



ASSIGNMENT 1: CRAWLER

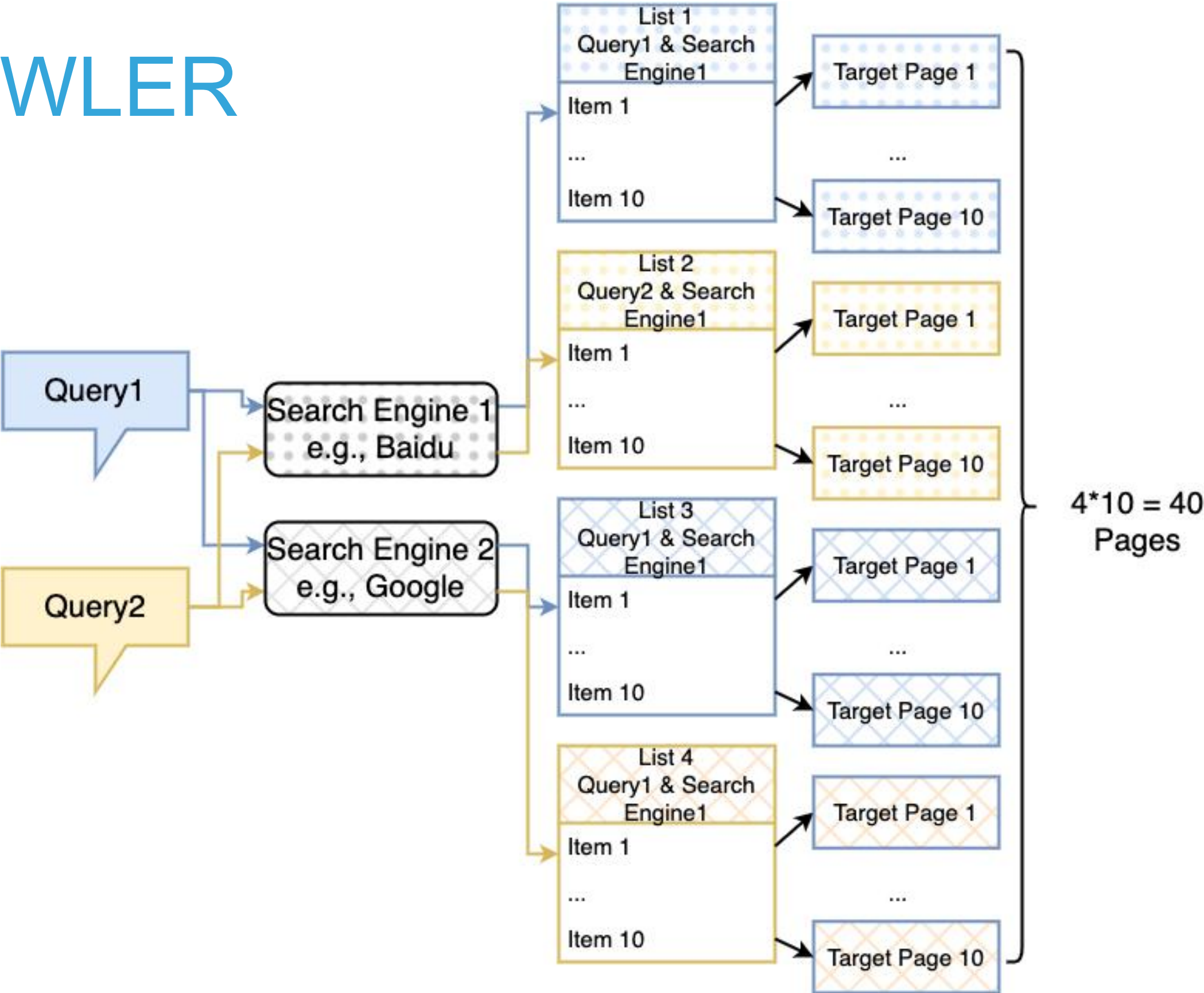
2023.2.28

WEB IR

CRAWLER

- ▶ Write a simplified prototype crawler to collect search engine results.
 - ▶ Design **2 queries** (at least 2 query terms for each query)
 - ▶ Submit each query to **2 Commercial Search Engines**
 - ▶ Google, Bing, Sogou
 - ▶ Or any other SE (that could search for *English results only*)
 - ▶ Collect the **top 10 results** in SERP (Search Engine Result Page) for each query
 - ▶ *Skip the Ads results*
 - ▶ *Don't save multimedia data*
 - ▶ $2 Q * 2 SE * 10 = 40$ landing pages
 - ▶ *If the results of the two search engines are the same, no need to repeat crawling 10 landing pages, but need to submit SERP results (json and screenshot) for both SE respectively.*

CRAWLER



DOCUMENTS TO BE SUBMITTED

► 4 kinds of files

- A Query design (query, description) file
- 4 SERP screenshots (2 for each query)
- 4 Search engines' result info (title, result link, text snippet) files, 2 per query
 - *Format each line contains result in JSON*
- 40 Targeting landing pages, one file per page
 - *Textual data only, including html/htm and pdf files, but dynamic card, img and video in the page is not required*
 - *Compress the targeting pages in one zip file*
- A README file (optional)
 - *Honor code, Extra explanation on your code, Advice and suggestions for the assignment...*

+ **source code** (Java, JavaScript, Perl, Python, PHP, Scala,)

DOCUMENTS TO BE SUBMITTED

- ▶ HW1_studentID.zip
 - ▶ QD_studentID.json (1 query design file)
 - ▶ SE_SearchEngineName_QueryNumber_studentID.json (4 SERP info files)
 - ▶ *e.g. SE_BAIDU_1_2022XXXXXX.json*
 - ▶ SE_SearchEngineName_QueryNumber_studentID.png (4 SERP screenshots)
 - ▶ *e.g. SE_BAIDU_1_2022XXXXXX.png*
- ▶ Target_Pages.zip (1 zip file)
 - ▶ TP_SearchEngineName_queryNo_rankNo_studentID.html
 - ▶ *e.g. TP_BAIDU_1_1_2022XXXXXX.png*
 - ▶ 10 targeting landing pages * 2 SEs * 2 queries
- ▶ source code (crawler, 1 zip file or 1 code file)
- ▶ README.txt(.md,.pdf...), optional

SCORING DETAILS

- ▶ Different code languages and third-party libraries are allowed.
- ▶ If you have difficulties with writing a crawler code, you can manually download all info and pages. **(-15% of total scores)**.
- ▶ Please follow the requirements on the previous page to organize your submissions. Disordered file trees and naming will result in certain points deduction. **(-1~10% of total scores)**
- ▶ Late submission of assignments will result **in a 10% reduction in total scores per day**, with a maximum delay of no more than one week.
 - ▶ For special cases, please contact the TA.
- ▶ Other cases of non-compliance(e.g., lack of files) will lead to deduction of total scores, depending on the circumstances.
- ▶ You are encouraged to ask or answer others' questions or ask for advice. Simply copy and pasting code from other students or from the Internet is strictly **not allowed**.

DETAILS

HW 1

STEP 1: QUERY DESIGN

► QD_studentID.json (1 query design file)

- Design 2 queries (based on your true information need in past several months)
 - Query (to submit to the SE)
 - Description (to show your information need in detail so that the others could understand what you want)

Query	Description
cheap flights Beijing to Dalian	Want to know the cheap flights to go to Dalian from Beijing
merge list of lists python	Find the function call or calls to merge a list containing several other list in the programing language python
Python or Ruby ?	I am looking for pro and cons, and a comparison, of the two programming language Python and Ruby
Crawler Python Google	Want to know about how to write a crawler for Google in python

* please do not use the same query as the examples.

STEP 1 DETAILS: QUERY DESIGN FILE

- ▶ File name convention QD_studentID.json

QD_2016999999.json

- ▶ Each line in the file is a JSON Object which contains following fields:
 - queryNum – Number (start from 1)
 - query - String
 - description - String

Example output for queries on previous slide

```
{"queryNum": 1, "query": "cheap flights Beijing to Dalian", "description": "Want to know the cheap flights to go to Dalian from Beijing"}
```

```
{"queryNum": 2, "query": "merge list of lists python", "description": "Find the function call or calls to merge a list containing several other list in the programming language python"}
```

STEP 2: SERP INFO

- ▶ Submit 2 queries to 2 Search engines.
 - A query has to contain at least 2 terms (more is better)
 - Search engines are Google, Bing, Sogou. Or any other SE
 - Save information of the top 10 results (for each query, at each SE) in a file, where each line represents one result from SE result page
 - Make screenshot of the search engine's result page
 - 4 json + 4 screenshots.

STEP 2 DETAILS: SERP INFO FILE

- ▶ File name convention SE_“Search engine name”_“query number”_studentID.json
SE_BING_1_2016999999.json

Search engine names options: Google, BING, Sogou ...

- ▶ Each line in the SERP info file is a JSON Object which contains following fields:
 - rank – Number (start from 1)
 - title - String
 - url - String



Ads should be ignored (in ranking, and the above information is not required for ads.)

Only text information is required. Do not save multimedia data.

STEP 2 EXAMPLE: BING

► Query 1 Tsinghua University

SE_BING_1_2016999999.json

```
{“rank”: 1, “title”: “Tsinghua University”, “url”:  
“www.tsinghua.edu.cn/publish/thu2018en/index.html”}
```

```
{“rank”: 2, “title”: “Tsinghua University - Wikipedia”, “url”:  
“https://en.wikipedia.org/wiki/Tsinghua_University”}
```

```
{“rank”: 3, “title”: “Tsinghua University World University Rankings | THE”,  
“url”: “https://www.timeshighereducation.com/.../tsinghua-university”}
```



Please make sure that the results in the json file correspond to the results in the screenshot.

For Bing, please click “国际版” to get results in English

国内版 国际版

tsinghua university

All Images Videos

翻译成中文 关闭取词

What would you like to know about this university?

tsinghua university acceptance rate tsinghua university library tsinghua university admission

Tsinghua University
www.tsinghua.edu.cn/publish/thu2018en/index.html
 Qiu Yong meets CUHK President Rocky S. Tuan and signs MOU on Undergraduate Dual Degree Programs
 2019.03.01

International Students
 Overview. Tsinghua University has been active in promoting the internationalization of ...

Undergraduate
 Tsinghua University offers undergraduate students a "liberal education of breadth": the ...

Graduate
 Graduate education stands at the very heart of Tsinghua's mission as an open ...

General Information
 Tsinghua University was established in 1911, originally under the name "Tsinghua ...

Faculty & Staff
 Faculty & Staff. Learn about the Tsinghua faculty and staff, and what it's like to be a ...

Contact
 E-mail: iso@tsinghua.edu.cn, issc@tsinghua.edu.cn For Official Visits ...

Search results from tsinghua.edu.cn

Tsinghua University - Wikipedia
https://en.wikipedia.org/wiki/Tsinghua_University
Location: Haidian District, Beijing, People's Rep... **Campus:** Urban, 395 hectares (980 acres)
Mascot: Curator the Scholar Cat (unofficial mas... **Students:** 36,300

Overview History Academics Student life Campus People


Tsinghua University is a major research university in Beijing, and a member of the elite C9 League of Chinese universities. Since its establishment in 1911, it has graduated numerous Chinese leaders in politics, business, academia, and culture. Reflecting its motto of Self-Discipline and Social Commitment, Tsinghua University is dedicated to academic excellence, advancing the well-being of Chinese society, and global development. Tsinghua is perennially ranked as one of the top academic institut

See more on en.wikipedia.org · Text under CC-BY-SA license

Tsinghua University World University Rankings | THE
<https://www.timeshighereducation.com/.../tsinghua-university>
 About **Tsinghua University** The campus of Tsinghua University is situated on the site of the former imperial gardens of the Qing Dynasty, and surrounded by a number of historical sites in northwest Beijing.


Please make sure that the results in the json file correspond to the results in the screenshot.
(delete the results whose class do not contain "b_algo" in your screenshot)

News about Tsinghua University




Chinese university students disciplined over rainbow flags file lawsuit against Education ...

South China Morning Post · 4d · on



Chinese university students sue education ministry after being disciplined over ...

Channel NewsA... · 4d



Want To Study Abroad? The 2023 Best Universities In Europe According To ...

Forbes · 10d

Tsinghua University in China - US News Best Global Universities
<https://www.usnews.com/.../tsinghua-university-503146> ✓
 Web Tsinghua University housing is available for both undergraduate and graduate student
 The university comprises numerous schools and departments, which offer programs ...

Tsinghua University : Rankings, Fees & Courses Details ...
<https://www.topuniversities.com/universities/tsinghua-university> ▾
 Web Tsinghua University was established as a preparatory school for the students' trips ab
 Today, Tsinghua's motto of "self-discipline and excellence" has taken it far. Most ...

Tsinghua University has been active in promoting the internationalization of its ...

Starting from 2017, Tsinghua begins to integrate programs/majors into divisions...

See results only from tsinghua.edu.cn

Tsinghua University : Rankings, Fees & Courses Details ...

<https://www.topuniversities.com/universities/tsinghua-university>

Tsinghua University was established in 1911 in the wake of the anti-colonialist Boxer Rebellion, which saw the US fine China \$30m as punishment. In 1909, President Theodore...

QS Top 50 Under 50 2020 NanYang Technological

Beijing QS Best Student Cities ranking: 32nd Home

Top Electrical Engineering Schools in 2015 **Click

Tsinghua University is a major public research university and a member of the C9 League. It is also a member of Double First Class University Plan, Project 985 and Project 211. Since its establishment in 1911, it has produced many notable leaders in science, engineering, politics, business, academia, and culture.

Wikipedia Instagram Facebook YouTube LinkedIn

School colors: White · Purple

Mascot: Tsinghua University Qing Xiaohua

正在发言: MinZhang

Chat with Bing

67个匹配

```

<li class="b_algo">
  <div class="b_title">
    <h2>Tsinghua University : Rankings, Fees & Courses Details ...</h2>
    <div class="b_surfix b_secondary_text nowrap scs_exp_sizz3 b_loadtr">
      <a class="scs_icn b_hide" tabindex="1" aria-expanded="False" aria-label="Additional Results" href="javascript:void(0)" h="ID=SERP,5170.1">
        <span class="scs_cls b_hide" tabindex="0" aria-label="Show less"></span>
        <span class="scs_arw" tabindex="0" aria-label="Show more"></span>
      </a>
      <div data-priority="2" data-sc-metadata="{&quot;entity&quot;:&quot;Tsinghua University&quot;,&quot;scenarios&quot;:&quot;2,6&quot;,&quot;url&quot;:&quot;https://www.topuniversities.com/universities/tsinghua-university&quot;}" data-sc-iid="SERP.5621" class="scs_c scs_load b_hide cnt_vis_hid"></div>
    </div>
    <div class="b_caption">
      <div class="b_attribution" u="1|5062|4798526748491864|kup0Z81WBCeGHg1STBTxQLuCd1YyiuYq">
        <cite>
          https://
          <strong>www.topuniversities.com</strong>
          /universities/
          <strong>tsinghua-university</strong>
        </cite>
        <span class="c_tlbxTrg">
          <span class="c_tlbxH" H="BASE: CACHEDPAGEDEFAULT" K="SERP,5171.1"></span>
        </span>
      </div>
    </div>
  </li>

```

```

> </li>
> <li class="b_ans b_mop b_nwsAns" h="SERP,5473.1"
data-bm="7">...</li>
> <li class="b_algo" data-bm="8">...</li>
> <li class="b_algo" data-bm="9">...</li>
> <li class="b_algo" data-bm="10">...</li>
> <li class="b_algo" data-bm="11">...</li>
> <li class="b_algo" data-bm="12">...</li>
> <li class="b_algo" data-bm="13">...</li>
> <li class="b_algo" data-bm="14">...</li>
> <li class="b_ans" h="SERP,5613.1" data-bm="15">...</li>
== $0
> <li class="b_pag" data-bm="16">...</li>
> <li id="nfs_post" class="fb_helptop" data-bm="17">

```

STEP 2 DETAILS: A TOOL FOR SCREENSHOT

► Recommendation

- An Extension in Chrome – “Full Page Screen Capture”

扩展程序

[更多扩展程序](#)

已添加



Full Page Screen Capture

提供方: mrcoles.com

Capture a screenshot of your current page in entirety and reliably—without reques

★★★★★ 27,737 开发者工具

评分

STEP 3: CRAWL TARGET PAGES

- ▶ Crawl the whole target page, for each url collected in previous step
- ▶ Do not save multimedia, only text info is required.

- ▶ File name convention

CD_SearchEngineName_queryNo_rankNo_studentID.html

TP_BING_1_1_2016999999.html, ...

TP_BING_2_1_2016999999.html, ...

TP_BAIDU_1_1_2016999999.html, ...

TP_BAIDU_2_1_2016999999.html, ...

DEADLINE:

8:00pm (UTC+08:00, Beijing Time) Mar.5th , Sunday
Submit your homework to web learning platform of thu:

[Http://learn.tsinghua.edu.cn.](http://learn.tsinghua.edu.cn)

our course section “Assignment”(课程作业)

QUESTIONS ?