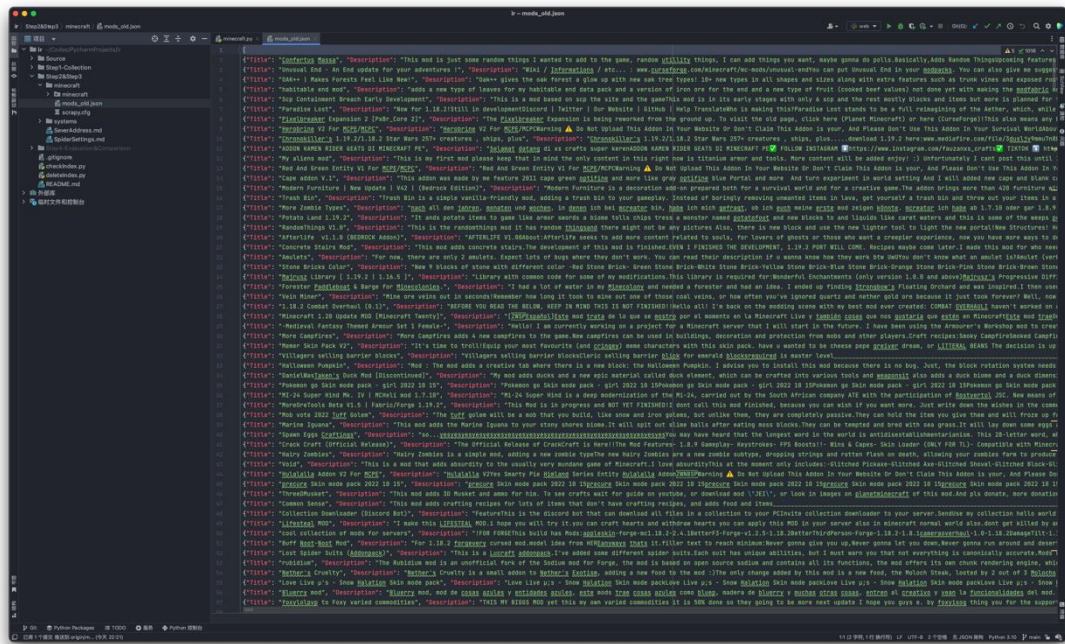
After obtaining the data, we also need to clean the data, such as data type conversion, string concatenation and so on. This is all set up in the **pipeline.py**.

```
import scrapy
import time

class MinicraftPipeline:
    """
    scrapy pipeline
    """
    def process_item(self, item, spider):
        description = ''
        if item.get('description'):
            description = item['description']
        if item.get('views'):
            views = int(item['views'].replace(',',''))
        if item.get('downloads'):
            downloads = int(item['downloads'].replace(',',''))
        if item.get('tags'):
            tags = item['tags']
        return item
```

P-1.5

The image below shows the data (**mods_old.json**) we got from the crawl. There are 950 pieces of data here.



P-1.6

Step 2. Implement a Baseline IR System:

- This part is completed by Wenbo Liu and Xuda Han.
 - i. Wenbo Liu - Search Interface
 - ii. Xuda Han – Web Page

Index the test collection using the default setting of Elasticsearch

To implement the baseline information retrieval system, we should upload the collections to Elasticsearch and index them firstly. Using the properties of scrapy, we can upload our data to Elasticsearch while crawling. To do this, we need to add the server address, index name, username, password and Elasticsearch pipeline to the file **setting.py** (P-2.1).

```
BOT_NAME = 'minecraft'

SPIDER_MODULES = ['minecraft.spiders']
NEWSPIDER_MODULE = 'minecraft.spiders'

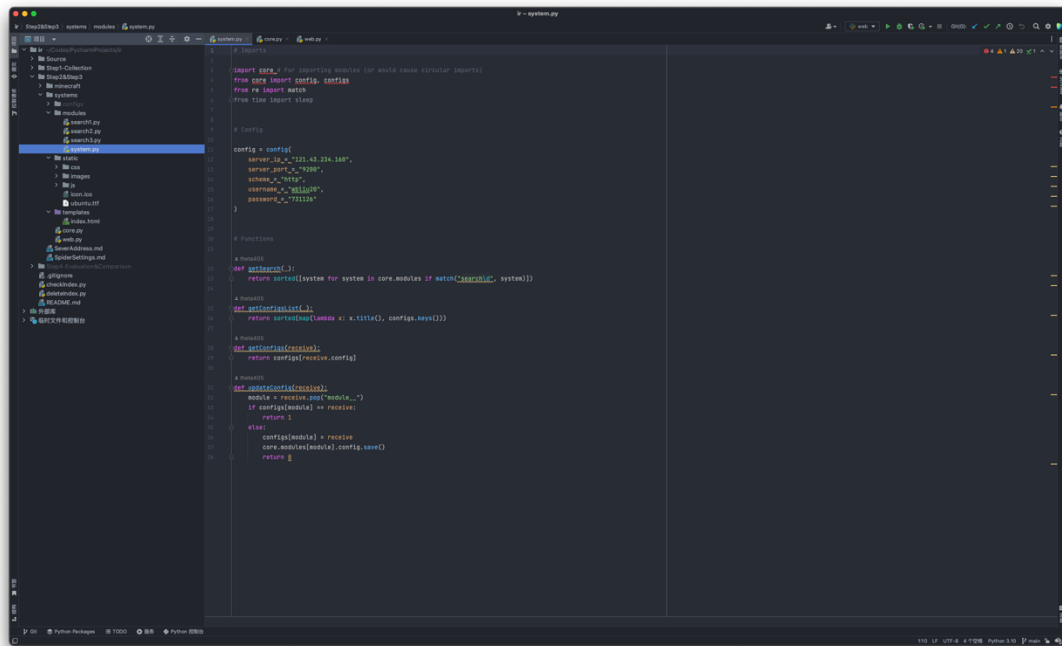
FEED_EXPORT_ENCODING = 'utf-8'
# new: 'http://wbllyu20:731126@210.26.48.81:9201'
# old: 'http://wbllyu20:731126@121.43.234.160:9200'
ELASTICSEARCH_SERVERS = 'http://wbllyu20:731126@121.43.234.160:9200'
ELASTICSEARCH_INDEX = 'wbllyu20_team_project'
ELASTICSEARCH_TYPE = '_doc'

# Crawl responsibly by identifying yourself (and your website) on the user-agent
USER_AGENT = 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/605.1.15 (KHTML, like Gecko) ' \
'Version/16.1 Safari/605.1.15'
```

P-2.1

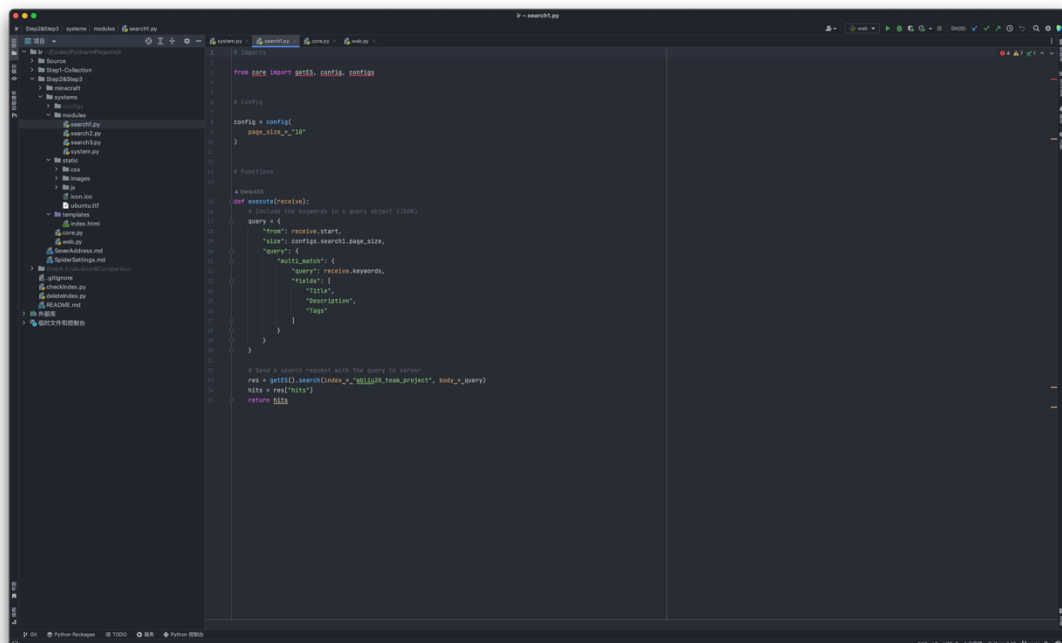
Use the Elasticsearch Python API to create a search interface

To create a search interface, we use Elasticsearch module (version 7.12.0) of Python. First, define Elasticsearch configuration information in the **system.py** file (P-2.2), such as the server address, index, account, password, etc. Other functions in the file are used to implement the features of the web page, such as system selection, saving the configuration file, updating the configuration file, and so on. We will show the specific features in the web page.



P-2. 2

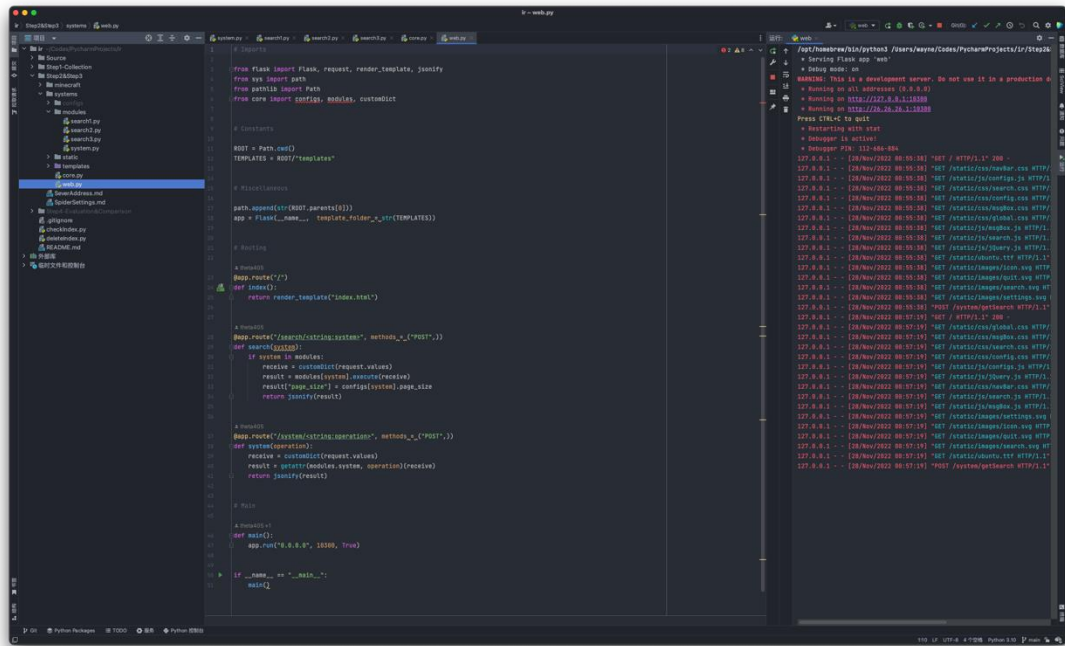
Secondly, we define the query statement of the baseline information retrieval system and the display range of each result page in **search1.py** (P-2.3). We will match the results using the default similarity algorithm in the title, description and tags fields.



P-2. 3

The default similarity algorithm of Elasticsearch is BM25.

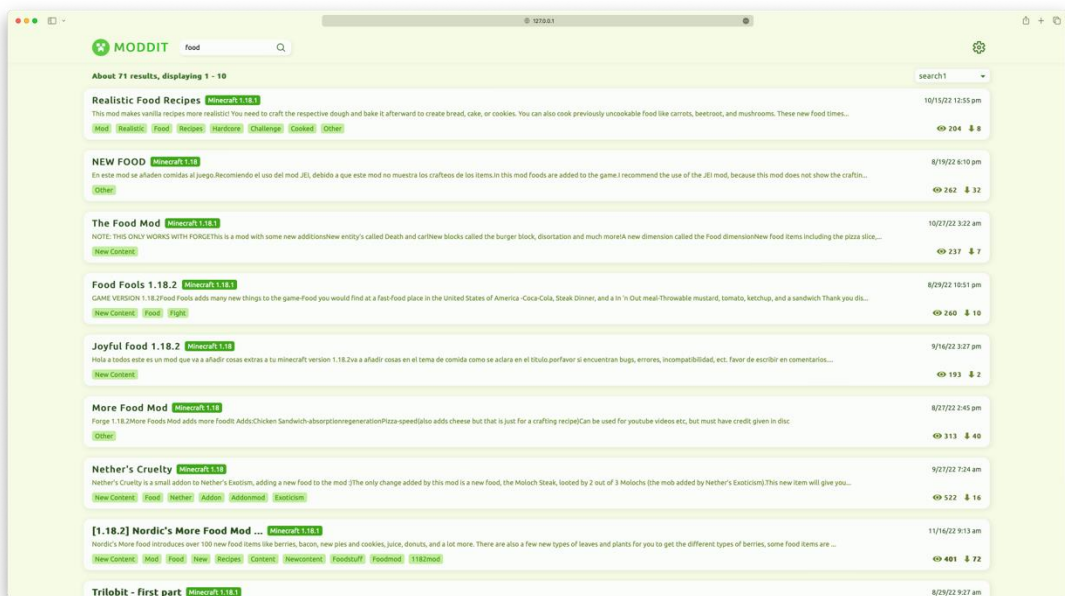
Finally, we initialize the configuration in **core.py** (P-2.4) and link to the Elasticsearch server.



P-2. 6

Test and make sure the interface and search work correctly

To test and make sure the interface and search work correctly, we use the query statement “food” to search food mods of Minecraft community on Elasticsearch sever. Returned results are shown below.



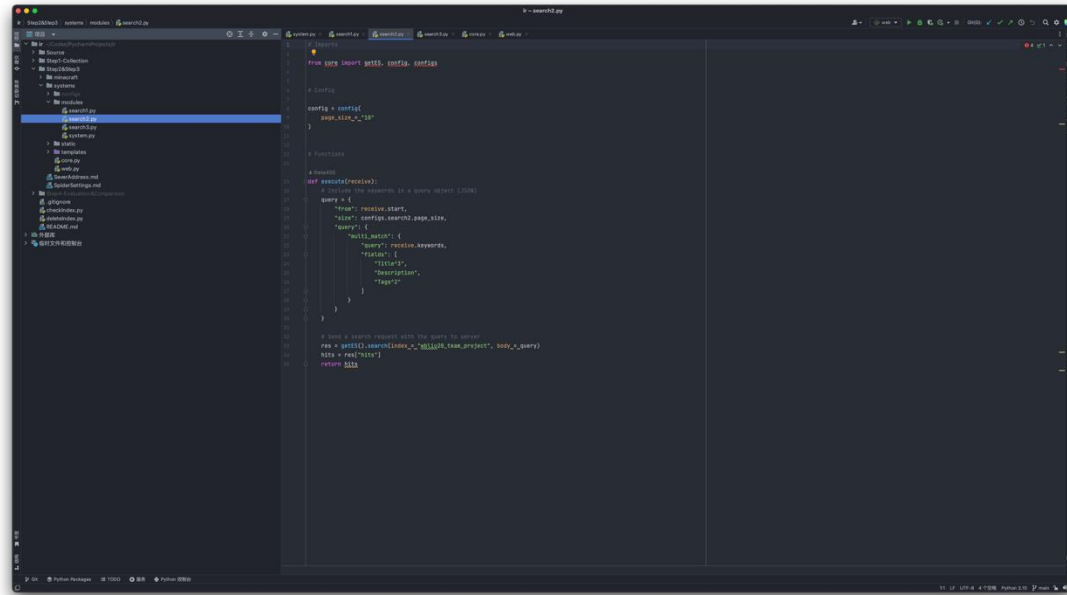
P-2. 7

Step 3. Enhance the Base System:

- This part is completed by Wenbo Liu and Xuda Han.

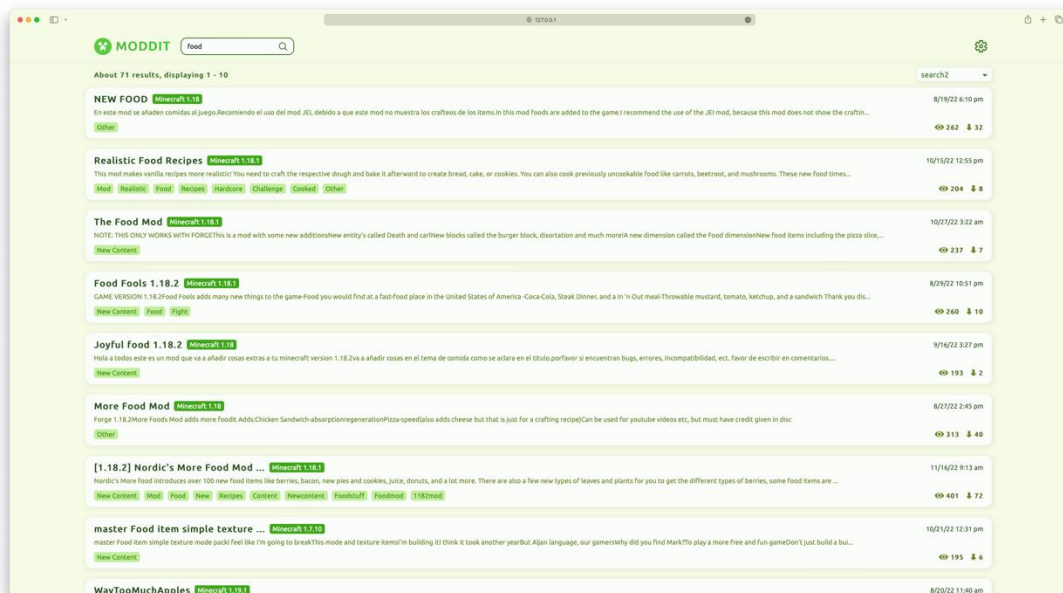
The default similarity algorithm of Elasticsearch is BM25. We will continue to use BM25 in System 2 and System 3. The difference is that we will change the weights of fields in index and the way of sorting.

When we query with the default Settings, the boost values for 'title', 'description' and 'tags' are all 1. However, we note that the description is long, and even if it contains the terms of the query, its contents are not necessarily related to it. Therefore, we set different boost values for different fields to give more weight to 'title' and 'tags' in queries. Below is our setup for the second information retrieval system (P-3.1).



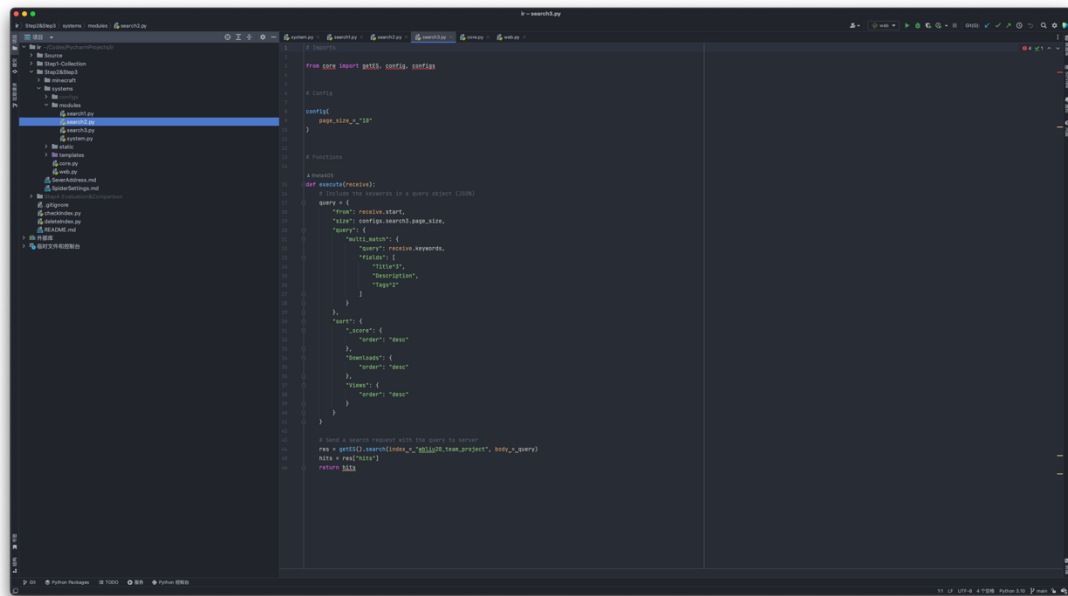
P-3. 1

Use system 2 to search food mods of Minecraft community on Elasticsearch sever. Returned results are shown below.



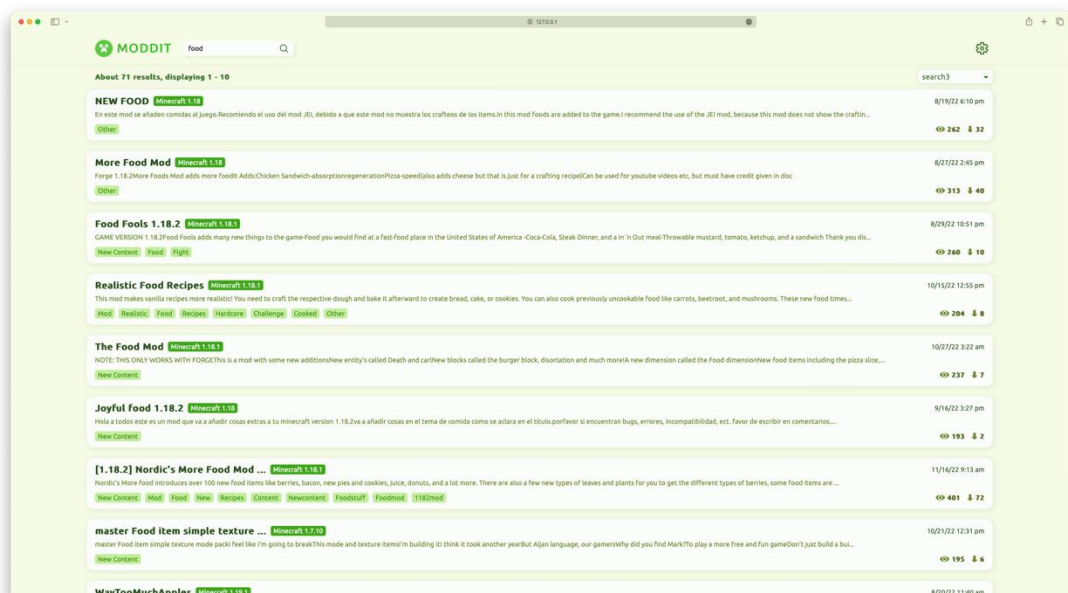
P-3. 2

What if two results have the same score? When we came across two results with the same score, we chose to sort the results by downloads first. When the number of downloads was the same, we chose to sort the results by the number of views. Below is our setup for the third information retrieval system (P-3.3).



P-3.3

Use system 3 to search food mods of Minecraft community on Elasticsearch sever. Returned results are shown below.



P-3.4

Step 4. Evaluate and Compare the Performances:

- This part is completed by Yifei Guo.

Run 6 different queries and make relevant judgments for the top 10 hits

Query 1: food

Retrieved	System1	System2	System3
Item1	R	R	R
Item2	R	R	R
Item3	R	R	R
Item4	R	R	R
Item5	R	R	R
Item6	R	R	R
Item7	R	R	R
Item8	R	N	N
Item9	N	R	R
Item10	N	R	R

Query 2: sword

Retrieved	System1	System2	System3
Item1	R	R	R
Item2	R	R	R
Item3	R	R	R
Item4	R	R	R
Item5	R	R	R
Item6	R	R	R
Item7	R	R	R
Item8	N	R	R
Item9	R	R	R
Item10	R	R	R

Query 3: pickaxe

Retrieved	System1	System2	System3
Item1	R	R	R
Item2	R	R	R
Item3	R	R	R
Item4	R	R	R
Item5	N	R	R
Item6	N	N	N
Item7	R	N	N
Item8	N	R	R
Item9	R	N	N
Item10	N	N	N

Query 4: gold

Retrieved	System1	System2	System3
Item1	R	R	R
Item2	R	R	R

Item3	R	R	R
Item4	R	R	R
Item5	N	N	N
Item6	N	N	N
Item7	R	R	R
Item8	N	N	N
Item9	N	N	N
Item10	R	N	N

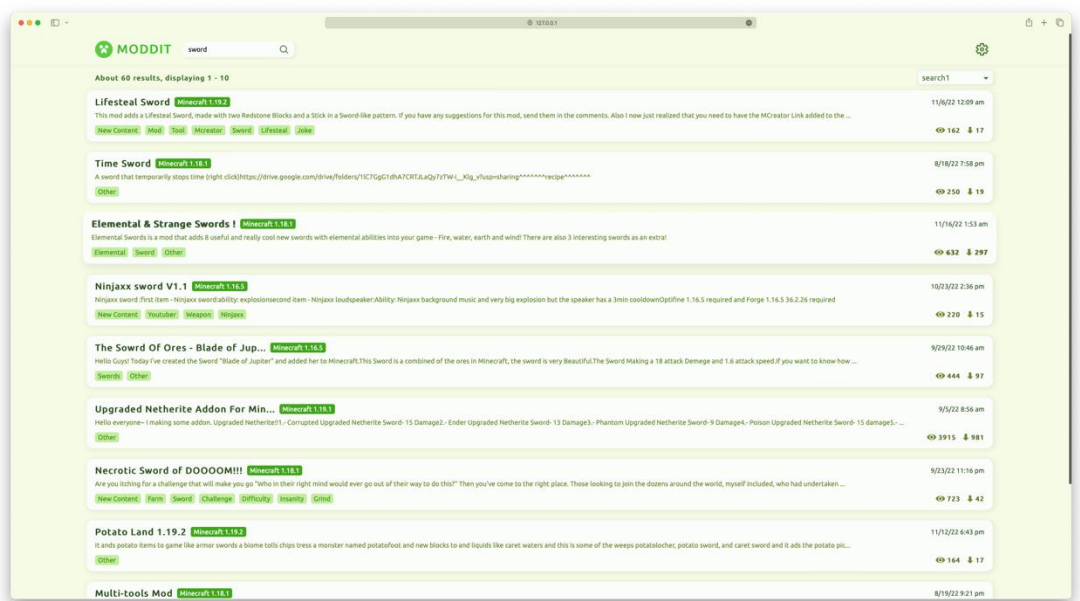
Query 5: potion

Retrieved	System1	System2	System3
Item1	R	R	R
Item2	R	R	R
Item3	R	R	R
Item4	R	R	R
Item5	N	N	N
Item6	R	N	N
Item7	N	R	R
Item8	R	N	N
Item9	N	N	N
Item10	N	R	R

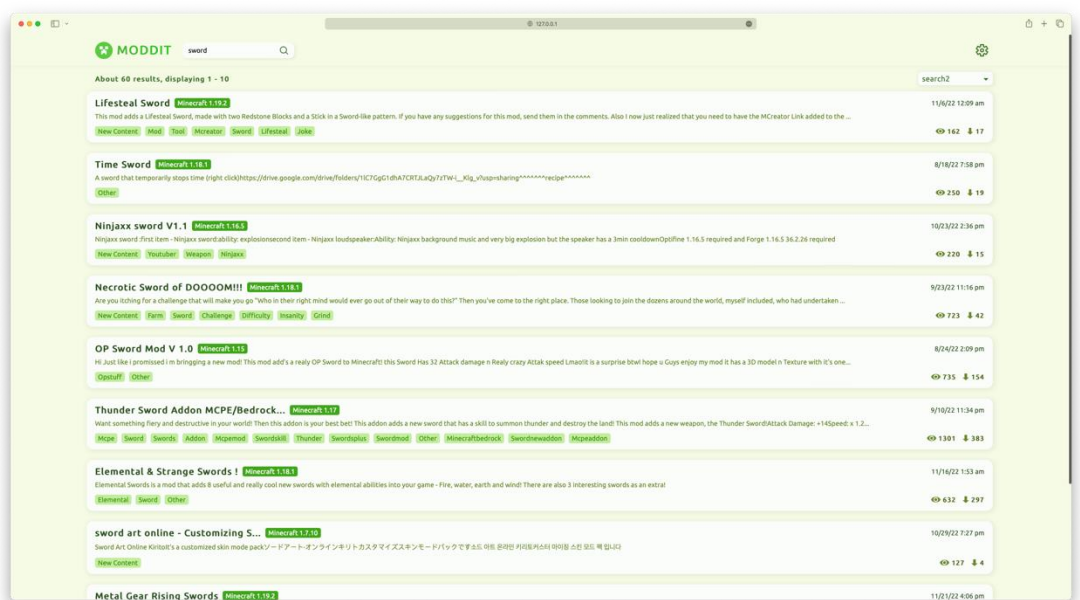
Query 6: dimension

Retrieved	System1	System2	System3
Item1	R	R	R
Item2	R	R	R
Item3	R	R	R
Item4	R	R	R
Item5	N	R	R
Item6	R	R	R
Item7	R	R	R
Item8	R	R	R
Item9	N	R	R
Item10	R	R	R

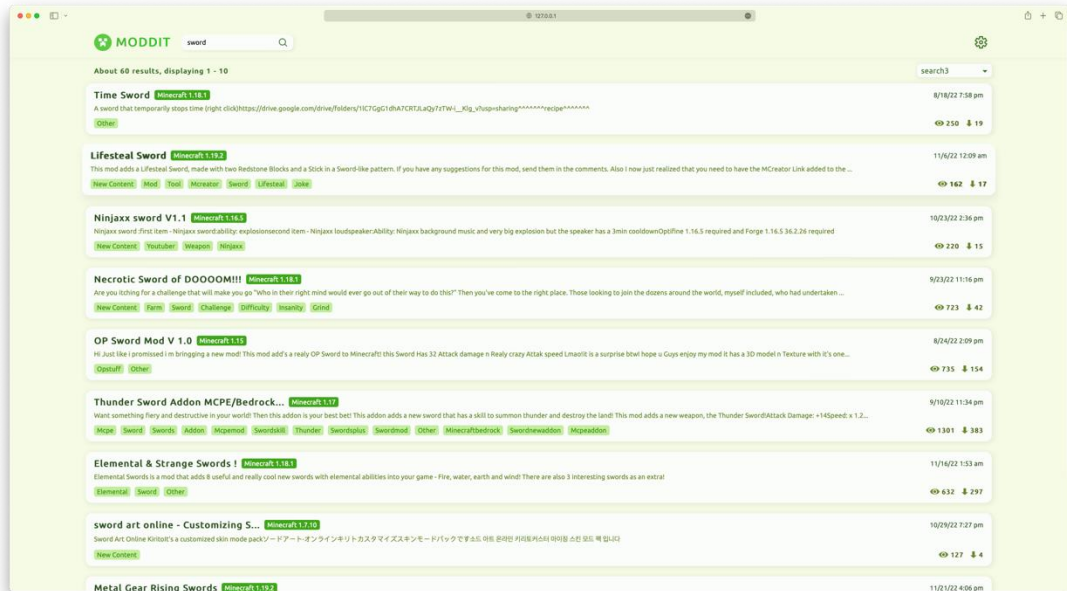
Screen shots of query results (query: sword)



P-4. 1



P-4. 2



P-4. 3

Compare results across 3 systems

By comparing the query results of the three systems, we can find that the accuracy of system 2 and system 3 is improved compared with system 1, but there seems to be no obvious gap between system 2 and system 3. In the case of the same accuracy, the DCG (Discounted Cumulative Gain) of system 2 and system 3 seems to be higher, that is, when the same number of related documents are returned, related documents rank higher in system 2 and system 3.

Calculate MAP for each of the queries on each system

1) Average precision for each query on each system:

Average Precision	Query 1	Query 2	Query 3	Query 4	Query 5	Query 6
System 1	1.0	0.9765	0.8968	0.8857	0.9306	0.9207
System 2	0.9765	1.0	0.9583	0.9429	0.8857	1.0
System 3	0.9765	1.0	0.9583	0.9429	0.8857	1.0

2) DCG for each query on each system:

DCG	Query 1	Query 2	Query 3	Query 4	Query 5	Query 6
System 1	4.6380	4.9212	3.8026	3.7882	3.8511	4.5084
System 2	4.9212	5.2545	3.8949	3.4871	3.7882	5.2545
System 3	4.9212	5.2545	3.8949	3.4871	3.7882	5.2545

3) MAP for each system:

- System 1: $MAP = (1.0 + 0.9765 + 0.8968 + 0.8857 + 0.9306 + 0.9207) / 6 = 0.93505$
- System 2: $MAP = (0.9765 + 1.0 + 0.9583 + 0.9429 + 0.8857 + 1.0) / 6 = 0.9606$
- System 3: $MAP = (0.9765 + 1.0 + 0.9583 + 0.9429 + 0.8857 + 1.0) / 6 = 0.9606$

Discuss and report the comparison results

It can be seen from the above calculation results that the MAP of system 2 and system 3 is slightly larger than that of system 1. This indicates that systems 2 and 3 have improved performance over systems 1. At the same time, we also calculated the DCG value of each query statement on each system. The results show that in most cases, the DCG values of system 2 and system 3 are greater than those of system 1. This indicates that the relevant documents rank higher in the search results of systems 2 and 3.

What We Have Learned

Through this project, we have a deeper understanding of the information retrieval system.

First, during the development of the project, we learned the structure of a complete information retrieval system, how to use scrapy framework to crawl and clean data and build indexes and store data in elastic search, and how to use flask framework to build simple web pages and establish connections with servers.

Second, when evaluating the performance of different systems, we learned the methods to evaluate the performance of different systems and learned to use these methods to conduct a simple analysis of the system. On top of the MAP calculation, we also calculated the DCG value of each query to measure other performance of the system.

Of course, we think there are still problems with our system. Comparing the query results shows that systems 2 and 3 are always not better than system 1, suggesting that our improvements are not applicable in all cases. This is where we think the system should be improved.