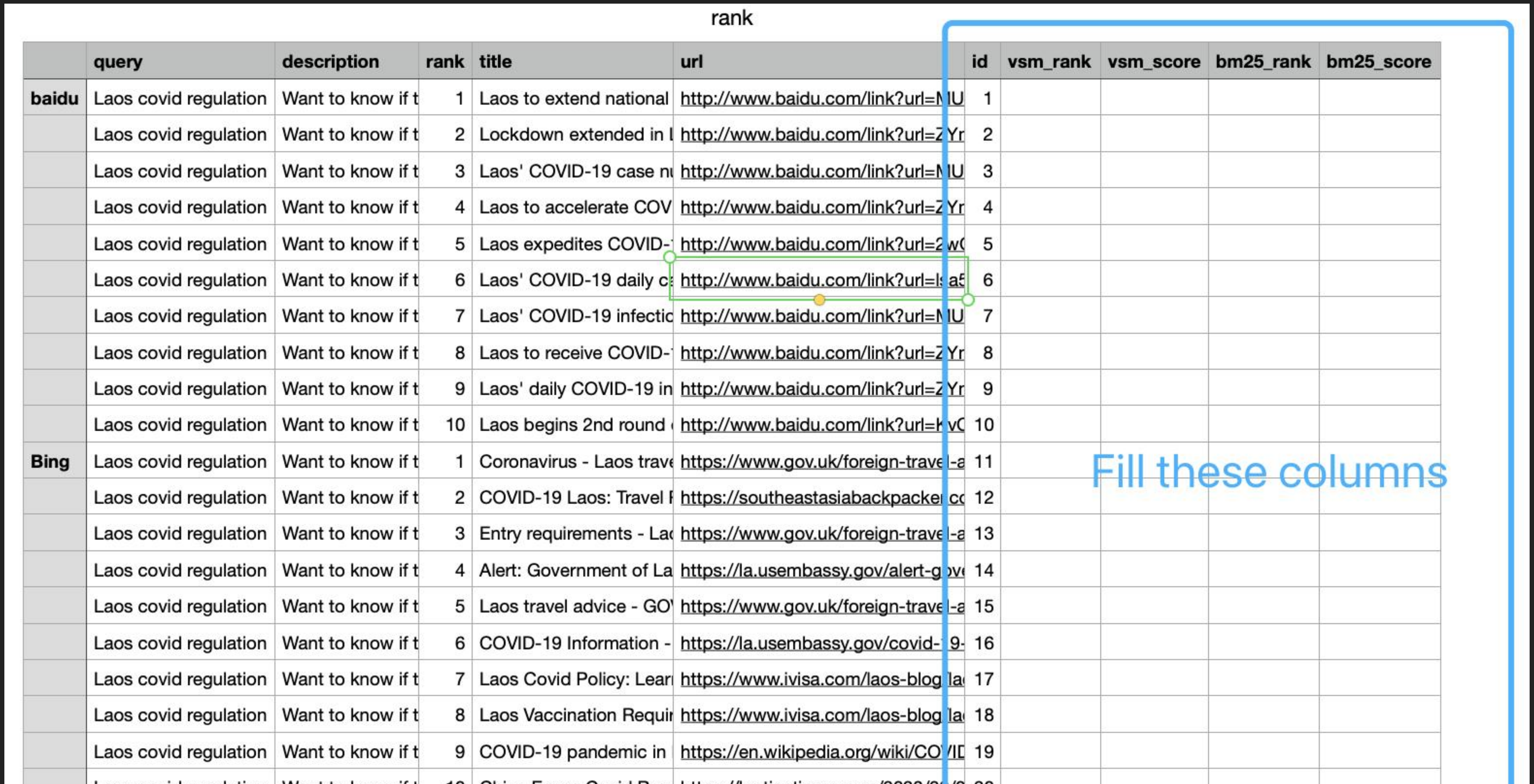
ASSIGNMENT 3: RANKING MODELS

KEY - OVERVIEW

- Develop application to rank crawled data with Vector Space Model and Okapi BM25
- Rank documents (results crawled in Homework 1)



rank

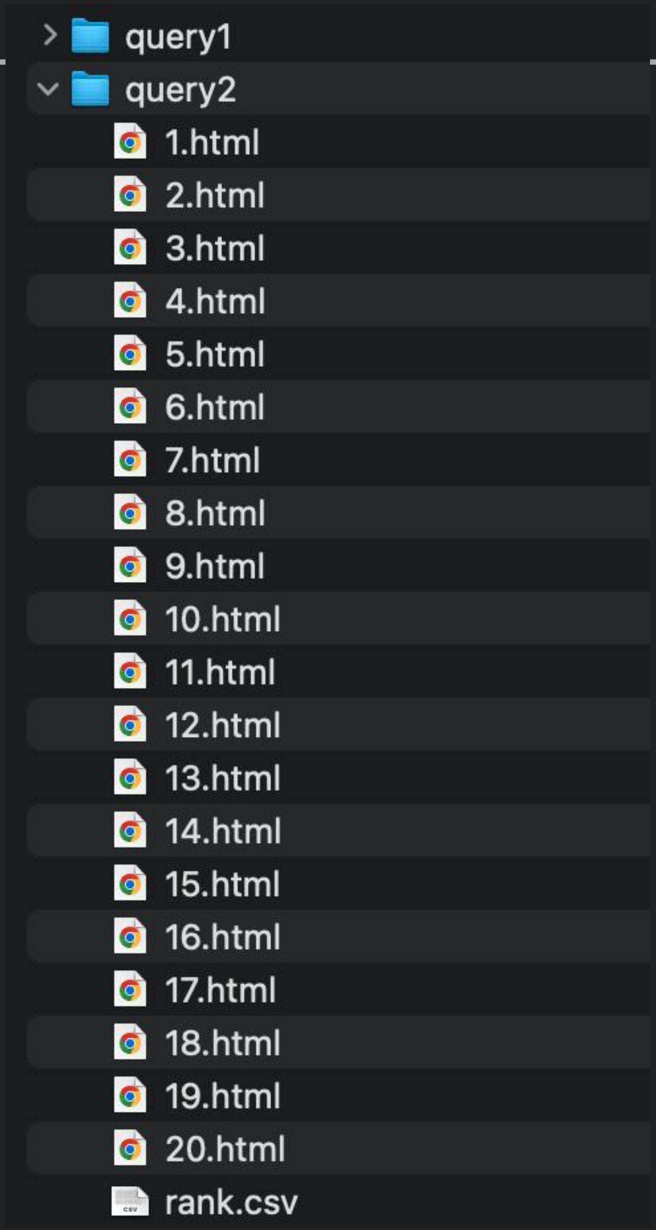
	query	description	rank	title	url	id	vsm_rank	vsm_score	bm25_rank	bm25_score
baidu	Laos covid regulation	Want to know if t	1	Laos to extend national	http://www.baidu.com/link?url=MU	1				
	Laos covid regulation	Want to know if t	2	Lockdown extended in l	http://www.baidu.com/link?url=ZYr	2				
	Laos covid regulation	Want to know if t	3	Laos' COVID-19 case nu	http://www.baidu.com/link?url=MU	3				
	Laos covid regulation	Want to know if t	4	Laos to accelerate COV	http://www.baidu.com/link?url=ZYr	4				
	Laos covid regulation	Want to know if t	5	Laos expedites COVID-	http://www.baidu.com/link?url=2wC	5				
	Laos covid regulation	Want to know if t	6	Laos' COVID-19 daily ca	http://www.baidu.com/link?url=Isa5	6				
	Laos covid regulation	Want to know if t	7	Laos' COVID-19 infectio	http://www.baidu.com/link?url=MU	7				
	Laos covid regulation	Want to know if t	8	Laos to receive COVID-	http://www.baidu.com/link?url=ZYr	8				
	Laos covid regulation	Want to know if t	9	Laos' daily COVID-19 in	http://www.baidu.com/link?url=ZYr	9				
	Laos covid regulation	Want to know if t	10	Laos begins 2nd round	http://www.baidu.com/link?url=KvC	10				
Bing	Laos covid regulation	Want to know if t	1	Coronavirus - Laos trave	https://www.gov.uk/foreign-travel-a	11				
	Laos covid regulation	Want to know if t	2	COVID-19 Laos: Travel F	https://southeastasiabackpacker.cc	12				
	Laos covid regulation	Want to know if t	3	Entry requirements - Lac	https://www.gov.uk/foreign-travel-a	13				
	Laos covid regulation	Want to know if t	4	Alert: Government of La	https://la.usembassy.gov/alert-gove	14				
	Laos covid regulation	Want to know if t	5	Laos travel advice - GO'	https://www.gov.uk/foreign-travel-a	15				
	Laos covid regulation	Want to know if t	6	COVID-19 Information -	https://la.usembassy.gov/covid-19-	16				
	Laos covid regulation	Want to know if t	7	Laos Covid Policy: Lear	https://www.ivisa.com/laos-blog/la	17				
	Laos covid regulation	Want to know if t	8	Laos Vaccination Requir	https://www.ivisa.com/laos-blog/la	18				
	Laos covid regulation	Want to know if t	9	COVID-19 pandemic in	https://en.wikipedia.org/wiki/COVID	19				
	Laos covid regulation	Want to know if t	10	[illegible]	[illegible]	20				

Fill these columns

REQUIREMENTS

- Files :
  - for every query
  - 20 candidate documents
  - a simple python file to help read csv files(not necessary to use)

- Fill rank.csv in each query folder.
  - rank: start from 1
- Compress the result files in one ZIP archive



```
> query1
v query2
    1.html
    2.html
    3.html
    4.html
    5.html
    6.html
    7.html
    8.html
    9.html
    10.html
    11.html
    12.html
    13.html
    14.html
    15.html
    16.html
    17.html
    18.html
    19.html
    20.html
    rank.csv
```

+ Source code (C++, C#, Java, JavaScript, Perl, Python, PHP, Scala, ……)

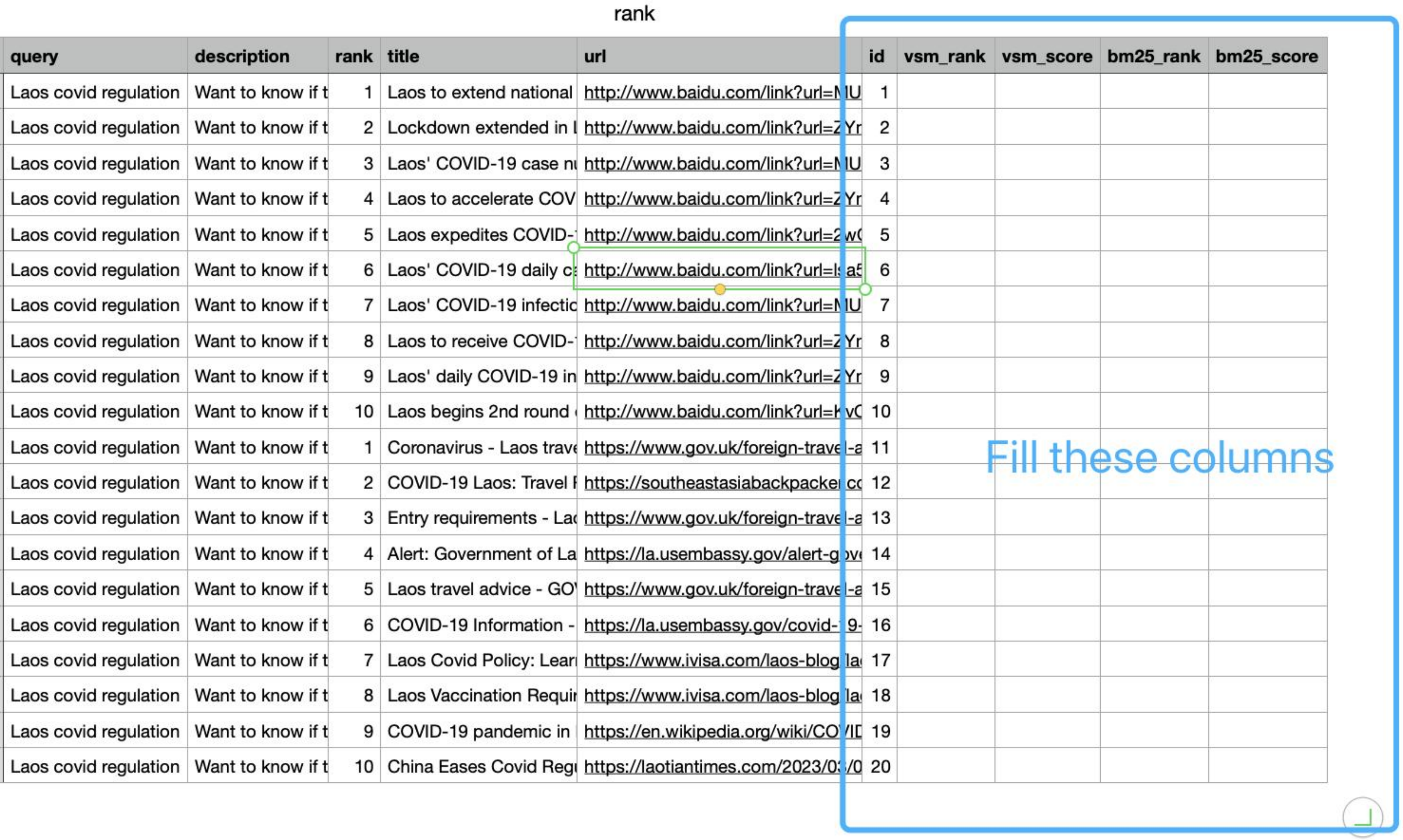
- Ranking models

- README (Optional)
  - *Introduction about implementations and data treatments, how to run the code, suggestions to the assignment…*

DETAILS

STEPS

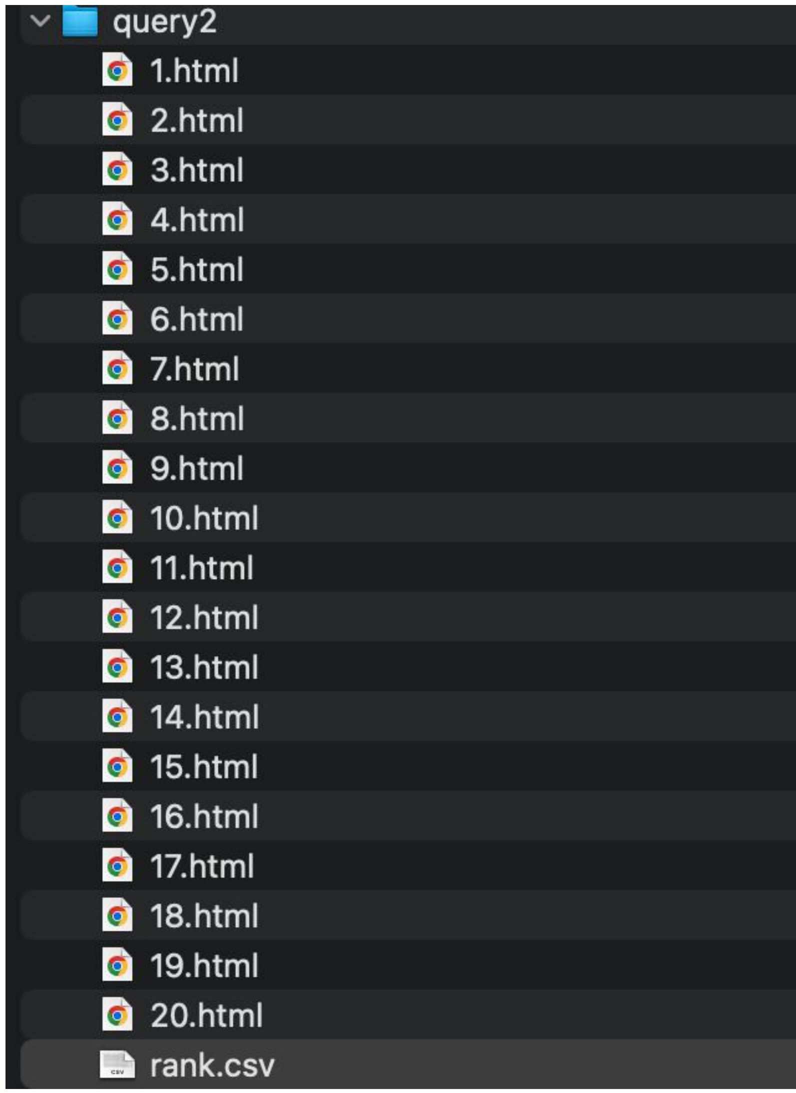
- Tare 5 folders “queryN”.
- For each query, “rank.csv” containing following fields for 20 documents
    - Id (1-20) – use this to find the html file in folder (-1 means the file is not found, skip these rows)



rank

query	description	rank	title	url	id	vsm_rank	vsm_score	bm25_rank	bm25_score
Laos covid regulation	Want to know if t	1	Laos to extend national	http://www.baidu.com/link?url=MU	1				
Laos covid regulation	Want to know if t	2	Lockdown extended in l	http://www.baidu.com/link?url=ZYr	2				
Laos covid regulation	Want to know if t	3	Laos' COVID-19 case nu	http://www.baidu.com/link?url=MU	3				
Laos covid regulation	Want to know if t	4	Laos to accelerate COV	http://www.baidu.com/link?url=ZYr	4				
Laos covid regulation	Want to know if t	5	Laos expedites COVID-1	http://www.baidu.com/link?url=2wC	5				
Laos covid regulation	Want to know if t	6	Laos' COVID-19 daily ca	http://www.baidu.com/link?url=Isa5	6				
Laos covid regulation	Want to know if t	7	Laos' COVID-19 infectic	http://www.baidu.com/link?url=MU	7				
Laos covid regulation	Want to know if t	8	Laos to receive COVID-1	http://www.baidu.com/link?url=ZYr	8				
Laos covid regulation	Want to know if t	9	Laos' daily COVID-19 in	http://www.baidu.com/link?url=ZYr	9				
Laos covid regulation	Want to know if t	10	Laos begins 2nd round (	http://www.baidu.com/link?url=KvC	10				
Laos covid regulation	Want to know if t	1	Coronavirus - Laos trave	https://www.gov.uk/foreign-travel-a	11				
Laos covid regulation	Want to know if t	2	COVID-19 Laos: Travel F	https://southeastasiabackpacker.cc	12				
Laos covid regulation	Want to know if t	3	Entry requirements - Lac	https://www.gov.uk/foreign-travel-a	13				
Laos covid regulation	Want to know if t	4	Alert: Government of La	https://la.usembassy.gov/alert-gove	14				
Laos covid regulation	Want to know if t	5	Laos travel advice - GO\	https://www.gov.uk/foreign-travel-a	15				
Laos covid regulation	Want to know if t	6	COVID-19 Information -	https://la.usembassy.gov/covid-19-	16				
Laos covid regulation	Want to know if t	7	Laos Covid Policy: Lear	https://www.ivisa.com/laos-blog/la	17				
Laos covid regulation	Want to know if t	8	Laos Vaccination Requir	https://www.ivisa.com/laos-blog/la	18				
Laos covid regulation	Want to know if t	9	COVID-19 pandemic in	https://en.wikipedia.org/wiki/COVID	19				
Laos covid regulation	Want to know if t	10	China Eases Covid Regu	https://laotiantimes.com/2023/03/0	20				

Fill these columns



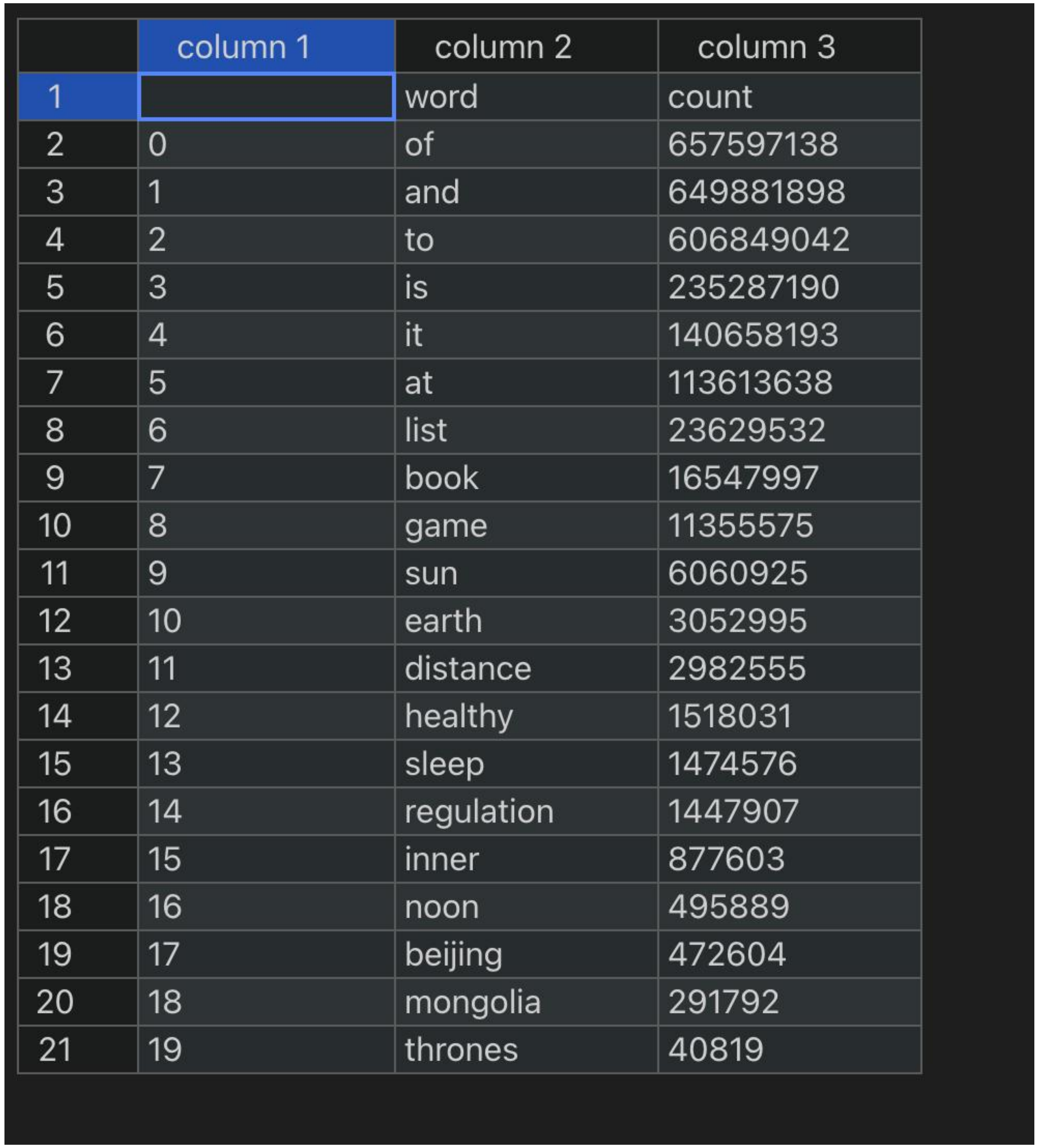
query2
1.html
2.html
3.html
4.html
5.html
6.html
7.html
8.html
9.html
10.html
11.html
12.html
13.html
14.html
15.html
16.html
17.html
18.html
19.html
20.html
rank.csv

STEPS

- Parse the html file to get the main body of the document. (Pure text, without html tag.)

- Scan the query and the document to get *tf, qtf, dl, etc.* (Note that there is no collection for your homework, so we provide a "*df.csv*" for *df* of terms in queries. And we set ***N = 100 billion***, ***avdl = 500*** for the collection.)

- Computing VSM score and BM25 score for each document.



	column 1	column 2	column 3
1		word	count
2	0	of	657597138
3	1	and	649881898
4	2	to	606849042
5	3	is	235287190
6	4	it	140658193
7	5	at	113613638
8	6	list	23629532
9	7	book	16547997
10	8	game	11355575
11	9	sun	6060925
12	10	earth	3052995
13	11	distance	2982555
14	12	healthy	1518031
15	13	sleep	1474576
16	14	regulation	1447907
17	15	inner	877603
18	16	noon	495889
19	17	beijing	472604
20	18	mongolia	291792
21	19	thrones	40819

NOTE

- Skip the rows with id=-1 (html not found)
- Some pages have a few Chinese characters, you can process it or just delete it.
- The main part of the algorithms should be implemented by yourself without simply using third-party packages.
- Although the results may not be exactly the same for each individual due to differences in data processing, correctness will be checked in a number of ways, such as whether there are consistent scores for the same documents and whether the ranking results match the general trend in all submissions.
  - You can indicate special treatments that may cause the results to differ from the general trend in README.

SUBMISSION

- HW3_StudentID.zip
  - query1/rank.csv
  - query2/rank.csv
  - …
  - query5/rank.csv

+ Source code (C++, C#, Java, JavaScript, Perl, Python, PHP, Scala, ……)
  - Details of VSM & BM25
- README(Optional)

DEADLINE:

10:00am (UTC+08:00, Beijing Time) Mar. 26, Sunday

Submit your homework to web learning platform of thu:

Http://learn.Tsinghua.Edu.Cn our course section “Assignment”(课程作业)

Also you can begin considering proposal for the final project ☺

QUESTIONS ?