

Πανεπιστήμιο Κρήτης
Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 Συστήματα Ανάκτησης Πληροφοριών
Εξάμηνο: Άνοιξη 2023

Γραπτή Αναφορά Έργου

Στοιχεία Φοιτητών

Μέλος	1 ^ο
Ονοματεπώνυμο	Ισίδωρος Χατζηχρυσός
ΑΜ	4338
Email	csd4338@csd.uoc.gr

Μέλος	2 ^ο
Ονοματεπώνυμο	Καραγιαννάκη Αντιγόνη
ΑΜ	4642
Email	csd4642@csd.uoc.gr

Πίνακας Περιεχομένων

1 ΕΙΣΑΓΩΓΗ	3
2 RUNS - ΧΡΟΝΟΙ	3
2.1 ΧΡΟΝΟΣ ΕΥΡΕΤΗΡΙΑΣΗΣ MEDICAL COLLECTION	3
2.2 Χρόνος για δημιουργία Result.txt	4
3 ΜΕΤΡΗΣΕΙΣ	5
3.1 bpref	5
3.2 AveP	6
3.3 NDCG	7
4 ΕΠΙΛΟΓΟΣ	7
5 ΑΝΑΦΟΡΕΣ	7

1 Εισαγωγή

Μετά την ολοκλήρωση της δεύτερης φάσης, όπου είχαν πραγματοποιηθεί αλλαγές και διορθώσεις στην πρώτη φάση, καταφέραμε και τρέξαμε το Medical Collection για να αξιολογήσουμε το σύστημα μας στην 3η φάση.

Αρχικά επειδή στην φάση δύο τρέχαμε χειροκίνητα τα queries του topic.xml το αλλάξαμε ώστε να τα τρέχει με αυτόματο τρόπο και να παράγει το result.txt αρχείο. Επιλέξαμε εξαρχείς τα summary για να υπάρχει μεγαλύτερη ακρίβεια επειδή τα description περιέχουν εκτενείς περιγραφές ίσως προκύψουν περισσότερα συναφή αρχεία.

Να σημειωθεί ότι εφόσον στο medical collection δεν υπάρχει αντιστοιχία και οργάνωση με type (diagnosis / test / treatment) αρχείου στις 3 κατηγορίες, τα αποτελέσματα που λαμβάνουμε είναι τα Top 1000 ranked με βάση του cos_similarity.

2 Runs - Χρόνοι

2.1 Χρόνος ευρετηρίασης Medical Collection

Ο χρόνος ευρετηρίασης της μεγάλης συλλογής (Medical Collection) :

Total time = 36406580 ms = 10 Hours 6 Minutes 46 Seconds

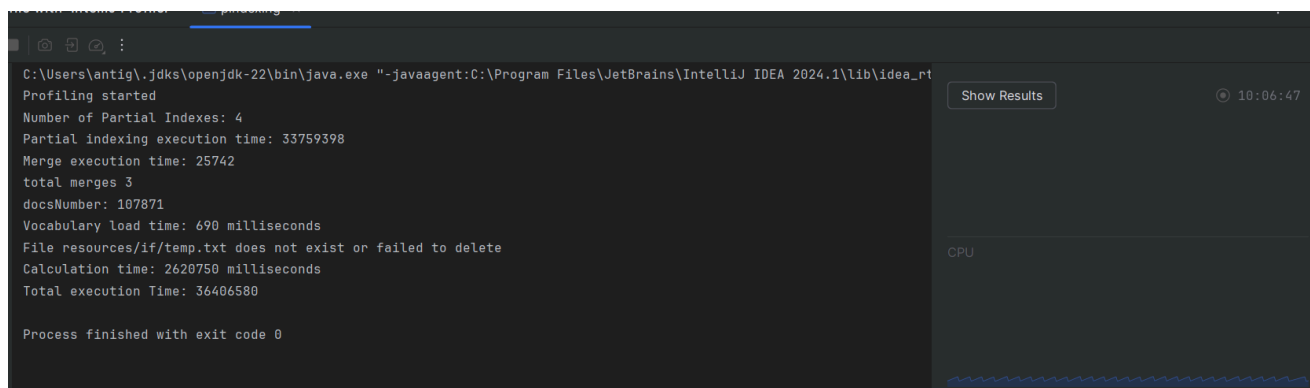
Total time = Partial indexing + Distance Calculation

Partial indexing time = 33759398 ms = 9 Hours 22 Minutes 39 Seconds

Distance Calculation = 2620750 ms = 43 Minutes 40 Seconds

Note: To run πραγματοποιήθηκε με profiler του intellij ο οποίος κάνει παραπάνω χρόνο διότι υπολογίζει του χρόνου σε κάθε σημείο όπου το πρόγραμμα εκτελείτε.

Επίσης για να υπάρχει μεγαλύτερη ακρίβεια στα calculation των norms του κάθε doc τα υπολογίζουμε αφού έχει ολοκληρωθεί το partial indexing & merging για να έχουμε τα σωστά df = document frequencys.



```
C:\Users\antig\jdk8\openjdk-22\bin\java.exe "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA 2024.1\lib\idea_rt...
Profiling started
Number of Partial Indexes: 4
Partial indexing execution time: 33759398
Merge execution time: 25742
total merges 3
docsNumber: 107871
Vocabulary load time: 690 milliseconds
File resources/lf/temp.txt does not exist or failed to delete
Calculation time: 2620750 milliseconds
Total execution Time: 36406580
Process finished with exit code 0
```

To minicollection ευρετηριάζεται περίπου σε 10 second

Pc specs:

ASUS VivoBook 15 2020, intel core i7 10th generation, nvidia geforce gtx Επεξεργαστής Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz 2.59 GHz Εγκατεστημένη RAM 16,0 GB (15,8 GB χρησιμοποιήσιμη) Λειτουργικό σύστημα 64 bit, επεξεργαστής τεχνολογίας x64

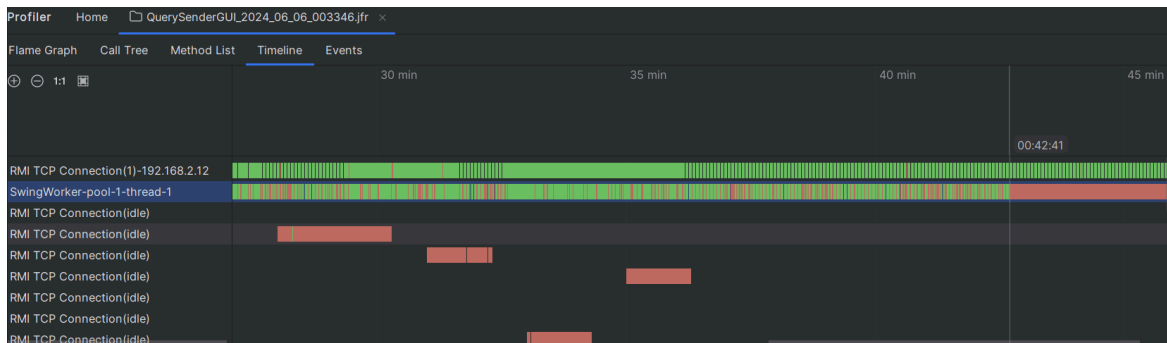
Heap size 4gb intellij και Threshold = 100.000

2.2 Χρόνος για δημιουργία Results.txt

Το result.txt περιέχει τα top 1000 docs retrieved για το κάθε ένα απο τα 30 topics. Σύνολο 30.000 εγγραφές.

- Ο χρόνος εκτέλεσης με την επιλογή των summaries από τα topics ως επερωτήσεις ήταν περίπου 42 λεπτά το οποίο φαίνεται στην φωτογραφία παρακάτω.

Το πρόγραμμα εκτελείται μέσω του gui interface και έγινε profiling μέσω του intellij.



Εικόνα Α , Timeline από το Run του QuerySelector , το οποίο παράγει το results.txt

Βλέπουμε ότι ο stringworker σταμάτησε να μην δίνει responses στα 42 λεπτά.

```
public void sendQueries() { 1 usage  ⚡ antigoniiliii
    SwingWorker<Void, Void> worker = new SwingWorker<Void, Void>() {  ⚡ antigoniiliii
        @Override  ⚡ antigoniiliii
        protected Void doInBackground() throws Exception {
            try {
                //key 1-diagnosis , value - keimeno(query)
                HashMap<String, String> queries = IrQualityEvaluator.readTopicsXML( summaryOrDescription
                int totalQueries = queries.size();
                int currentQuery = 0;

                for (String query : queries.values()) {
                    String key = queries.keySet().stream().filter(k -> queries.get(k).equals(query))
                    String[] parts = key.split( regex: "-" );
                    parts[1] = parts[1].toLowerCase();
                    System.out.println(parts[1]);
                    // store the key of the query
                }
            }
        }
    }
}
```

- Από την άλλη , ο χρόνος εκτέλεσης με την επιλογή των descriptions από τα topics ως επερωτήσεις ήταν περίπου 41 λεπτά (2.501.890 ~ 41,69 minutes).

3 Μετρήσεις

TOPIC_NO	BPREF_VALUE	AVEP_VALUE	NDCG_VALUE	Sum
1	0,5607985480944	0,6304195412842	0,8065952729780	1,9978133623566
2	0,6089743589744	0,8722637598751	0,6518613869065	2,1330995057560
3	0,3440453686200	0,4316875934471	0,6891184109491	1,4648513730162
4	0,4775910364146	0,7517360215377	0,8459985092943	2,0753255672466
5	0,00	0,2883301048088	0,3761312483040	0,66
6	0,5817704426107	0,7565707835958	0,6956023671033	2,0339435933098
7	0,6645424836601	0,7121066379319	0,8873508955874	2,2640000171794
8	0,5093099671413	0,7813250108475	0,6515793448232	1,9422143228120
9	0,0000000000000	0,1764799450270	0,3030439334566	0,4795238784836
10	0,1877777777778	0,3902469277588	0,6607358812862	2,1330995057560
11	0,00	0,2219840369240	0,3891670994780	0,61
12	0,00	0,2602988553617	0,3510144887140	0,61
13	0,00	0,2831525078118	0,4374380378301	0,72
14	0,4232954545455	0,4593163811543	0,5419766766137	1,4245885123135
15	0,6948099415205	0,7755821333146	0,8372486228377	2,3076406976728
16	0,00	0,0636260716896	0,3057908088541	0,37
17	0,7212843429328	0,6525169219493	0,8217792147109	2,1955804795930
18	0,6113769271664	0,6670777224455	0,7703092857089	2,0487639353208
19	0,6203389830508	0,7722868845606	0,6857572578797	2,0783831254911
20	0,4860074626866	0,6424757652147	0,6434897807070	1,7719730086083
21	0,5232447817837	0,6929217632905	0,7474115255523	1,9635780706265
22	0,00	0,2827543896612	0,3931787386989	0,68
23	0,00	0,3340699909472	0,4596424323260	0,79
24	0,5712526931363	0,6000714610937	0,7162003533073	1,8875245075373
25	0,00	0,0944238156438	0,3495149120894	0,44
26	0,5466570466570	0,6358951034823	0,9087140053385	2,0912661554778
27	0,6988338192420	0,9700891920499	0,8606745344582	2,5295975457501
28	0,2729766803841	0,4147244080747	0,5549294001747	1,2426304886335
29	0,4372123602893	0,4965933586853	0,7915572097460	1,7253629287206
30	0,5975765306122	0,6899864692268	0,8525145164032	2,1400775162422
Median	0,4817992495506	0,6152455011890	0,6732465695830	1,7702913203225
Average	0,3713225669100	0,5267004519565	0,6328775384039	1,5309005572704
Min	0,00	0,0636260716896	0,3030439334566	0,3694168805437
Max	0,7212843429328	0,9700891920499	0,9087140053385	2,5295975457501

Εικόνα Β , Πίνακας με τα *evaluation results*

TOPIC_NO	- αριθμός topic (1 - 30 , από topic.xml)
BPREF_VALUE	- Binary Preference Μετρική [αναφ. 3.1]
AVEP_VALUE	- Average Precision Μετρική [αναφ. 3.2]
NDCG_VALUE	- NDCG Μετρική [αναφ. 3.3]

Παρατηρούμε αρνητικές τιμές στην περίπτωση που σημαίνει ότι το σύστημα λαμβάνει Bpref score = 0.

3.1 bpref

Η μετρική bpref (binary preference) είναι σχεδιασμένη για την αξιολόγηση συστημάτων αναζήτησης, ειδικά όταν υπάρχουν ατελή δεδομένα αξιολόγησης. Υπολογίζει την

απόδοση του συστήματος βασισμένη σε σχετικές και μη σχετικές εγγραφές.

Υπολογισμός:

- Bpref λαμβάνει υπόψη μόνο τις σχετικές εγγραφές και μετρά το πόσες φορές μια σχετική εγγραφή κατατάσσεται υψηλότερα από μια μη σχετική εγγραφή.
- Η φόρμουλα για το bpref είναι:

$$bpref = \frac{1}{R} * \sum_r \left(1 - \frac{|number\ of\ non-relevant\ documents\ ranked\ higher\ than\ r|}{min(R, N)} \right)$$

όπου ,

R είναι ο αριθμός των σχετικών ευρεθέντων εγγραφών (Relevant Docs)

N είναι ο αριθμός των μη σχετικών εγγραφών (Non-Relevant Docs)

r είναι ένα σχετικό ευρεθέν αρχείο (Relevant Retrieved doc)

$\sum R$ - άθροισμα που συμβαίνει r φορές (r - number of relevant documents)

$min(R, N)$ - βρίσκει το minimum Set , όποιο είναι μικρότερο (R or N).

Η χρήση του εξωτερικού συντελεστή $\frac{1}{R}$ συμβάλλει στην “φράξη” της μέτρησης μεταξύ ορίων , έτσι ώστε να είναι εύκολες οι συγκρίσεις.

3.2 AveP

Η Μέση Ακρίβεια (AP ή AveP) είναι μια ευρέως χρησιμοποιούμενη μέτρηση στην ανάκτηση πληροφοριών που ποσοτικοποιεί την ποιότητα των ταξινομημένων αποτελεσμάτων ενός συστήματος ανάκτησης . Αυτό συμβαίνει μέσω του τρόπου που υπολογίζεται η μετρική.

Υπολογισμός:

- Αρχικά, εξετάζεται ένα ερωτήματα που έχει ως αποτέλεσμα την ανάκτηση ενός συνόλου στοιχείων (results),
- Υπολογίζεται η ακρίβεια για κάθε θέση όπου ανακτάται ένα σχετικό στοιχείο στη λίστα με τα αποτελέσματα, και τέλος
- Υπολογίζεται ο Μέσος όρος αυτών των τιμών ακριβείας, αλλά μόνο για τις θέσεις όπου ανακτήθηκαν σχετικά στοιχεία.

Η Μέση Ακρίβεια εστιάζει στο πόσο καλά κατατάσσονται τα σχετικά στοιχεία σε αυτήν τη λίστα των αποτελεσμάτων. Ουσιαστικά μετρά την περιοχή κάτω από την καμπύλη ανάκλησης ακριβείας για κάθε ερώτημα (topic) .

$$AveP = \frac{1}{R} * \sum_{k=1}^R (precision\ at\ position\ k * relevance\ at\ position\ k)$$

όπου ,

R είναι ο αριθμός των σχετικών ευρεθέντων εγγραφών (Relevant Docs)
 $relevance$ at position k είναι το κατά πόσον είναι σχετικό ένα έγγραφο k (
 επιλογές: Σχετικό , Μη-Σχετικό - binary)
 $precision$ at position k είναι η ακρίβεια δοθέντως μόνο των τοπ k Ευρεθέντων
 Αρχείων

3.3 NDCG

Το Normalized Discounted Cumulative Gain (**NDCG**) είναι μια μέτρηση που αξιολογεί την ποιότητα των συστημάτων ανάκτησης πληροφοριών και όχι μόνο. Το NDCG βοηθά στη μέτρηση της ικανότητας ενός αλγόριθμου να ταξινομεί στοιχεία με βάση τη συνάφεια.

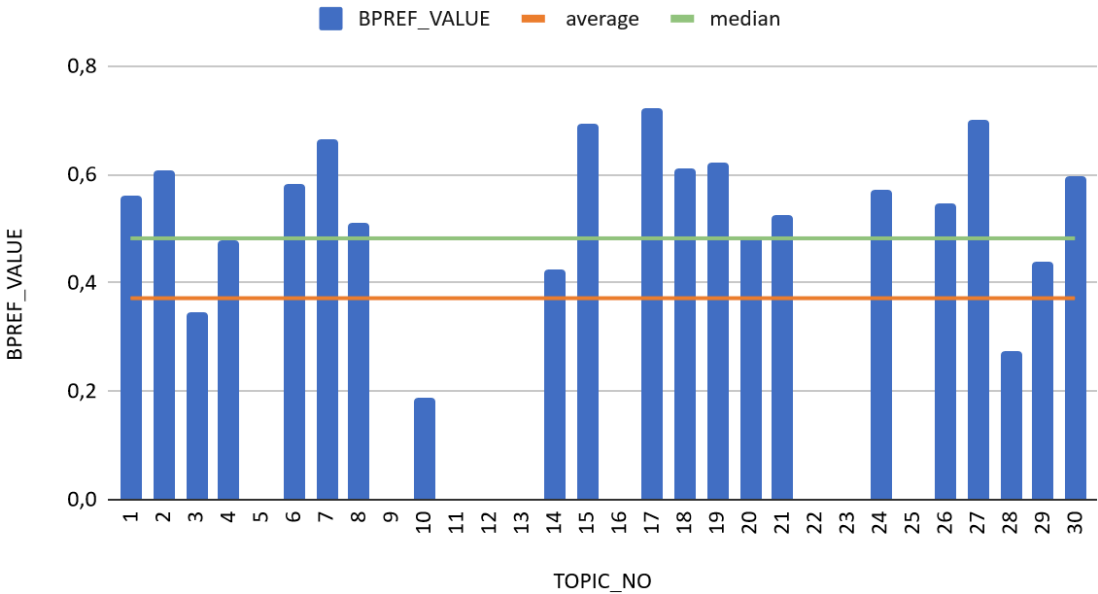
Τα πολύ σχετικά έγγραφα είναι πιο χρήσιμα όταν εμφανίζονται νωρίτερα σε μια λίστα αποτελεσμάτων μηχανής αναζήτησης (έχουν υψηλότερες βαθμολογίες) Τα εξαιρετικά σχετικά έγγραφα είναι πιο χρήσιμα από τα οριακά σχετικά έγγραφα, τα οποία με τη σειρά τους είναι πιο χρήσιμα από τα μη σχετικά έγγραφα.

$$nDCG_p = \frac{DCG_p}{IDCG_p}, \quad IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

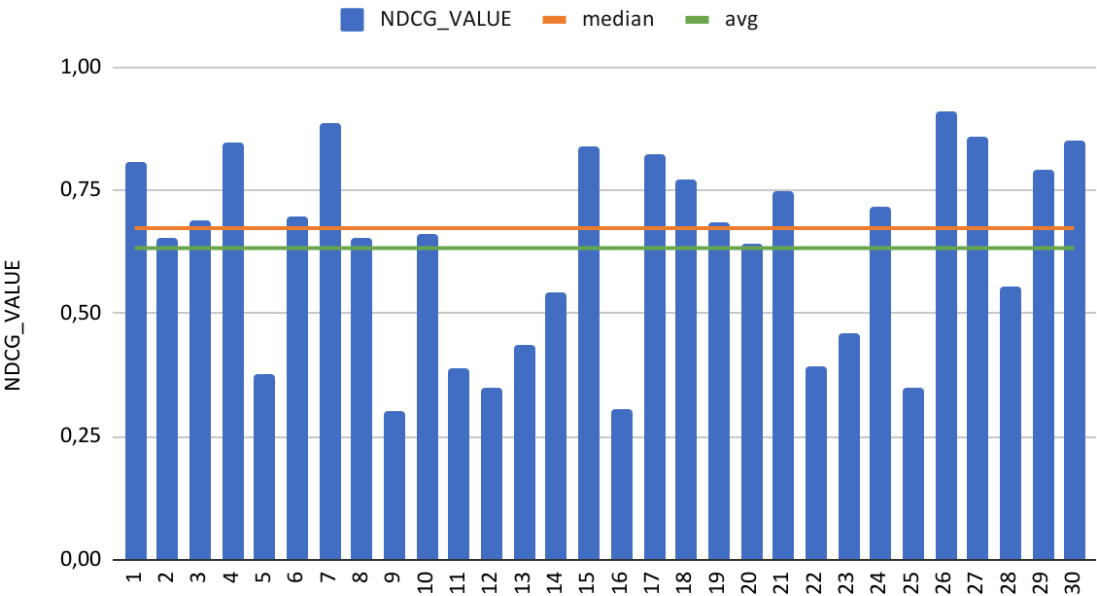
$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i + 1)} = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(i + 1)}$$

3.4 Graphs

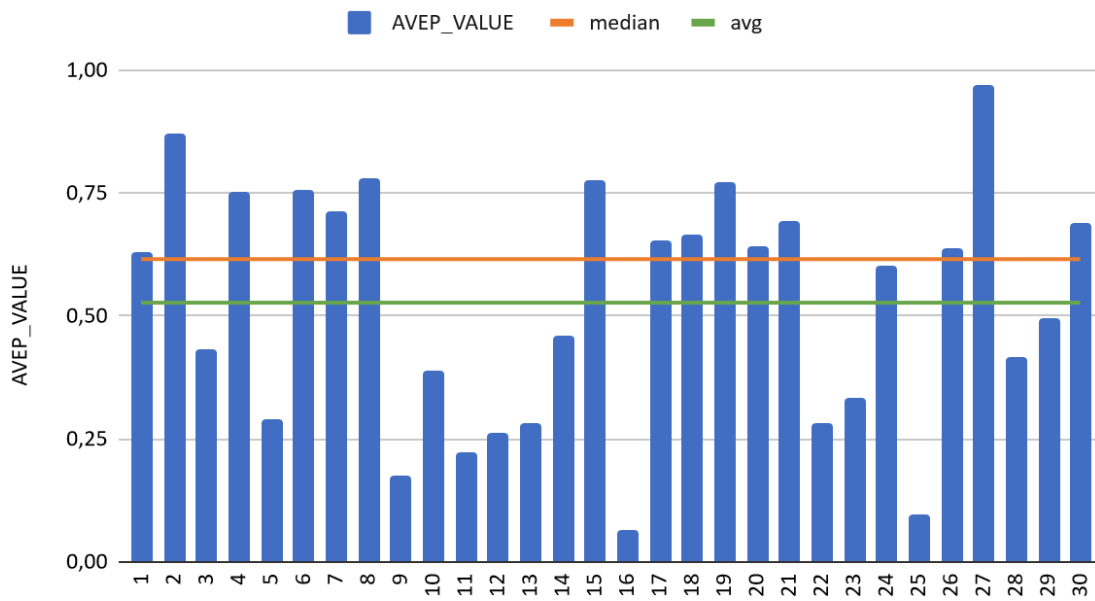
BPREF_VALUE



NDCG_VALUE

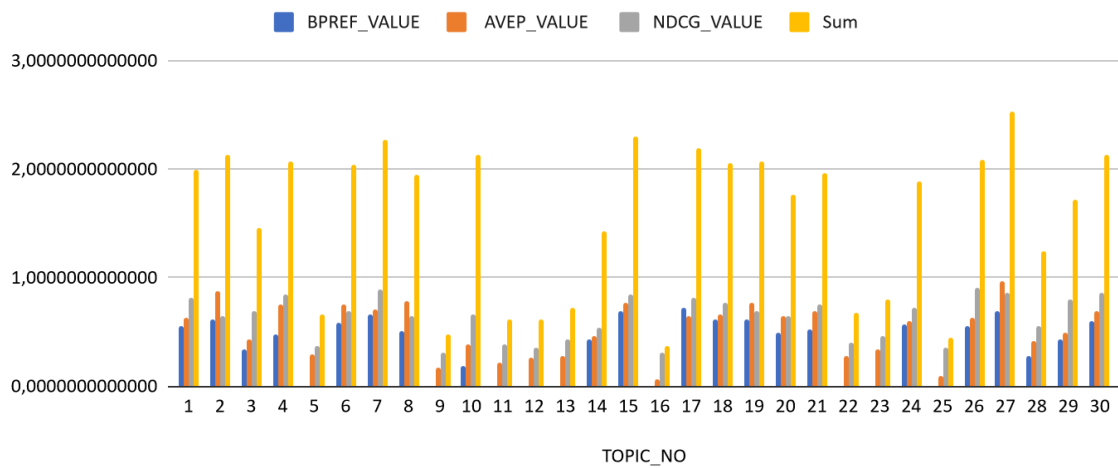


AVEP_VALUE



Sum των metric για υπολογισμός score διαγωνισμού

Sum



✚ evaluation_results

4 Επίλογος

Συνοψίζοντας , αφότου αναλογιστήκαμε τα αποτελέσματα της αξιολόγησής μας , όπως και τους επιμέρους χρόνους εκτέλεσης , είμαστε ικανοποιημένοι σε γενικές γραμμές και από του αλλά και από τα αποτελέσματα.

Θα θέλαμε να επεκτείνουμε και να ολοκληρώσουμε το gui του συστήματος με καλύτερο και πιο δομημένο τρόπο και να βρούμε σημεία που μπορούν να γίνουν βελτιστοποιήσεις ως προς τις επιδόσεις.

5 Αναφορές

- [1] Chris Buckley. trec eval IR evaluation package.
- [2] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 33–40, 2000.
- [3] https://www.cs.helsinki.fi/u/hahonen/irm07/lectures/irm07_2.pdf
- [4] <https://people.cs.georgetown.edu/~nazli/classes/ir-Slides/Evaluation-12.pdf>
- [5] *Ανάκτηση Πληροφορίας*, 2η Έκδοση, Κωδικός Βιβλίου στον Εύδοξο: 41954965, Έκδοση: 2η/2014, Συγγραφείς: *Baeza-Yates Ricardo, Ribeiro-Neto Berthier*
Chatgpt: Χρήση για γρήγορες λύσεις, συμβουλές και προτάσεις