Πανεπιστήμιο Κρήτης

Τμήμα Επιστήμης Υπολογιστών

ΗΥ463 Συστήματα Ανάκτησης Πληροφοριών

Εξάμηνο: Άνοιξη 2021

Γραπτή Αναφορά Έργου

Στοιχεία Φοιτητών

Μέλος	1ο
Ονοματωπώνυμο	Εμμανουήλ Σμυρνάκης
ΑΜ	Csdp1215
Email	msmyrnakis@csd.uoc.gr

Μέλος	2ο
Ονοματωπώνυμο	Ευάγγελος Σουγιάς
ΑΜ	Csd3479
Email	csd3479@csd.uoc.gr

Πίνακας Περιεχομένων

1 Εισαγωγή

Στην πρώτη φάση του Project υλοποιήσαμε τον μηχανισμό ευρετηρίασης και τον μηχανισμό απάντησης ερωτήσεων ακολουθώντας όλα τα βήματα της εκφώνησης πιστά.

Στην δεύτερη φάση του Project υλοποιήσαμε τον μηχανισμό αξιολόγησης όπου και φτιάξαμε τα απαραίτητα αρχεία που ζητούνται και καταγράψαμε τις απαραίτητες μετρικές.

Λόγω περιορισμού μνήμης δεν καταφέραμε να ευρετηριάσουμε και να τεστάρουμε την MedicalCollection ενώ για να δοκιμάσουμε τις μετρικές μας πάνω στην MiniCollection υλοποιήσαμε το qrels1.txt .

Αποφασίσαμε να εργαστούμε από κοινού καθόλη την διάρκεια της εργασίας και δεν την χωρίσαμε σε κομμάτια

2 Υλοποίηση

2.1 Α΄Φάση

Υλοποιήσαμε γραφική διεπαφή τόσο για την διαδικασία ευρετηρίασης και την αποτίμηση επερωτήσεων.

2.1.1 Ευρετηρίαση

Στην διαδικασία της Ευρετηρίασης υλοποιήσαμε με την σειρά:

1. Ανάγνωση και αποθήκευση σε μια δομή δεδομένων των ελληνικών και λατινικών stopwords.
2. Ανάγνωση nxml αρχείων από όλους τους φαέλους αναδρομικά και διαχωρισμός των αρχείων σε λέξεις.
3. Υποστηρίξαμε την διαδικασία του Stemming και την αποθήκευση των αποτελεσμάτων της διαδικασίας.
4. Δημιουργία VocabularyFile αποτελούμενο από όλες τις μοναδικές λέξεις που περιέχονται στα αρχεία.
5. Δημιουργία DocumentFile το οποίο περιέχει όλα τα αρχεία του ευρετηρίου και κάποια στοιχεία για αυτά.
6. Δημιουργία ενος PostingFile με τα στοιχεία που μας ζητούνται και ένωση των VocabulayFile με τα PostingFile και το PostingFile με το DocumentFile.

2.1.2 Αποτίμηση Επερωτήσεων

Στην διαδικασία της Αποτίμησης Επερωτήσεων υλοποιήσαμε με την σειρά:

1. Υλοποίηση διαδικασίας εύρεσης λέξεων αφού φορτώσαμε στην μνήμη το VocabularyFile.
2. Επιστροφή απαντήσεων ως προς το διανυσματικό μοντέλο.
3. Παραγωγή επερωτήσεων σε φυσική γλώσσα και επιστροφή απάντησης με το μονοπάτι αρχείου και βαθμό ομοιότητας.

2.2 Β΄Φάση

Στην δεύτερη φάση του Project υλοποιήσαμε τον μηχανισμό αξιολόγησης όπου και φτιάξαμε τα απαραίτητα αρχεία που ζητούνται και καταγράψαμε τις απαραίτητες μετρικές.

Χαράκτηριστικά :

- Δημιουργήσαμε ένα qrels1.txt για να τρέξουμε μετρήσεις στο MiniCollection καθορίζοντας το relevance του κάθε αρχείου με το topic χρησιμοποιώντας random από το 0 εως το 2.
- Αποθηκεύσαμε τα κορυφαία αποτελέσματα στο αρχείο results.txt
- Αποθηκεύσαμε τις μετρήσεις μας σε TSV αρχείο όπως ζητηθηκε στην εκφώνηση eval_results.txt
- Δημιουργήσαμε γραφική διεπαφή για την έναρξη της αξιολόγησης αφού έχει ολοκληρωθεί η ευρετηρίαση.

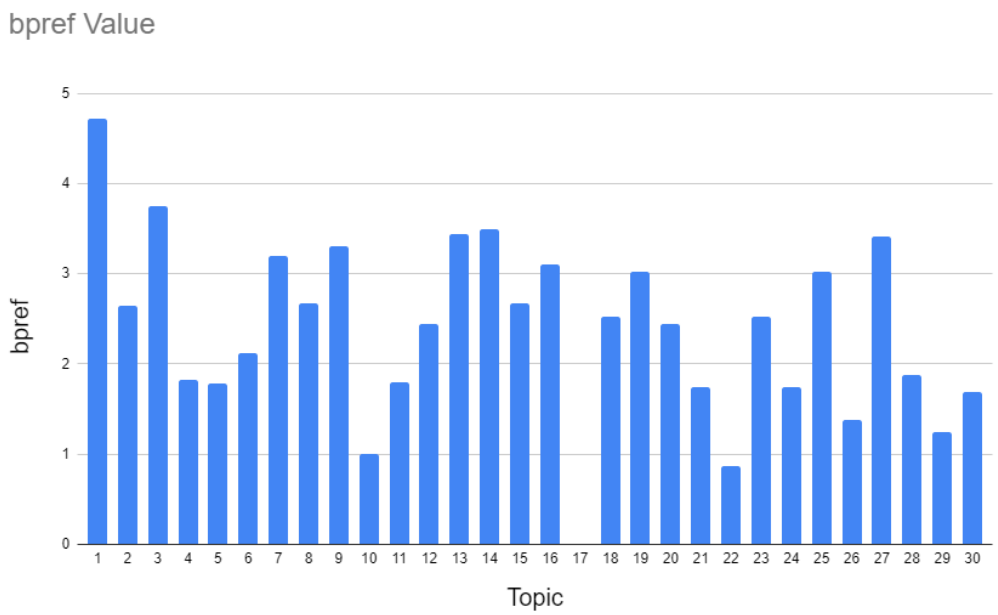
3 Μετρήσεις

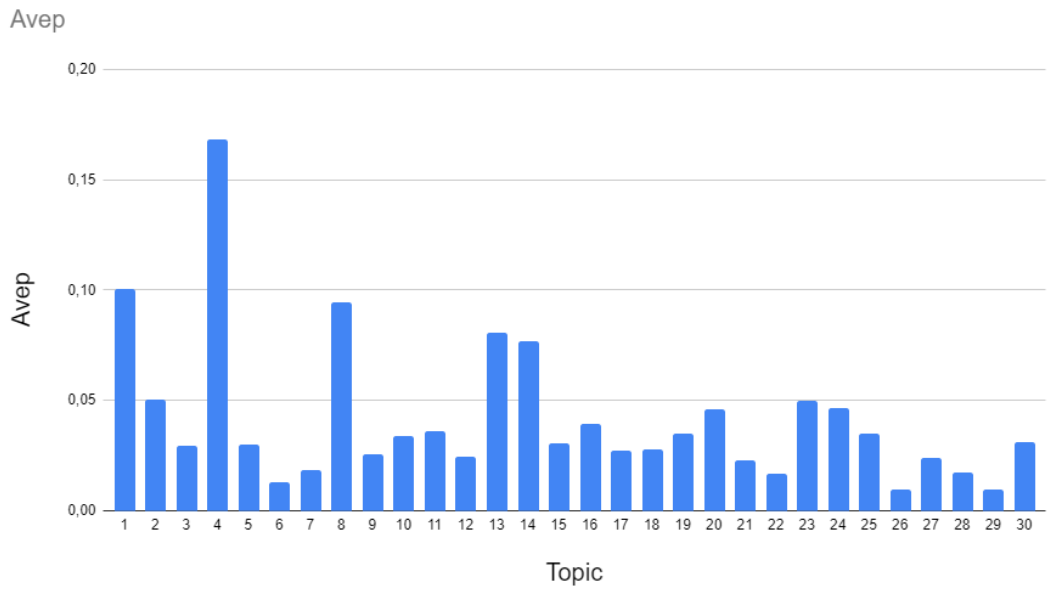
Οι χρόνοι ευρετηρίασης του miniCollection είναι ≃ 2235 που είναι περίπου 37.25 minutes.Αντίστοιχα ο χρόνος αποτίμησης επερωτήσεων ειναι απο 3-7 seconds.

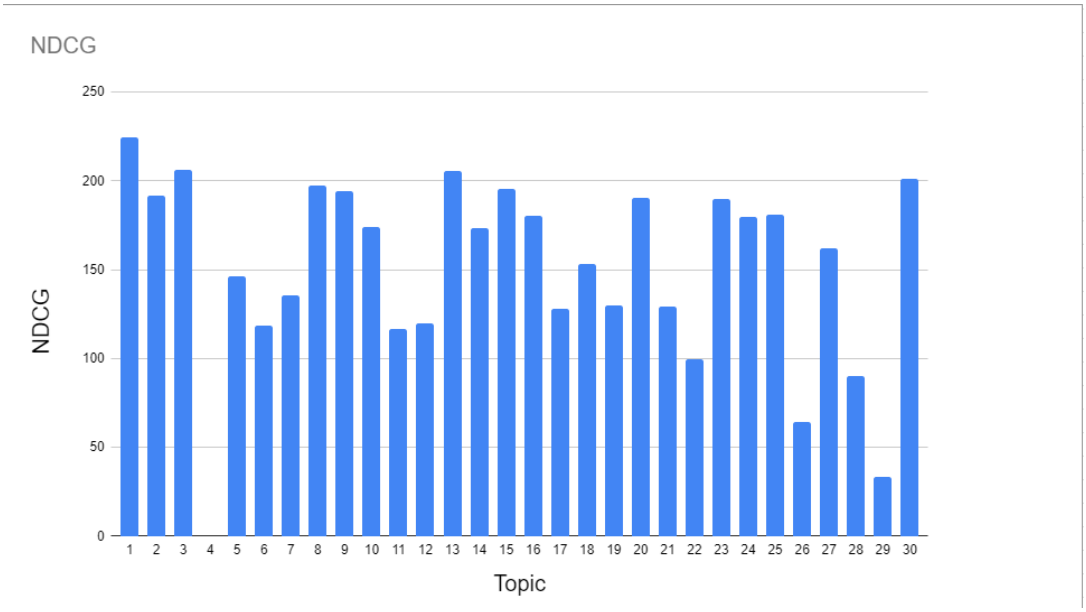
Υπολογίσαμε τα στατιστικά στοιχεία median,average,min και max και τα παραθέτουμε στον παρακάτω πίνακα

	BPREF	AVEP	NDCG
Average	2,384	0,042	150,454
Median	2,488	0,031	167,776
Max	4,722	0,168	224,592
Min	0	0,010	0

Αντίστοιχα για την κατασκευή των γραφημάτων χρησιμοποίησαμε τις δυνατότητες του Google Sheets και σας παραθέτουμε τα παρακάτω διαγράμματα.







4 Επίλογος

Εκτέλεσαμε πιστά τις οδηγίες των εκφωνήσεων για την δημιουργία ενός συστήματος ανάκτησης πληροφοριών και ενός υποσυστήματος για την αξιολόγηση του. Βρήκαμε την εργασία αυτή αρκετά ενδιαφέρουσα και αφιερώσαμε αρκέτο χρόνο στην κατανοήση και υλοποίηση της.

5 Αναφορές

Χρησιμοποιήσαμε αρκετές γνώσεις από την εμπειρία μας στην γλώσσα Java στο τεχνικό κομμάτι ενώ στο θεωρητικό αντλήσαμε γνώσεις από το μάθημα και τις σημειώσεις.

6 Οδηγίες

Για την πρώτη φάση δημιουργήσαμε ένα queryevaluator.jar το οποίο τρέχει από το command line με την εντολή ”java -jar queryevaluator” και πρέπει να είναι στο ίδιο directory με το “5_Resources_Corpus” directory ώστε να διαβάσει τα topics. Δεν περιλαμβάνεται το MiniCollection στο “5_Resources_Corpus” καθώς έβγαινε μεγάλο το .zip

Αντίστοιχα για την δεύτερη φάση πρέπει να ακολουθηθούν οι προηγούμενες οδηγίες όσον αφορά το “5_Resources_Corpus” directory και για την εκκίνηση της αξιολόγησης(γραφικό περιβάλλον) πρέπει ο χρήστης να περιμένει το τέλος του Indexing.