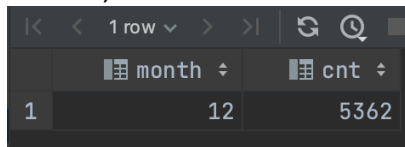


Report

21700303 Junhyung Park
21700352 Mideum Seo
21701062 Changhee Han

I. TASKS

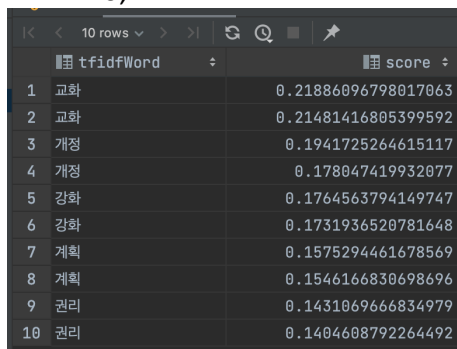
1. SELECT month(post_date) AS month, COUNT(*) AS cnt
FROM post
GROUP BY month
ORDER BY cnt DESC
LIMIT 1;



	month	cnt
1	12	5362

[2023-06-19 00:41:32] 1 row retrieved starting from 1 in 62 ms (execution: 44 ms, fetching: 18 ms)

2. SELECT tfidfWord, Score
FROM
(SELECT hash_key, savedDocHashKey, count(savedDocHashKey) AS cnt
FROM mydocs LEFT JOIN post ON mydocs.savedDocHashKey=post.hash_key
WHERE post_date LIKE '2011%'
GROUP BY savedDocHashKey
ORDER BY cnt DESC
LIMIT 1) AS H, document AS D, frequency AS F
WHERE H.hash_key = D.hash_key AND D.doc_title = F.docTitle
ORDER BY score desc
LIMIT 10;



	tfidfWord	score
1	교화	0.21886096798017063
2	교화	0.21481416805399592
3	개정	0.1941725264615117
4	개정	0.178047419932077
5	강화	0.1764563794149747
6	강화	0.1731936520781648
7	계획	0.1575294461678569
8	계획	0.1546166830698696
9	권리	0.1431069666834979
10	권리	0.1404608792264492

[2023-06-19 00:44:07] 10 rows retrieved starting from 1 in 36 ms (execution: 23 ms, fetching: 13 ms)

3. SELECT docTitle
FROM
(SELECT savedDocHashKey, COUNT(*) AS cnt
FROM mydocs
WHERE savedUser LIKE '%handong.ac.kr'
GROUP BY savedDocHashKey
ORDER BY cnt ASC) AS H, document AS D, frequency AS F, similarity AS S

WHERE H.savedDocHashKey = D.hash_key AND D.doc_title = F.docTitle AND F.docID = S.docID

0 rows
docTitle
[2023-06-19 02:18:35] 0 rows retrieved in 63 ms (execution: 51 ms, fetching: 12 ms)

- SELECT tfidfWord, Score
FROM
(SELECT hash_key, savedDocHashKey, count(savedDocHashKey) AS cnt
FROM mydocs LEFT JOIN post ON mydocs.savedDocHashKey=post.hash_key
WHERE post_writer = '조한범'
GROUP BY savedDocHashKey
ORDER BY cnt DESC
LIMIT 1,1) AS H, document AS D, frequency AS F
WHERE H.hash_key = D.hash_key AND D.doc_title = F.docTitle
ORDER BY score desc
LIMIT 6;

<

>

6 rows

<

>

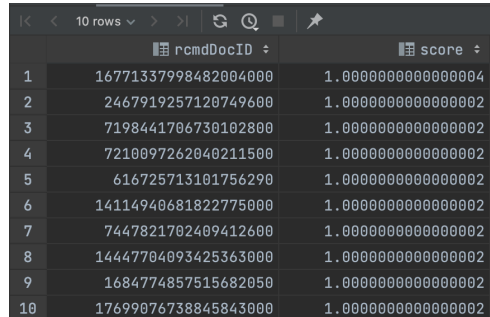
tfidfWord	score
1 경제	0.19900737329302598
2 경제	0.19020710061668764
3 광복절	0.13204479006257422
4 경축사	0.1266646169337994
5 광복절	0.11831780520431984
6 경축사	0.11349693892164234

[2023-06-19 00:49:17] 6 rows retrieved starting from 1 in 129 ms (execution: 117 ms, fetching: 12 ms)

- SELECT tfidfWord, COUNT(*) AS word_count
FROM frequency
GROUP BY tfidfword;

<

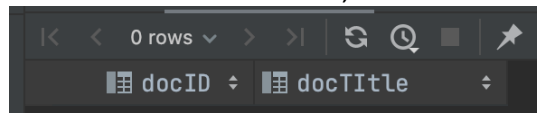
6. SELECT rcmdDocID, Score
FROM
 (SELECT docID
 FROM frequency
 WHERE tfidfWord = '개인'
 ORDER BY Score DESC) AS H, similarity AS S
WHERE H.docID = S.docID
ORDER BY Score desc
LIMIT 10;



	rcmdDocID	score
1	1677133798482004000	1.0000000000000004
2	2467919257120749600	1.0000000000000002
3	7198441706730102800	1.0000000000000002
4	7210097262040211500	1.0000000000000002
5	616725713101756290	1.0000000000000002
6	14114940681822775000	1.0000000000000002
7	7447821702409412600	1.0000000000000002
8	14447704093425363000	1.0000000000000002
9	1684774857515682050	1.0000000000000002
10	17699076738845843000	1.0000000000000002

[2023-06-19 00:51:38] 10 rows retrieved starting from 1 in 35 ms (execution: 16 ms, fetching: 19 ms)

7. SELECT *
FROM
 (SELECT *
 FROM frequency
 WHERE tfidfWord >SOME
 (SELECT tfidfWord
 FROM
 (SELECT tfidfWord, COUNT(*) AS cnt
 FROM frequency
 GROUP BY tfidfWord
 ORDER BY cnt DESC
 LIMIT 10) AS A)
 ORDER BY Score DESC
 LIMIT 199,1) AS B, similarity AS S
WHERE B.docID = S.docID;

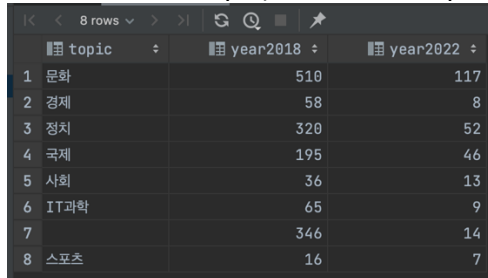


docID	docTitle
-------	----------

[2023-06-19 02:17:32] 0 rows retrieved in 3 s 709 ms (execution: 3 s 697 ms, fetching: 12 ms)

8. SELECT topic2018 AS topic, year2018, year2022
FROM
 (SELECT topic AS topic2018, COUNT(topic) AS year2018
 FROM post
 WHERE post_date LIKE '2018%'
 GROUP BY topic) AS A JOIN

```
(SELECT topic AS topic2022, COUNT(topic) AS year2022
FROM post
WHERE post_date LIKE '2022%'
GROUP BY topic) AS B ON A.topic2018 = B.topic2022;
```



	topic	year2018	year2022
1	문화	510	117
2	경제	58	8
3	정치	320	52
4	국제	195	46
5	사회	36	13
6	IT과학	65	9
7		346	14
8	스포츠	16	7

[2023-06-19 01:29:36] 8 rows retrieved starting from 1 in 39 ms (execution: 27 ms, fetching: 12 ms)

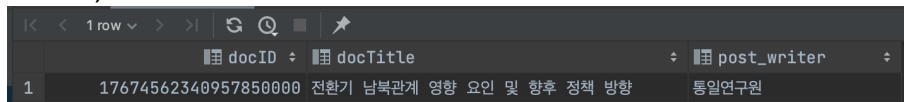
9. SELECT S.docID
FROM
(SELECT hash_key, LENGTH(post_title) AS len
FROM post
WHERE post_date LIKE '2018%'
ORDER BY len DESC
LIMIT 9,1) AS H, document AS D, frequency F, similarity AS S
WHERE H.hash_key=D.hash_key AND D.doc_title=F.docTitle AND F.docID=S.docID



docID

[2023-06-19 02:24:57] 0 rows retrieved in 130 ms (execution: 115 ms, fetching: 15 ms)

10. SELECT docID, docTitle, post_writer
FROM frequency LEFT JOIN post ON frequency.docTitle = post.post_title
WHERE tfidfWrod = '관계' AND post_title_first_char = 'ㅈ'
ORDER BY Score DESC
LIMIT 1;



	docID	docTitle	post_writer
1	17674562340957850000	전환기 남북관계 영향 요인 및 향후 정책 방향	통일연구원

[2023-06-19 01:14:40] 1 row retrieved starting from 1 in 235 ms (execution: 223 ms, fetching: 12 ms)

11. SELECT *
FROM
(SELECT LEFT(post_date, 4) A 강경, COUNT(*) AS 강경 cnt
FROM post
WHERE post_body LIKE '%강경%'
GROUP BY 강경

```

ORDER BY 강경 ASC) AS GG JOIN
(SELECT LEFT(post_date, 4) AS 대화, COUNT(*) AS 대화 cnt
FROM post
WHERE post_body LIKE '%대화%'
GROUP BY 대화
ORDER BY 대화 ASC) AS DH ON GG.강경 = DH.대화;

```

강경	강경cnt	대화	대화cnt
1 2003	2 2003	5	
2 2004	1 2004	22	
3 2005	2 2005	12	
4 2006	32 2006	212	
5 2007	22 2007	172	
6 2008	30 2008	145	
7 2009	87 2009	315	
8 2010	30 2010	71	
9 2011	4 2011	54	
10 2012	23 2012	75	
11 2013	13 2013	50	
12 2014	4 2014	31	
13 2015	7 2015	76	
14 2016	4 2016	28	
15 2017	17 2017	56	
16 2018	47 2018	94	
17 2019	8 2019	45	
18 2020	6 2020	42	
19 2021	5 2021	65	
20 2022	5 2022	58	
21 2023	4 2023	11	

[2023-06-19 01:39:40] 21 rows retrieved starting from 1 in 405 ms (execution: 391 ms, fetching: 14 ms)

II. Description

- First, the normalized tables in phase 1 were denormalized using the parts connected to each primary key. A table with the same primary key was joined with it, and a table without the same column was denormalized through the union command. As a result, the volume of the data became very large, but it was expected that the columns would be included in the same table to speed up the execution time. Also, we tried to reduce the execution time through indexing by table. We have indexed vast amounts of newly acquired data such as Doctitle, score, etc. so that it can be found faster.

III. Summary

- Database
 - Size: 56,979,104.0 KB
- Tables
 - Board Size: 16.0 KB
 - Denormal_freqsim Size: 56,593.408.0 KB
 - Document Size: 14912.0 KB
 - File Size: 178,832.0 KB
 - Frequency Size: 97,920.0 KB
 - Mydocs Size: 1,552.0 KB
 - Post Size: 39,536.0KB
 - Reference Size: 6,672.0 KB

- ix. Similarity Size: 2,575.0 KB
- x. Users Size: 16.0 KB