Programming HW 1

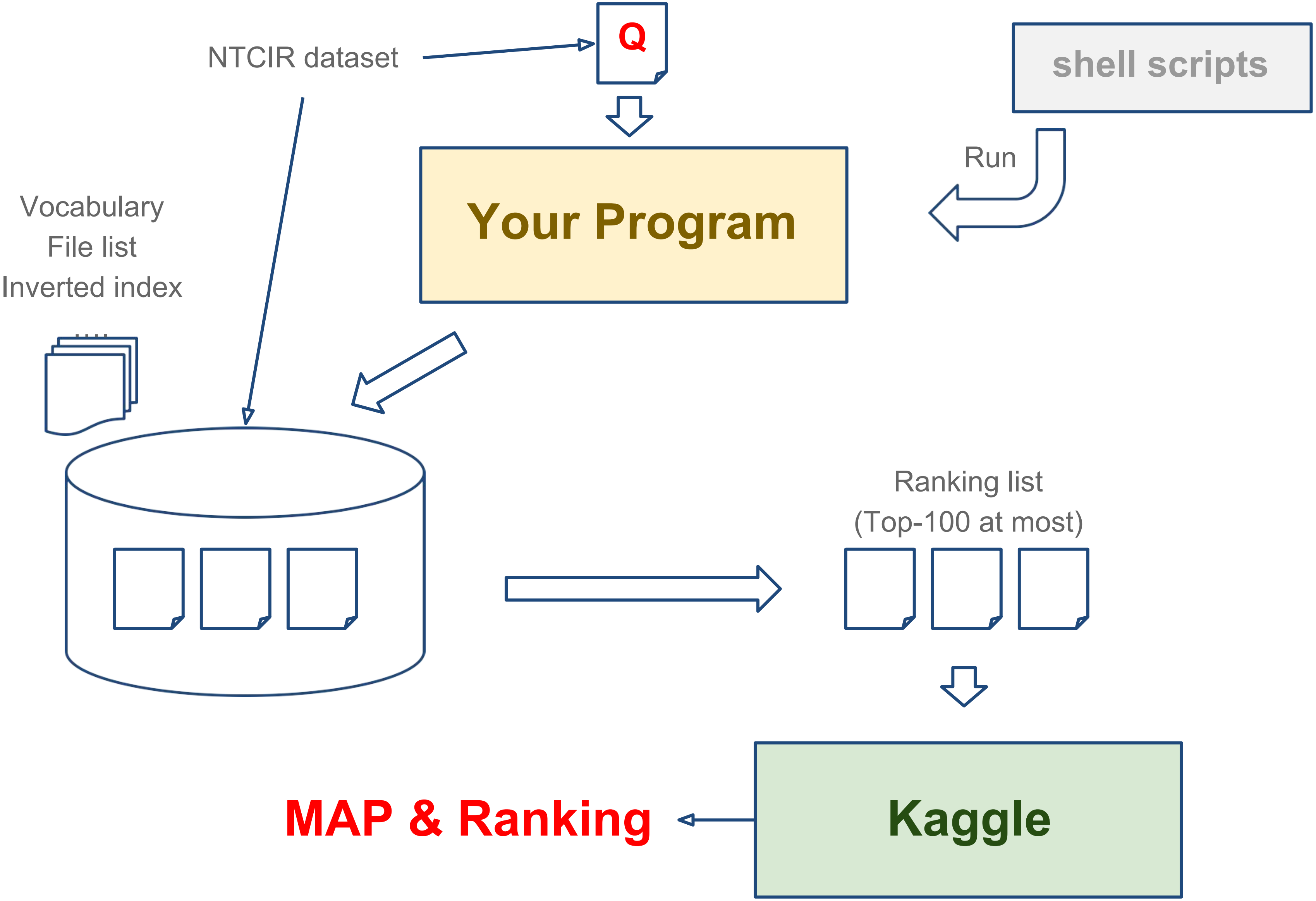
Web Retrieval and Mining spring 2018

Introduction

- In this homework, you are asked to implement **a small information retrieval system**.

- We will give you a bunch of Chinese news articles and several queries in NTCIR format, and your task is to find the relevant documents among these articles according to the given queries.

- You should implement the retrieval system by **Vector Space Model** (VSM) with **Rocchio Relevance Feedback** (pseudo version).



NTCIR dataset

Q

shell scripts

Run

Your Program

Vocabulary
File list
Inverted index
....

Ranking list
(Top-100 at most)

MAP & Ranking

Kaggle

Kaggle

https://www.kaggle.com/c/ntucsie-wm2018-vsm

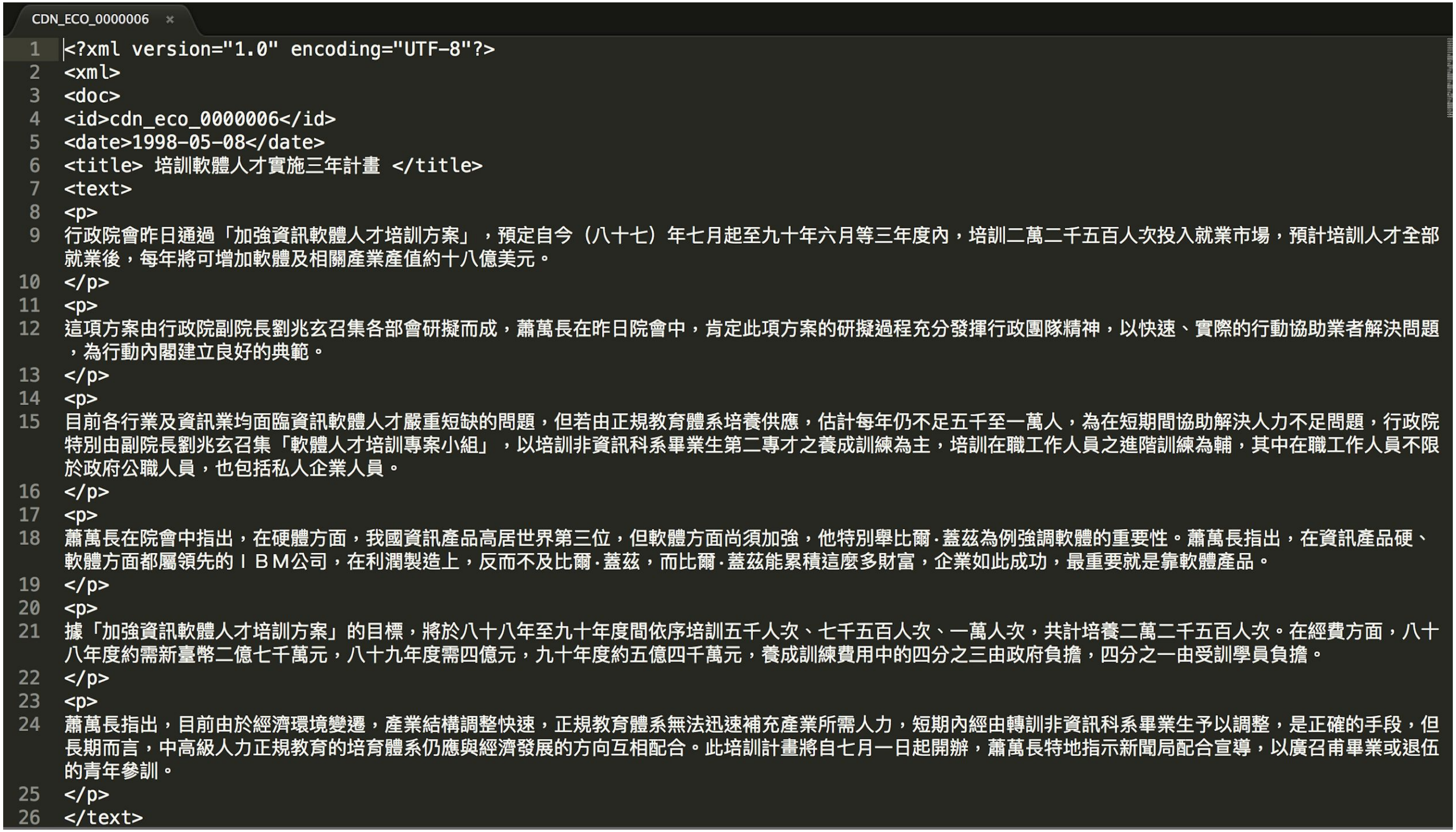
Data

1. CIRB010.tar.gz
2. model.tar.gz
3. user_aggreement_form_for_ntcir.pdf
4. query_train.xml
5. ans_train.csv
6. query_test.xml
7. script.tar.gz
8. sample_submission.csv

NTCIR Document Format

- The NTCIR document format conforms to XML 1.0
- The root element is <xml>, it contains exactly one <doc> tag.
- A <doc> tag represents exactly one newswire article, in which several sub-elements are used to specify different type of information:
  - <id>: An unique document ID.
  - <date>: The publication date.
  - <title>: The title of the article.
  - <text>: The content of the article, which may include one or more passages enclosed in <p> tags.

CIRB010.tar.gz



```
<?xml version="1.0" encoding="UTF-8"?>
<xml>
<doc>
<id>cdn_eco_0000006</id>
<date>1998-05-08</date>
<title> 培訓軟體人才實施三年計畫 </title>
<text>
<p>
行政院會昨日通過「加強資訊軟體人才培訓方案」，預定自今（八十七）年七月起至九十年六月等三年度內，培訓二萬二千五百人次投入就業市場，預計培訓人才全部就業後，每年將可增加軟體及相關產業產值約十八億美元。
</p>
<p>
這項方案由行政院副院長劉兆玄召集各部會研擬而成，蕭萬長在昨日院會中，肯定此項方案的研擬過程充分發揮行政團隊精神，以快速、實際的行動協助業者解決問題，為行動內閣建立良好的典範。
</p>
<p>
目前各行業及資訊業均面臨資訊軟體人才嚴重短缺的問題，但若由正規教育體系培養供應，估計每年仍不足五千至一萬人，為在短期間協助解決人力不足問題，行政院特別由副院長劉兆玄召集「軟體人才培訓專案小組」，以培訓非資訊科系畢業生第二專才之養成訓練為主，培訓在職工作人員之進階訓練為輔，其中在職工作人員不限於政府公職人員，也包括私人企業人員。
</p>
<p>
蕭萬長在院會中指出，在硬體方面，我國資訊產品高居世界第三位，但軟體方面尚須加強，他特別舉比爾·蓋茲為例強調軟體的重要性。蕭萬長指出，在資訊產品硬、軟體方面都屬領先的ＩＢＭ公司，在利潤製造上，反而不及比爾·蓋茲，而比爾·蓋茲能累積這麼多財富，企業如此成功，最重要就是靠軟體產品。
</p>
<p>
據「加強資訊軟體人才培訓方案」的目標，將於八十八年至九十年度間依序培訓五千人次、七千五百人次、一萬人次，共計培養二萬二千五百人次。在經費方面，八十八年度約需新臺幣二億七千萬元，八十九年度需四億元，九十年度約五億四千萬元，養成訓練費用中的四分之三由政府負擔，四分之一由受訓學員負擔。
</p>
<p>
蕭萬長指出，目前由於經濟環境變遷，產業結構調整快速，正規教育體系無法迅速補充產業所需人力，短期內經由轉訓非資訊科系畢業生予以調整，是正確的手段，但長期而言，中高級人力正規教育的培育體系仍應與經濟發展的方向互相配合。此培訓計畫將自七月一日起開辦，蕭萬長特地指示新聞局配合宣導，以廣召甫畢業或退伍的青年參訓。
</p>
</text>
```

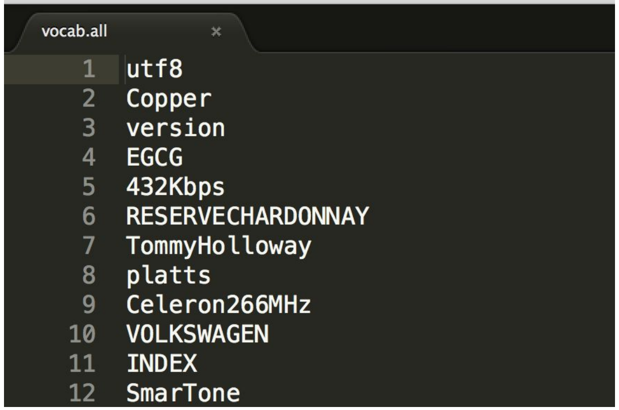
NTCIR Document Set

- Please sign up the user agreement form and hand it to the TAs (at R302 of CSIE dept) in order to use this corpus.
  - Note that you'll get no point if you don't sign up the user agreement form.

- We have indexed the NTCIR documents and produced three model files for you:
  - vocab.all
  - file-list
  - inverted-file

Format: *vocab.all*

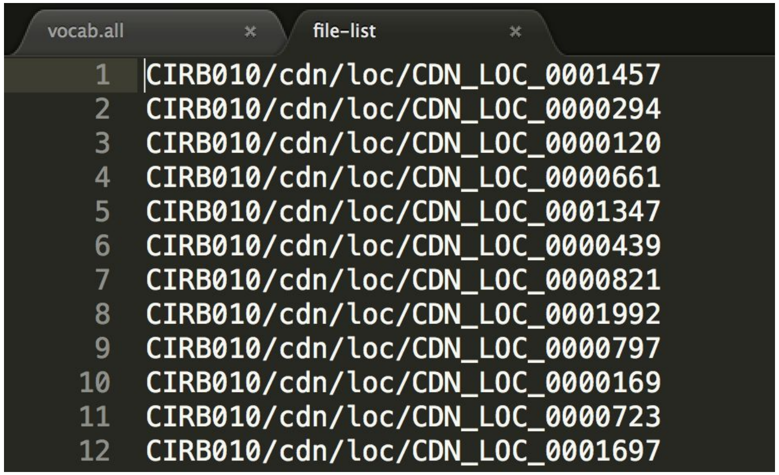
- The file contains all vocabularies in NTCIR documents.
- The first line is character encoding format.
- Each line of the following is a vocabulary.
- Vocabularies are case-sensitive.
- Each vocabulary will have a **vocab_id** according to its line number(start from 0).
  - E.g. the **vocab_id** of **"Valentine"** is **1**; **Powell** is **2**.



```
utf8
Copper
version
EGCG
432Kbps
RESERVECHARDONNAY
TommyHolloway
platts
Celeron266MHz
VOLKSWAGEN
INDEX
SmarTone
```

Format: *file-list*

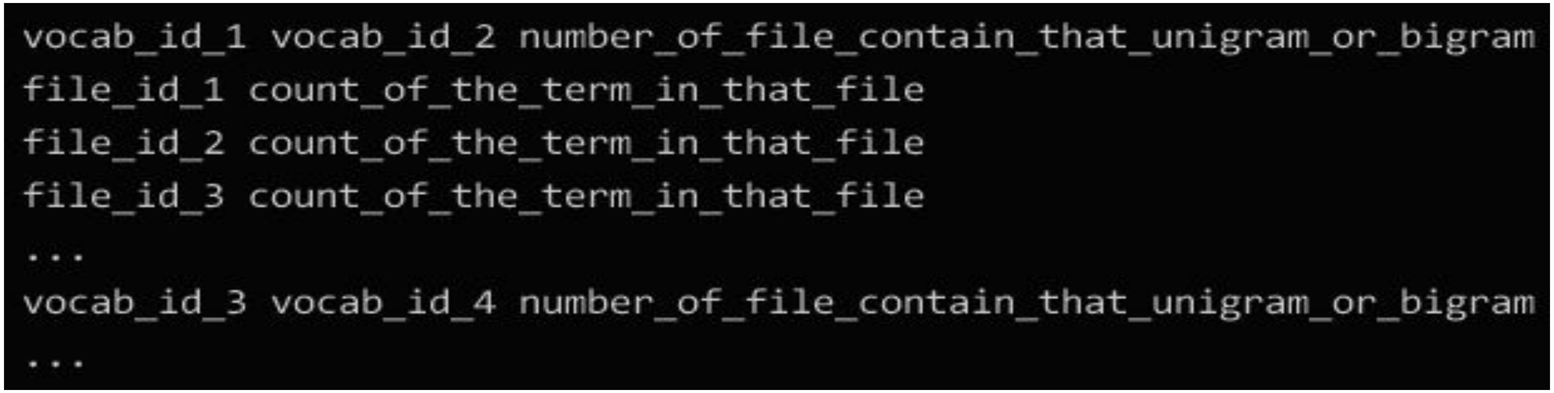
- This is a list of all NTCIR documents.
- Each line denotes a document which has its line number (start from 0) as its **file_id**.
  - E.g. **./CIRB010/cdn/chi/cdn_chi_0001457** has **file_id 0**, **./CIRB010/cdn/chi/cdn_chi_0000294** has **file_id 1**.



```
 1  CIRB010/cdn/loc/CDN_LOC_0001457
 2  CIRB010/cdn/loc/CDN_LOC_0000294
 3  CIRB010/cdn/loc/CDN_LOC_0000120
 4  CIRB010/cdn/loc/CDN_LOC_0000661
 5  CIRB010/cdn/loc/CDN_LOC_0001347
 6  CIRB010/cdn/loc/CDN_LOC_0000439
 7  CIRB010/cdn/loc/CDN_LOC_0000821
 8  CIRB010/cdn/loc/CDN_LOC_0001992
 9  CIRB010/cdn/loc/CDN_LOC_0000797
10  CIRB010/cdn/loc/CDN_LOC_0000169
11  CIRB010/cdn/loc/CDN_LOC_0000723
12  CIRB010/cdn/loc/CDN_LOC_0001697
```

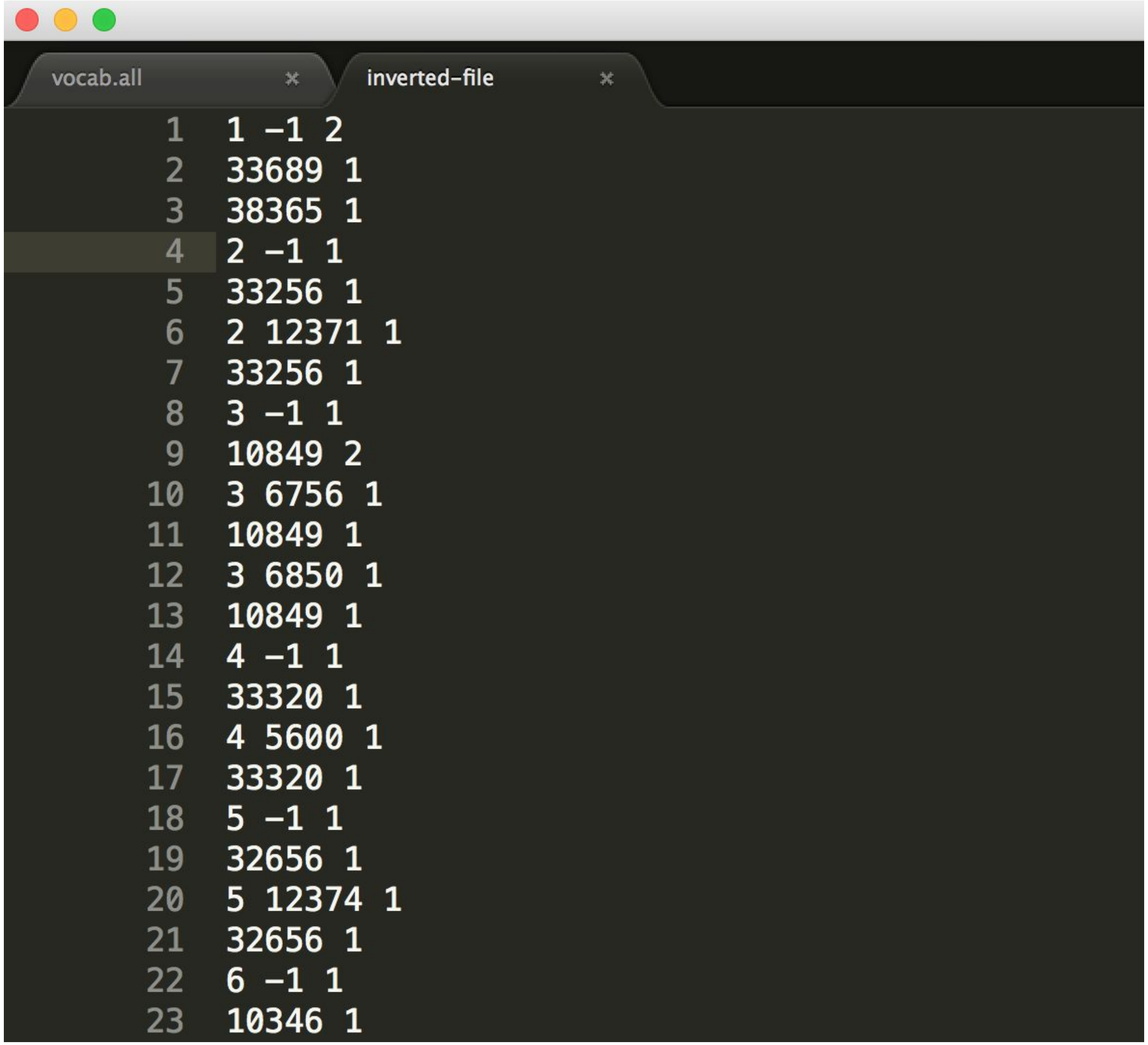
Format: *inverted-file*

- **vocab_id** and **file_id** referred from **vocab.all** and **file-list**.
- **vocab_id_1 vocab_id_2** denotes an unigram when **vocab_id_2==-1** or a bigram when **vocab_id_2**!=**-1**.
- If there are **N** files containing **vocab_id_1 vocab_id_2**, there will be the number **N** next to **vocab_id_2**, followed by **N** lines that display the counts of this term in each file.



```
vocab_id_1 vocab_id_2 number_of_file_contain_that_unigram_or_bigram
file_id_1 count_of_the_term_in_that_file
file_id_2 count_of_the_term_in_that_file
file_id_3 count_of_the_term_in_that_file
...
vocab_id_3 vocab_id_4 number_of_file_contain_that_unigram_or_bigram
...
```

inverted-file

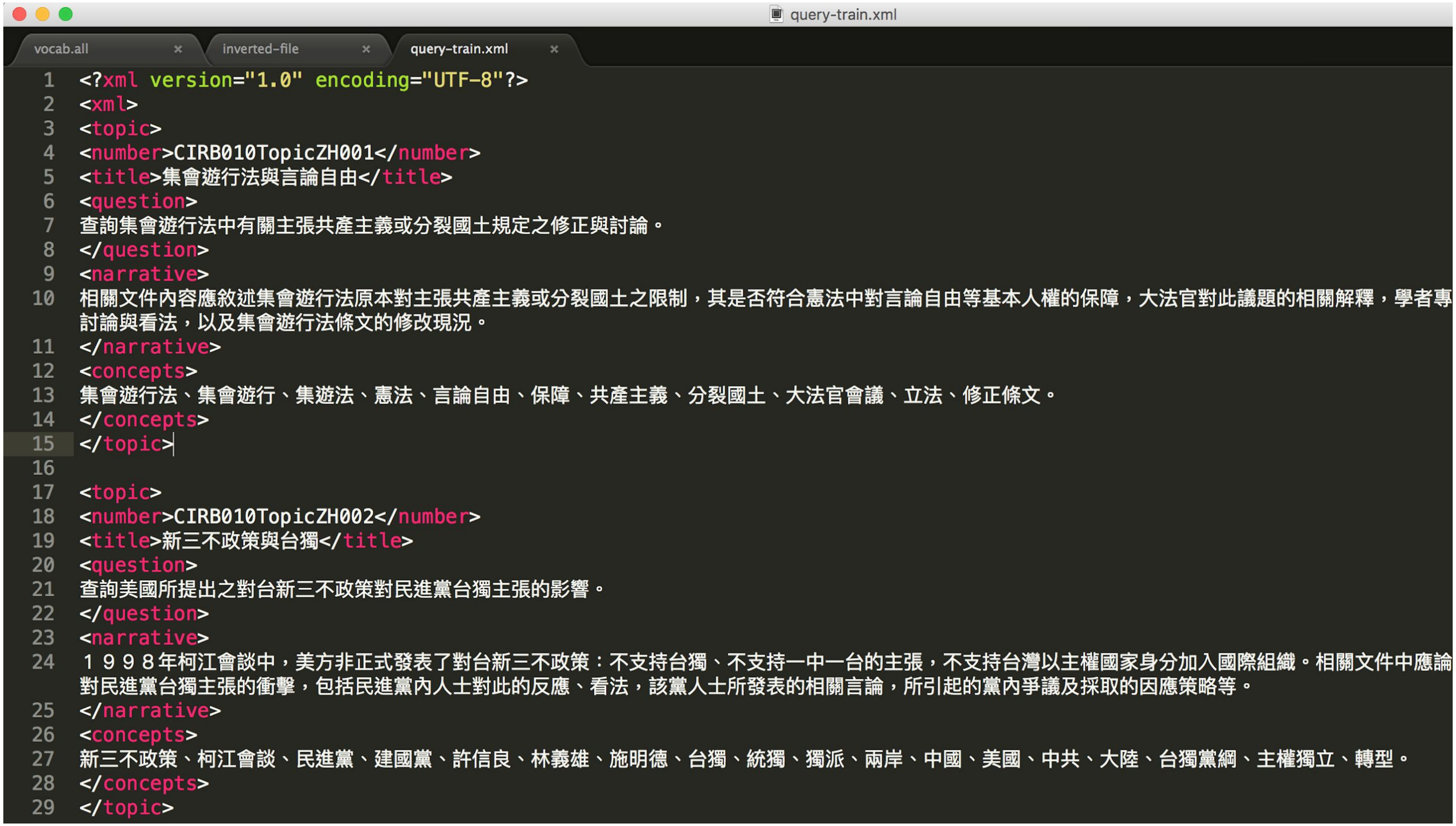


```
1 -1 2
33689 1
38365 1
2 -1 1
33256 1
2 12371 1
33256 1
3 -1 1
10849 2
3 6756 1
10849 1
3 6850 1
10849 1
4 -1 1
33320 1
4 5600 1
33320 1
5 -1 1
32656 1
5 12374 1
32656 1
6 -1 1
10346 1
```

Query File Format

- The NTCIR topic format conforms to XML 1.0
- The file contains multiple topics, each of them is enclosed in a <topic> tag. In each topic, different types of information are specified by the following tags:
    - <number>: The topic number.
    - <title>: The topic title.
    - <question>: A short description about the query topic.
    - <narrative>: Even more verbose descriptions about the topic.
    - <concepts>: A set of keywords that can be used in retrieval about the topic.
- You have to retrieve several relevant documents for each topic.
- All the content of **title**, **question**, **narrative**, and **concepts** can be used as the query of the topic, it's **your own choice** to decide which part(s) you want to use·

Query File Format



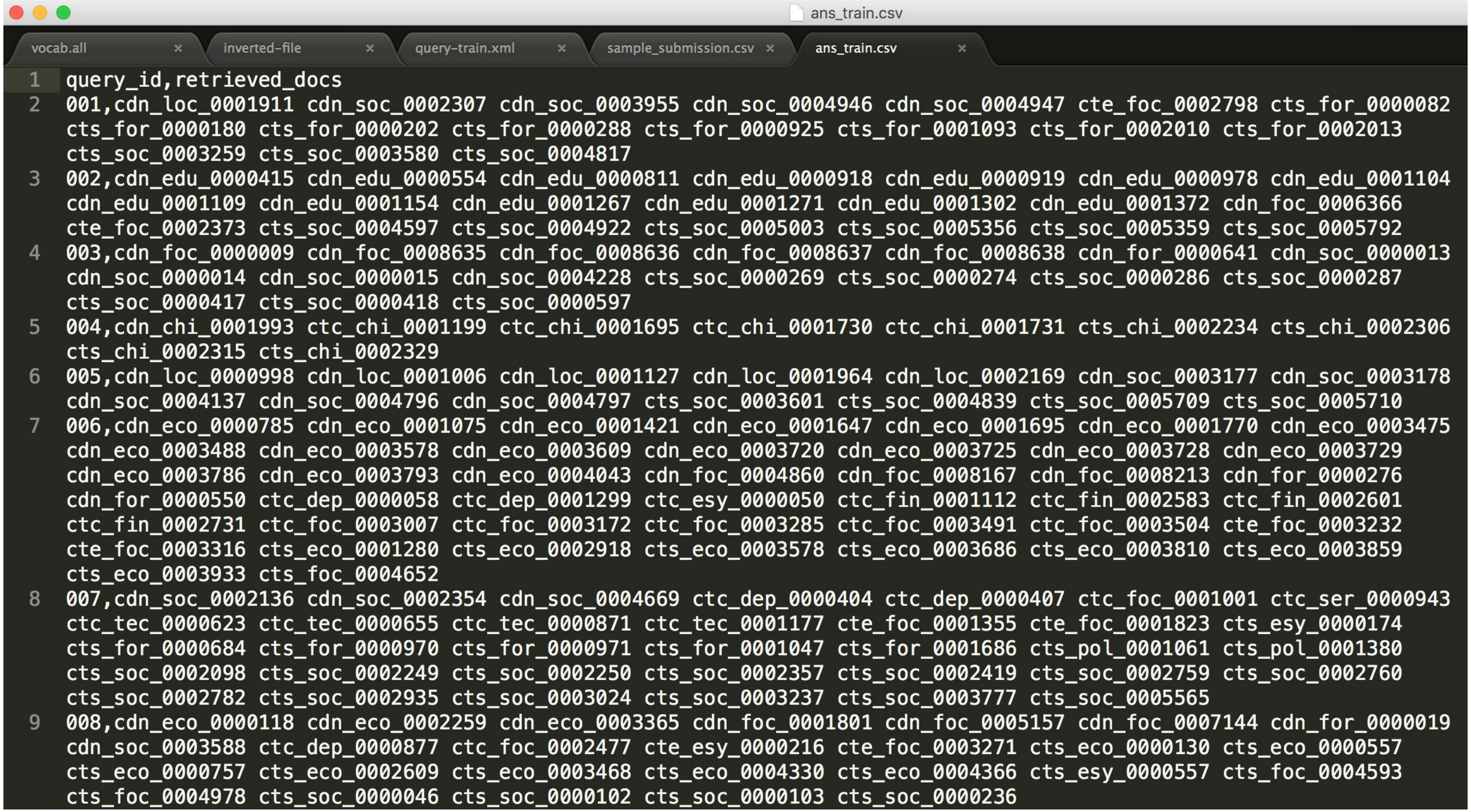
```xml
<?xml version="1.0" encoding="UTF-8"?>
<xml>
<topic>
<number>CIRB010TopicZH001</number>
<title>集會遊行法與言論自由</title>
<question>
查詢集會遊行法中有關主張共產主義或分裂國土規定之修正與討論。
</question>
<narrative>
相關文件內容應敘述集會遊行法原本對主張共產主義或分裂國土之限制，其是否符合憲法中對言論自由等基本人權的保障，大法官對此議題的相關解釋，學者專
討論與看法，以及集會遊行法條文的修改現況。
</narrative>
<concepts>
集會遊行法、集會遊行、集遊法、憲法、言論自由、保障、共產主義、分裂國土、大法官會議、立法、修正條文。
</concepts>
</topic>

<topic>
<number>CIRB010TopicZH002</number>
<title>新三不政策與台獨</title>
<question>
查詢美國所提出之對台新三不政策對民進黨台獨主張的影響。
</question>
<narrative>
１９９８年柯江會談中，美方非正式發表了對台新三不政策：不支持台獨、不支持一中一台的主張，不支持台灣以主權國家身分加入國際組織。相關文件中應論
對民進黨台獨主張的衝擊，包括民進黨內人士對此的反應、看法，該黨人士所發表的相關言論，所引起的黨內爭議及採取的因應策略等。
</narrative>
<concepts>
新三不政策、柯江會談、民進黨、建國黨、許信良、林義雄、施明德、台獨、統獨、獨派、兩岸、中國、美國、中共、大陸、台獨黨綱、主權獨立、轉型。
</concepts>
</topic>
```

Ranking List Format

- Each line contains a topic and a document relevant to this topic.
- First column: **topic number**, which is the last three digits in <number>…</number> tag in the query xml file.
- Second column: **document_id**, which is the string in <id>…</id> tag in the NTCIR document. Please note it should be in lowercase.
- The two columns should be separated by a space.
- There will be several lines for the same topic if there are many.
- Note that the lines of a same topic should be sorted by their ranks, which means that, take topic **001** for example, **cdn_foc_0004185** should be more relevant to **001** than **cdn_foc_0004196**.
- You can return at most 100 documents for each topic.

Ranking List Format



```
query_id,retrieved_docs
001,cdn_loc_0001911 cdn_soc_0002307 cdn_soc_0003955 cdn_soc_0004946 cdn_soc_0004947 cte_foc_0002798 cts_for_0000082
cts_for_0000180 cts_for_0000202 cts_for_0000288 cts_for_0000925 cts_for_0001093 cts_for_0002010 cts_for_0002013
cts_soc_0003259 cts_soc_0003580 cts_soc_0004817
002,cdn_edu_0000415 cdn_edu_0000554 cdn_edu_0000811 cdn_edu_0000918 cdn_edu_0000919 cdn_edu_0000978 cdn_edu_0001104
cdn_edu_0001109 cdn_edu_0001154 cdn_edu_0001267 cdn_edu_0001271 cdn_edu_0001302 cdn_edu_0001372 cdn_foc_0006366
cte_foc_0002373 cts_soc_0004597 cts_soc_0004922 cts_soc_0005003 cts_soc_0005356 cts_soc_0005359 cts_soc_0005792
003,cdn_foc_0000009 cdn_foc_0008635 cdn_foc_0008636 cdn_foc_0008637 cdn_foc_0008638 cdn_for_0000641 cdn_soc_0000013
cdn_soc_0000014 cdn_soc_0000015 cdn_soc_0004228 cts_soc_0000269 cts_soc_0000274 cts_soc_0000286 cts_soc_0000287
cts_soc_0000417 cts_soc_0000418 cts_soc_0000597
004,cdn_chi_0001993 ctc_chi_0001199 ctc_chi_0001695 ctc_chi_0001730 ctc_chi_0001731 cts_chi_0002234 cts_chi_0002306
cts_chi_0002315 cts_chi_0002329
005,cdn_loc_0000998 cdn_loc_0001006 cdn_loc_0001127 cdn_loc_0001964 cdn_loc_0002169 cdn_soc_0003177 cdn_soc_0003178
cdn_soc_0004137 cdn_soc_0004796 cdn_soc_0004797 cts_soc_0003601 cts_soc_0004839 cts_soc_0005709 cts_soc_0005710
006,cdn_eco_0000785 cdn_eco_0001075 cdn_eco_0001421 cdn_eco_0001647 cdn_eco_0001695 cdn_eco_0001770 cdn_eco_0003475
cdn_eco_0003488 cdn_eco_0003578 cdn_eco_0003609 cdn_eco_0003720 cdn_eco_0003725 cdn_eco_0003728 cdn_eco_0003729
cdn_eco_0003786 cdn_eco_0003793 cdn_eco_0004043 cdn_foc_0004860 cdn_foc_0008167 cdn_foc_0008213 cdn_for_0000276
cdn_for_0000550 ctc_dep_0000058 ctc_dep_0001299 ctc_esy_0000050 ctc_fin_0001112 ctc_fin_0002583 ctc_fin_0002601
ctc_fin_0002731 ctc_foc_0003007 ctc_foc_0003172 ctc_foc_0003285 ctc_foc_0003491 ctc_foc_0003504 cte_foc_0003232
cte_foc_0003316 cts_eco_0001280 cts_eco_0002918 cts_eco_0003578 cts_eco_0003686 cts_eco_0003810 cts_eco_0003859
cts_eco_0003933 cts_foc_0004652
007,cdn_soc_0002136 cdn_soc_0002354 cdn_soc_0004669 ctc_dep_0000404 ctc_dep_0000407 ctc_foc_0001001 ctc_ser_0000943
ctc_tec_0000623 ctc_tec_0000655 ctc_tec_0000871 ctc_tec_0001177 cte_foc_0001355 cte_foc_0001823 cts_esy_0000174
cts_for_0000684 cts_for_0000970 cts_for_0000971 cts_for_0001047 cts_for_0001686 cts_pol_0001061 cts_pol_0001380
cts_soc_0002098 cts_soc_0002249 cts_soc_0002250 cts_soc_0002357 cts_soc_0002419 cts_soc_0002759 cts_soc_0002760
cts_soc_0002782 cts_soc_0002935 cts_soc_0003024 cts_soc_0003237 cts_soc_0003777 cts_soc_0005565
008,cdn_eco_0000118 cdn_eco_0002259 cdn_eco_0003365 cdn_foc_0001801 cdn_foc_0005157 cdn_foc_0007144 cdn_for_0000019
cdn_soc_0003588 ctc_dep_0000877 ctc_foc_0002477 cte_esy_0000216 cte_foc_0003271 cts_eco_0000130 cts_eco_0000557
cts_eco_0000757 cts_eco_0002609 cts_eco_0003468 cts_eco_0004330 cts_eco_0004366 cts_esy_0000557 cts_foc_0004593
cts_foc_0004978 cts_soc_0000046 cts_soc_0000102 cts_soc_0000103 cts_soc_0000236
```

Program IO

- Your program is required to support input of a **query file**, and output a **ranking list**.
- We provide 30 query files for you as inputs.(We provide only 10 answers, 10 queries for public scoreboard and 10 queries for private scoreboard)
- There is no restriction to the programming language you use, but make sure your program is **executable on R217 workstation**.
- Using the third party tools directly for VSM or Relevance Feedback is prohibited.

Program Execution Details

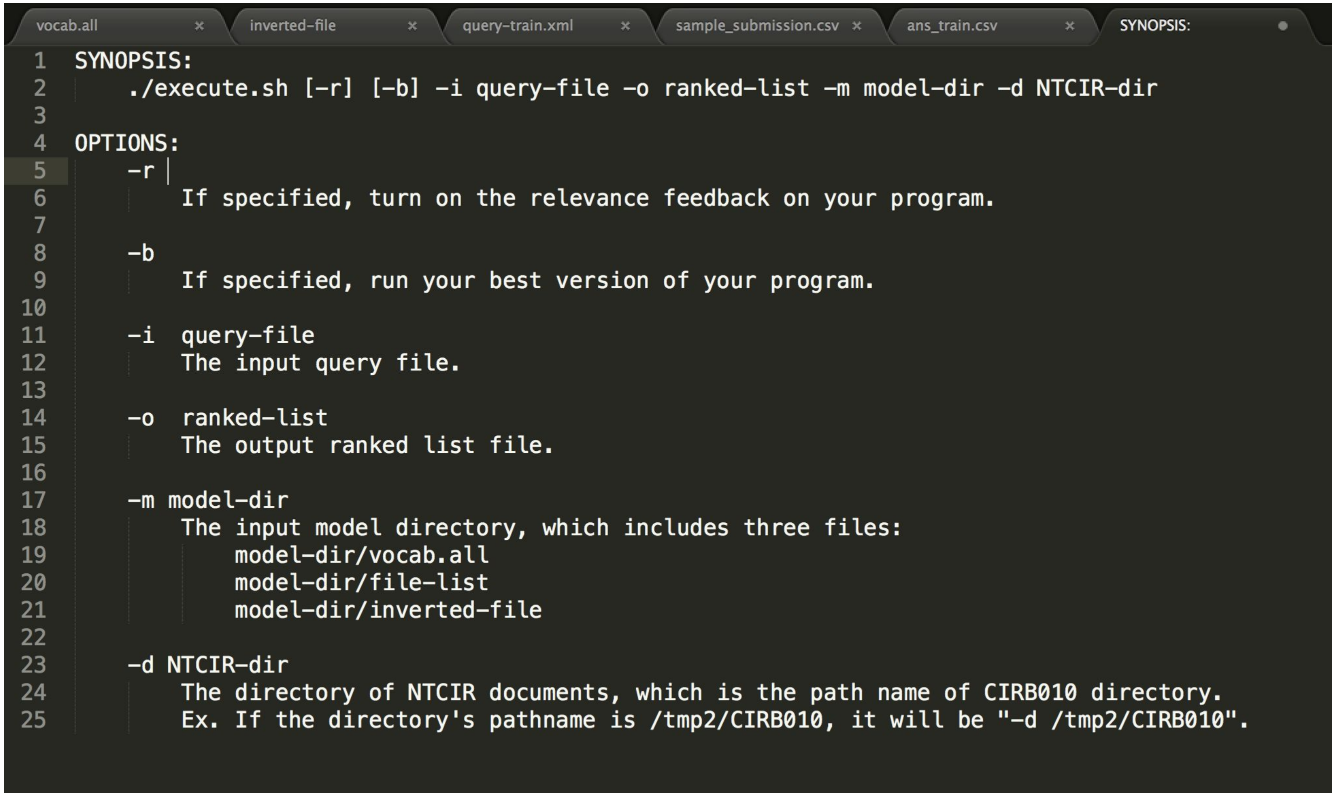
- You are given two shell scripts to compile and run your program.

- You should edit these two scripts according to how you implement this assignment.

- When testing your program, we will execute the following commands **on R217 workstation**, please make sure your program is executable on the workstation.
  - $./compile.sh
  - $./execute.sh -option1 value1 -option2 value2...

Program Execution Details

- Here are the required options that must be supported by your program. (Options without default values are guaranteed to be specified when we test your program.)



```
SYNOPSIS:
    ./execute.sh [-r] [-b] -i query-file -o ranked-list -m model-dir -d NTCIR-dir

OPTIONS:
    -r
        If specified, turn on the relevance feedback on your program.

    -b
        If specified, run your best version of your program.

    -i  query-file
        The input query file.

    -o  ranked-list
        The output ranked list file.

    -m model-dir
        The input model directory, which includes three files:
            model-dir/vocab.all
            model-dir/file-list
            model-dir/inverted-file

    -d NTCIR-dir
        The directory of NTCIR documents, which is the path name of CIRB010 directory.
        Ex. If the directory's pathname is /tmp2/CIRB010, it will be "-d /tmp2/CIRB010".
```

Evaluation

- We will use the **Mean Average Precision (MAP)** value to evaluate your ranking list.

- We provide an answer **ranking list** for **query-train.xml**.
  - There're two columns in the answer list, first is the **topic number**, and second is the **document_id** relevant to this topic.
  - You can use this answer list to check your system's performance.

Report (8%)

- Please write your report in a **Report.pdf** and put it into the submission zip. The report should contain the following content:
    - (2%) Describe your VSM (e.g., parameter….)
    - (2%) Describe your Rocchio Relevance Feedback (e.g., how do you define relevant documents, parameter…)
    - (3%) **Results of Experiments**
        - **MAP value under different parameters of VSM**
        - **Feedback vs. no Feedback**
        - Other experiments you tried
    - (1%) Discussion: what you learn in the homework.

Submission

- Please put report ,scripts and code into the directory named your **student ID**. Package this folder into a zip file and submit it to CEIBA, following is the structure and content of the zip:
- For example: R05922XXX.zip

```
+---R05922XXX(directory)
    +---report.pdf
    +---compile.sh
    +---execute.sh
    +---source/
```

(Note that you don't have to submit the model files and NTCIR documents)

Scoring(15 points)

- 2% for VSM model.
- 2% for Rocchio relevance feedback.
- 8% for your report.
- 3% for performance (2% for simple baseline, 1% for strong baseline on public scoreboard)

- Note that you'll get 0 point if
  1. you don’t have any record on kaggle
  2. you don't sign up the user_agreement_form
  3. you get caught cheating by kaggle
  4. TA can’t reproduce your work

Bonus

- Top-3 ranking at public scoreboard.
  - 1% for 1st-3rd
- Top-10 ranking at private scoreboard.
  - 3% for 1st-3rd
  - 2% for 4th-5th
  - 1% for 6th-10th

Bonus package limitation

1. Any external data is allowed except for ground truth !
   (pre-trained word2vec model)
2. Any package is allowed except for search engine package !
3. You need to describe your method in your report !

Rules

- Kaggle:
  - display name: 學號_ID (Ex: r05922032_嘿嘿)
  - 5 times submissions a day
  - 2 entries for private scoreboard submissions
- Deadline:
  - Kaggle: 2018/04/13 08:00:00
  - Ceiba: 2018/04/14 23:59:59
- Late policy: 10% per day
- Email to TAs: **ntucsiewm@gmail.com**