BM25-CTF Search Engine

By Group E

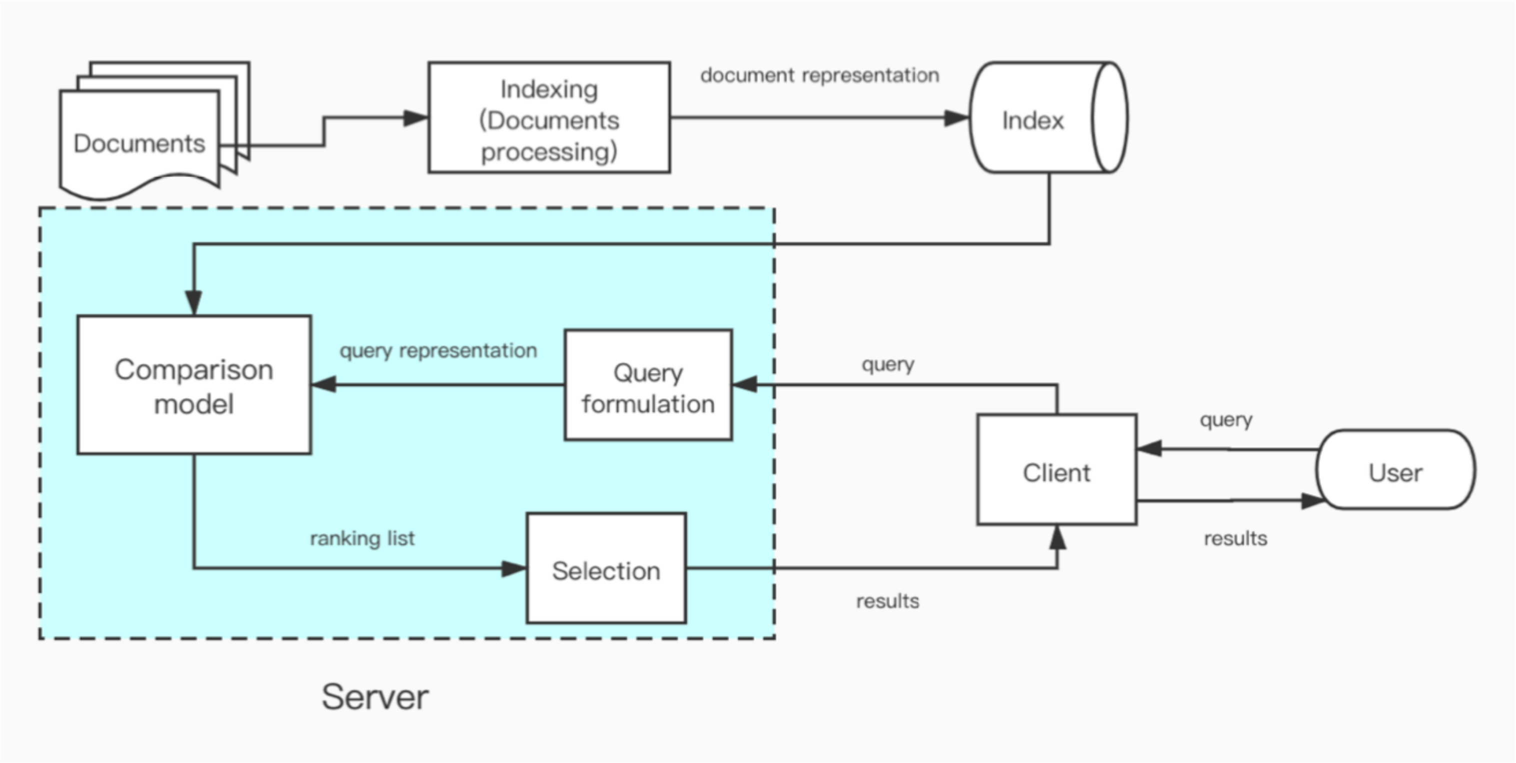
(Jingye Shang, Shangyu Luo, Yinghao Qin, ChungHao Ku)

Content

- Recap: architecture, retrieval function, methods (ChungHao)
- Data Processing (JingYe)
- Demo (Yinghao)
- Results and Analysis (ShangYu)

Recap

Search Engine Architecture



BM25-CTF

- BM25

$$\sum_{w \in Q} idf(w) \times \frac{(k_1+1) \times tf(w,d)}{k_1 \times K(d) + tf(w,d)} \times \frac{(k_3+1) \times tf(w,q)}{k_3 + tf(w,q)}$$

- BM25-CTF[1]

$$\sum_{w \in q} bidf(w) \times \frac{(k_1 + 1) \times btf(w,d)}{k_1 \times K(d) + btf(w,d)} \times \frac{(k_3 + 1) \times btf(w,q)}{k_3 + btf(w,q)}$$

- Why CTF?

[1] Jimenez, Sergio, et al. (2018). BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies. Journal of Intelligent & Fuzzy Systems. 34. 1-13. 10.3233/JIFS-169475.

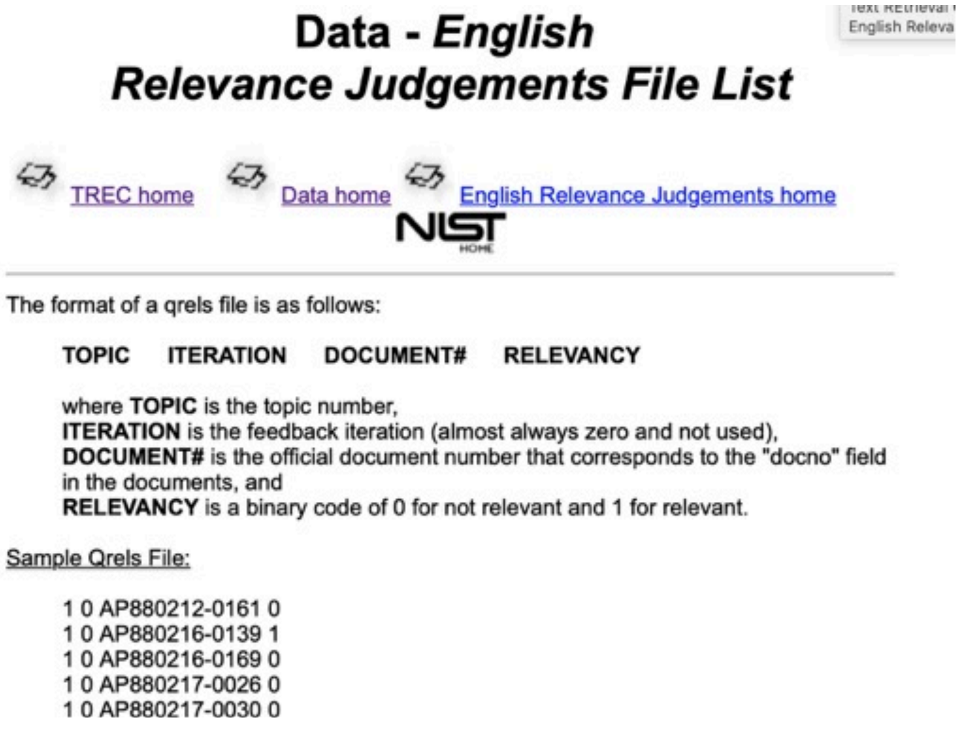
Software Modules

Five important modules to consider in the code development:

1. Query and Document parser
2. Query Processor
3. Ranking Function
4. Data Structures
5. Evaluations

Data processing

- Official data we used (open online resource): including topic ,iteration, document and relevancy.
- Link: [2] https://trec.nist.gov/data/qrels_eng/



Data - *English*
Relevance Judgements File List

The format of a qrels file is as follows:

TOPIC ITERATION DOCUMENT# RELEVANCY

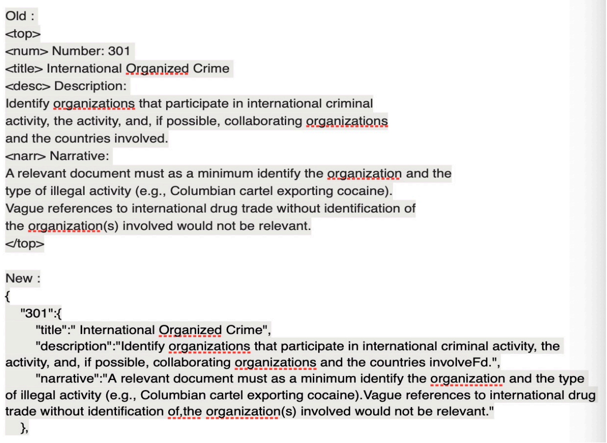
where **TOPIC** is the topic number,
ITERATION is the feedback iteration (almost always zero and not used),
DOCUMENT# is the official document number that corresponds to the "docno" field in the documents, and
RELEVANCY is a binary code of 0 for not relevant and 1 for relevant.

Sample Qrels File:

```
1 0 AP880212-0161 0
1 0 AP880216-0139 1
1 0 AP880216-0169 0
1 0 AP880217-0026 0
1 0 AP880217-0030 0
```

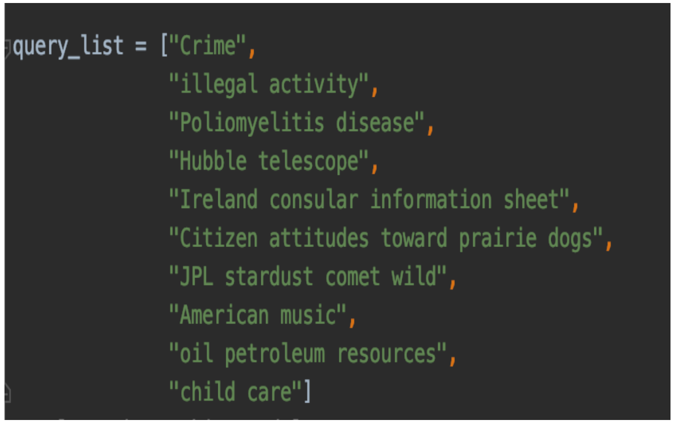
Data processing

- Data examples:

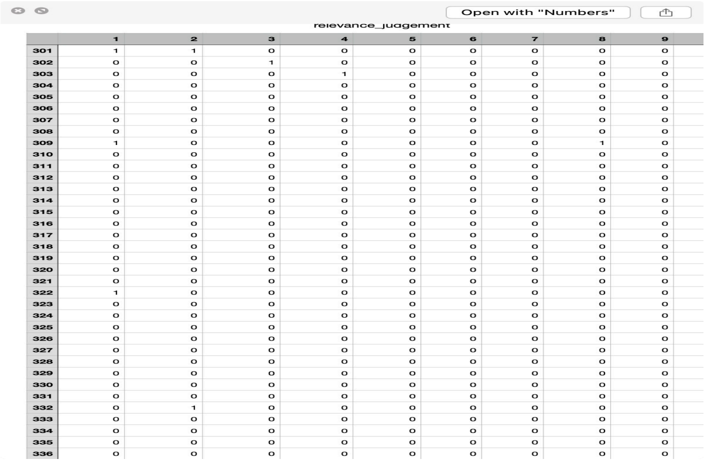


```
Old :
<top>
<num> Number: 301
<title> International Organized Crime
<desc> Description:
Identify organizations that participate in international criminal
activity, the activity, and, if possible, collaborating organizations
and the countries involved.
<narr> Narrative:
A relevant document must as a minimum identify the organization and the
type of illegal activity (e.g., Columbian cartel exporting cocaine).
Vague references to international drug trade without identification of
the organization(s) involved would not be relevant.
</top>

New :
{
   "301":{
      "title":" International Organized Crime",
      "description":"Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involveFd.",
      "narrative":"A relevant document must as a minimum identify the organization and the type of illegal activity (e.g., Columbian cartel exporting cocaine).Vague references to international drug trade without identification of,the organization(s) involved would not be relevant."
   },
```



```
query_list = ["Crime",
              "illegal activity",
              "Poliomyelitis disease",
              "Hubble telescope",
              "Ireland consular information sheet",
              "Citizen attitudes toward prairie dogs",
              "JPL stardust comet wild",
              "American music",
              "oil petroleum resources",
              "child care"]
```

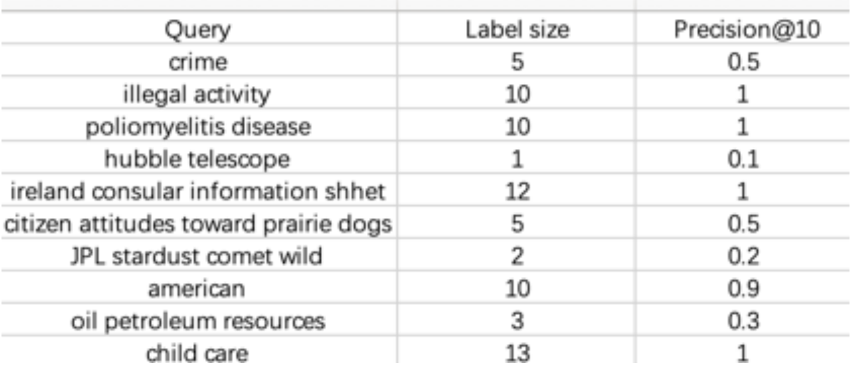


relevance_judgement

	1	2	3	4	5	6	7	8	9
301	1	1	0	0	0	0	0	0	0
302	0	0	1	0	0	0	0	0	0
303	0	0	0	1	0	0	0	0	0
304	0	0	0	0	0	0	0	0	0
305	0	0	0	0	0	0	0	0	0
306	0	0	0	0	0	0	0	0	0
307	0	0	0	0	0	0	0	0	0
308	0	0	0	0	0	0	0	0	0
309	1	0	0	0	0	0	0	1	0
310	0	0	0	0	0	0	0	0	0
311	0	0	0	0	0	0	0	0	0
312	0	0	0	0	0	0	0	0	0
313	0	0	0	0	0	0	0	0	0
314	0	0	0	0	0	0	0	0	0
315	0	0	0	0	0	0	0	0	0
316	0	0	0	0	0	0	0	0	0
317	0	0	0	0	0	0	0	0	0
318	0	0	0	0	0	0	0	0	0
319	0	0	0	0	0	0	0	0	0
320	0	0	0	0	0	0	0	0	0
321	0	0	0	0	0	0	0	0	0
322	1	0	0	0	0	0	0	0	0
323	0	0	0	0	0	0	0	0	0
324	0	0	0	0	0	0	0	0	0
325	0	0	0	0	0	0	0	0	0
326	0	0	0	0	0	0	0	0	0
327	0	0	0	0	0	0	0	0	0
328	0	0	0	0	0	0	0	0	0
329	0	0	0	0	0	0	0	0	0
330	0	0	0	0	0	0	0	0	0
331	0	0	0	0	0	0	0	0	0
332	0	1	0	0	0	0	0	0	0
333	0	0	0	0	0	0	0	0	0
334	0	0	0	0	0	0	0	0	0
335	0	0	0	0	0	0	0	0	0
336	0	0	0	0	0	0	0	0	0

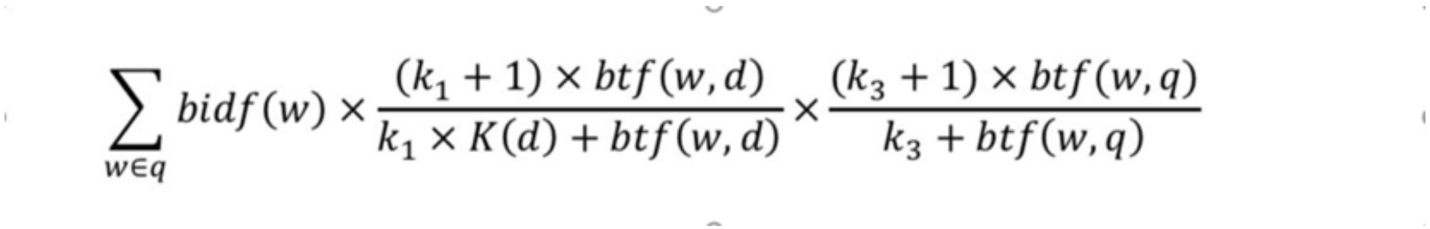
Results and Analysis

Stop-words removal And stemming	BM25	BM25-CTF（bidf	BM25-CTF（btf）	BM25-CTF（bidf-btf）
Yes NO	0.6252956404	0.6277956404	0.6215351167	0.6257017834
No Yes	0.4008121874	0.4049788541	0.4033121874	0.4062288541
Yes Yes	0.4008121874	0.4033121874	0.4008121874	0.4049788541
No NO	0.6252956404	0.6294623071	0.6230368367	0.6284535034



Query	Label size	Precision@10
crime	5	0.5
illegal activity	10	1
poliomyelitis disease	10	1
hubble telescope	1	0.1
ireland consular information shhet	12	1
citizen attitudes toward prairie dogs	5	0.5
JPL stardust comet wild	2	0.2
american	10	0.9
oil petroleum resources	3	0.3
child care	13	1

Optimization



$$\sum_{w \in q} bidf(w) \times \frac{(k_1 + 1) \times btf(w,d)}{k_1 \times K(d) + btf(w,d)} \times \frac{(k_3 + 1) \times btf(w,q)}{k_3 + btf(w,q)}$$

K_1, b, K_3？

Weight of term？

References

- *[1] Jimenez, Sergio, et al. (2018). BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies. Journal of Intelligent & Fuzzy Systems. 34. 1-13. 10.3233/JIFS-169475.*

- *[2] https://trec.nist.gov/data/qrels_eng/*