CE6155 Homework1 Report

student ID：112502911　　name：胡晉逸

Can answer in Chinese or English(可以使用中文或英文回答)

1. (15%) Please record the performance statistics of the web crawler, for example: how long does it take to crawl 100 web pages, how long does it take to crawl 500 web pages? Then, how do you speed up the process, and what is the resulting performance after speeding up?

Using the original sample code and my own computational resources, the crawling process takes 64.5 seconds to crawl 100 pages and 2394.28 seconds to crawl 500 pages. To speed up the crawling process, I applied multithreading on the crawler. The multithreading technique can utilize the CPU more while improving the speed of the overall crawling process. I tried different threading settings, and the results are shown below:

Settings	t_{100}	t_{500}
Baseline	64.5	2394.28
4 threads	27.8 (-56%)	838.16 (-65%)
16 threads	11.1 (-82%)	342.46 (-85%)

Normally, to implement multithreading, the sequence of input is already fixed. However, the set of input in our case is not fixed since we may append crawled links hiding in different

websites. Therefore, I defined a new function called $deque\_popleft(deque, num\_workers)$ to do multiple $popleft()$ operations in a single time. This way can allow me to process the threads with a limited set.

```
def deque_popleft(deque, num_workers):
    results = []
    for _ in range(num_workers):
        if deque:
            results.append(deque.popleft())
    return results
```

For the multithreading process, I used the class $ThreadPoolExecutor$ to handle the jobs. The crawling loop can be implemented in this way:

```
with ThreadPoolExecutor(max_workers=self.num_workers) as ex:
    end_flag = False
    while queue:
        processing_url = deque_popleft(queue, self.num_workers)
        future = [
            ex.submit(self.scrape, url, depth)
            for url, depth in processing_url
        ]
        for f in as_completed(future):
            ...
```

Lastly, I realize there are hangs while running this line:

```
resp = requests.get(url=url, timeout=5)
```

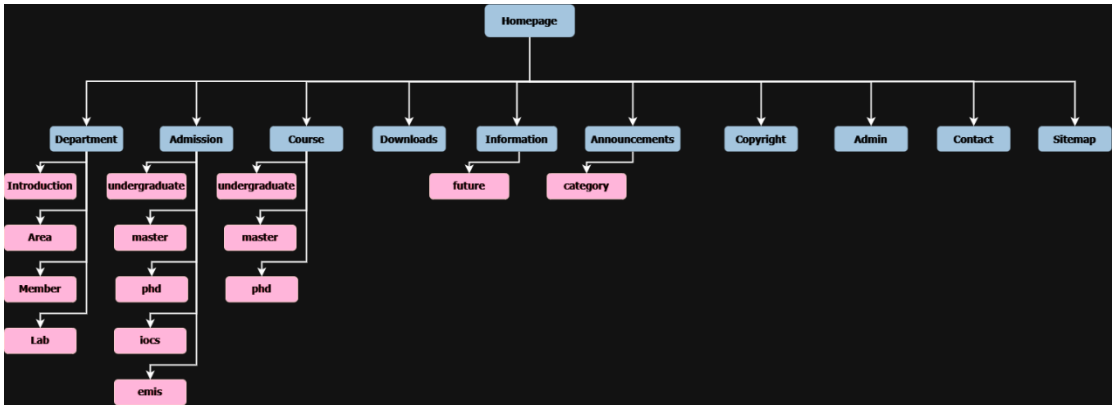
This happens because some websites are not responding to me and therefore I add a value for the parameter $timeout$.

After all the abovementioned modifications, the web crawler

can run smoothly and fast enough.

2. (15%) Based on the URLs you have crawled, analyze the structure and layout of the website, and draw an approximate sitemap of the website.

For the website of csie(https://www.csie.ncu.edu.tw), it has 10 categories for its pages: department, admission, course, downloads, information, announcements, copyright, admin, contact and sitemap. There are loops of access in the whole website as all websites are linked to the homepage.



3. (20%) Please compare the differences in query results using different query methods. Please attempt to use embedding models for vector search.

For this question, I will always use “中央大學 資工系” as the query input. For single field searching, “content” is the default field.

1. *match*:

```
Hung-Hsuan Chen - National Central University: http://www.ncu.edu.tw/~hhchen/
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/page/10/category/演講公告
系所介紹 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/department/introduction
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/1960a9ea69c55af4cd50c740fc73b9fa
Prof. Jorng-Tzong Horng: https://sites.google.com/db.csie.ncu.edu.tw/fdblab/home
系所介紹 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/department
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/8201aab7e5ccb25cbd13cc6ea9af3474
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/page/3/category/演講公告
NCU - HSCC: https://hscc.csie.ncu.edu.tw/
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/category/演講公告
```

2. *match_bool_prefix*:

```
Hung-Hsuan Chen - National Central University: http://www.ncu.edu.tw/~hhchen/
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/page/10/category/演講公告
系所介紹 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/department/introduction
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/1960a9ea69c55af4cd50c740fc73b9fa
Prof. Jorng-Tzong Horng: https://sites.google.com/db.csie.ncu.edu.tw/fdblab/home
系所介紹 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/department
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/8201aab7e5ccb25cbd13cc6ea9af3474
NCU - HSCC: https://hscc.csie.ncu.edu.tw/
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/category/演講公告
人工智慧與知識系統實驗室 AI&KSLab: https://www.ncuksl.com/
```

3. *match_phrase*:

```
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/page/4/category/徵才訊息
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/8201aab7e5ccb25cbd13cc6ea9af3474
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/1960a9ea69c55af4cd50c740fc73b9fa
H.T. Yang: https://sites.google.com/stonybrook.edu/haotyang/%E4%B8%AD%E6%96%87?authuser=0
```

4. *match_phrase_prefix*:

```
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/page/4/category/徵才訊息
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/8201aab7e5ccb25cbd13cc6ea9af3474
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/1960a9ea69c55af4cd50c740fc73b9fa
H.T. Yang: https://sites.google.com/stonybrook.edu/haotyang/%E4%B8%AD%E6%96%87?authuser=0
```

5. *combined_fields* with title and content:

```
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/page/10/category/演講公告
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/1960a9ea69c55af4cd50c740fc73b9fa
Hung-Hsuan Chen - National Central University: http://www.ncu.edu.tw/~hhchen/
系所介紹 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/department/introduction
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/8201aab7e5ccb25cbd13cc6ea9af3474
系所介紹 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/department
Prof. Jorng-Tzong Horng: https://sites.google.com/db.csie.ncu.edu.tw/fdblab/home

        國立中央大學-資訊電機學院 - College of Electrical Engineering & Computer Science
: http://www.ceecs.ncu.edu.tw/Academic_Journal.aspx
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/page/3/category/演講公告
公告 - 國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/announcement/category/演講公告
```

6. *multi_match* with title and content:

```
中央資工系學會: https://www.facebook.com/NCUCSIE?fref=ts
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#About
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#Department
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#Student_Info
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#Regulation_form
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#International exchange
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#C
```

Since our input can be in Chinese, the standard analyzer of

elasticsearch cannot tokenize the Chinese input properly. Each Chinese words can be expressed as a phrase. For example, “中央大學” will be tokenized as “中”, ”央”, ”大”, ”學” rather than “中央大學”. Using match_phrase or match_phrase_prefix can force elasticsearch to search the whole phrase “中央大學” in the text, which can lead to better results. Eventually, using a better analyzer should be the best solution for this issue. However there is also limitation like “中央大學” can be tokenized as “中央”, ”大學” even for Chinese tokenizer.

combined_fields and multi_match are similar but the former one has more limitations where the data types of query fields must be same.

I have attempted to use embedding model to perform vector search. I use all-MiniLM-L6-v2 as the embedding model and encode both title and content of website into text vectors. The dimension of the text vectors is 384.

Since I am adding new properties to the data structure, the mapping also needs to be edited.

The new mapping with text vectors:

```
"content_vector": {
    "type": "dense_vector",
    "dims": 384
},
"title_vector": {
    "type": "dense_vector",
    "dims": 384
}
```

The scripts for computing the text vectors are in

text_embedding. py

The results of knn search with title vectors:

```
NCU中央大學校外實習網: https://reurl.cc/mrGDAG
國立中央大學校園徵才網: https://campustalent.careercenter.ncu.edu.tw/orientation/student
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#Department
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#About
國立中央大學通訊工程學系: http://www.ce.ncu.edu.tw
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#Regulation_form
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#Student_Info
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#International exchange
國立中央大學資訊工程學系: https://www.csie.ncu.edu.tw/#C
```

The results of knn search with content vectors:

```
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2Fe63e754c6f9a30bda6881bd6f50c4da7
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2F4da730ed34fbe6633eeb57f6e1ed99ba
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2Fdd3489f81ec3c4488d718eb7ab893465
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2Fe411dd9a5167c20d56225ac8e2382e73
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2F41ce365be431b3edf8de84daa4d79336
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2F2f4913b60aecd864ad81ef818ead9673
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2Fb34a61c7cac04f197e5c2748daefd244
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2F8201aab7e5ccb25cbd13cc6ea9af3474
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2F901c5fd0b3cb47a2a5559556c0b61997
登入 Facebook | Facebook: https://www.facebook.com/sharer/sharer.php?u=http%3A%2F%2Fwww.csie.ncu.edu.tw%2Fannouncement%2F46c655cf03353209e33be8576f6031cc
```

The results of knn search with title vectors are much better

than with content vectors.