Πανεπιστήμιο Κρήτης
Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 Συστήματα Ανάκτησης Πληροφοριών
Εξάμηνο: Άνοιξη 2019

Γραπτή Αναφορά Έργου

Στοιχεία Φοιτητών

Μέλος	1ο
Ονοματωπώνυμο	Κατσιφαράκης Εμμανουήλ
ΑΜ	3195
Email	csd3195@csd.uoc.gr

Μέλος	2ο
Ονοματωπώνυμο	Κανάρια Ιωάννα
ΑΜ	3609
Email	csd3609@csd.uoc.gr

Πίνακας Περιεχομένων

- *Εισαγωγή*

Στο project του μαθήματος, ασχοληθήκαμε με την δημιουργία και την αξιολόγηση, ενός δικού μας Συστήματος Ανάκτησης Πληροφοριών. Αναλυτικότερα, στην Α' Φάση της εργασίας, κατασκευάσαμε ένα ανεστραμμένο ευρετήριο, πάνω στο οποίο κάναμε επερωτήσεις σε μια πραγματική συλλογή βιοϊατρικών άρθρων. Επιπροσθέτως, στην Β' Φάση της εργασίας, αξιολογήσαμε την αποτελεσματικότητα του συστήματος μας, βασιζόμενοι σε 30 ιατρικά θέματα και καταγράφοντας μετρήσεις, βάσει των μέτρων bpref (μέτρηση με βάση το άθροισμα του αριθμού των σχετικών εγγράφων που κατατάσσονται πριν από μη συναφή έγγραφα), AveP (υπολογισμος της τιμής της ακρίβειας σε κάθε δυνατό recall level) και nDCG (κανονικοποιημένο μειωμένο σωρευτικό κέρδος).

Ολοκληρώσαμε επιτυχώς όλα τα ερωτήματα της εκφώνησης για την Β' Φάση, ενώ διορθώσαμε την διαδικασία παραγωγής της νόρμας κάθε εγγράφου, την οποία δεν είχαμε επιτύχει στην Α' Φάση.

Τα αδύνατο σημείο της εργασίας μας, είναι ότι η διαδικασία της ευρετηρίασης δεν είναι ιδιαίτερα γρήγορη όπως επίσης ότι κατά την αξιολόγηση του συστήματος με βάση τα μέτρα bpref, AveP, nDCG, είχαμε πολλά μηδενικά αποτελέσματα.

Όσον αφορά τις χρονικές επιδόσεις του συστήματος, η διαδικασία επερωτήσεων διήρκησε 25 λεπτά και οι μετρήσεις έγιναν σε σύστημα:
CPU: AMD FX-8320 3.5Ghz
Ram: 8GB DDR3-1600MHz
SSD: 120Gb Samsung 840 Evo

Η εργασία έγινε συνεργατικά από τα μέλη της ομάδας, οπότε και τα δύο μέλη ασχολήθηκαν με όλα τα κομμάτια της εργασίας.

- *Τύποι και μέτρα που χρησιμοποιήθηκαν*

1. *Τύποι για την δημιουργία του ανεστραμμένου ευρετηρίου – Φάση Α*

<u>*Κανονικοποίηση*</u>
tf_{ij} =$freq_{ij}$ /max_k {$freqk_j$} ,
όπου maxk {freqkj} το μεγαλύτερο πλήθος εμφανίσεων ενός όρου στο έγγραφο j.

<u>*Βάρος όρου i στο έγγραφο j*</u>
w_{ij} = tf_{ij}*df_i,
όπου df_i είναι το πλήθος εγγράφων που περιέχουν τον όρο i.

<u>*Νόρμα εγγράφου*</u>
$|w_{ij}|=\sqrt{(\sum_{1<=i<=N}(w_{ij}^2))}$

2. *Μέτρα για την αξιολόγηση του συστήματος – Φάση Β*

<u>*bPref (binary preference-based measure) [1]*</u>

$$\text{bpref} = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{R}$$

όπου R είναι ο αριθμός των συναφών εγγράφων και r είναι το πρώτο συναφές έγγραφο που εντοπίζουμε μετά από n μη συναφή έγγραφα.

AveP (Average Precision)

$$AP = \frac{1}{\Sigma} \sum_{i=1}^{|E|} rel(i) P@i$$

$$P@k = \frac{|E_k \cap \Sigma|}{k}$$

όπου E είναι ο αριθμός των ευρεθέντων εγγράφων, $E_κ$ είναι το ευρεθέν έγγραφο στην θέση k, Σ είναι ο αριθμός των συναφών εγγράφων και rel(i) είναι μία συνάρτηση η οποία μας επιστρέφει 1 αν το έγγραφο i είναι συναφές, διαφορετικά 0.

nDCG(Normalized Discounted Cumulative Gain)

Για i < 2

$$DCG = \sum_{i=1}^{E} (rel(i))$$

Για i > 2

$$DCG = \sum_{i=1}^{E} \left(\frac{rel(i)}{log_2(i)}\right)$$

Για i < 2

$$IDCG = \sum_{i=1}^{\Sigma} (rel(i))$$

Για i > 2

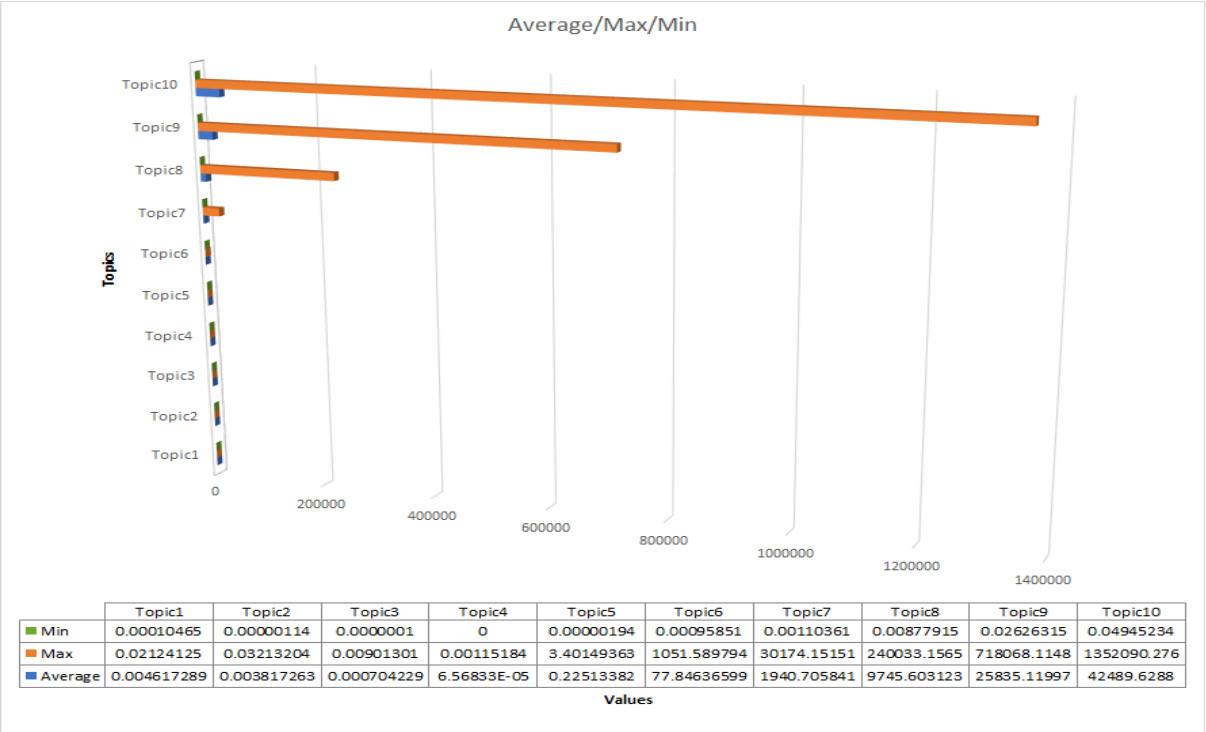
$$IDCG = \sum_{i=1}^{\Sigma} \left(\frac{rel(i)}{log_2(i)}\right)$$

$$NDCG = \frac{DCG}{\text{IDCG}}$$

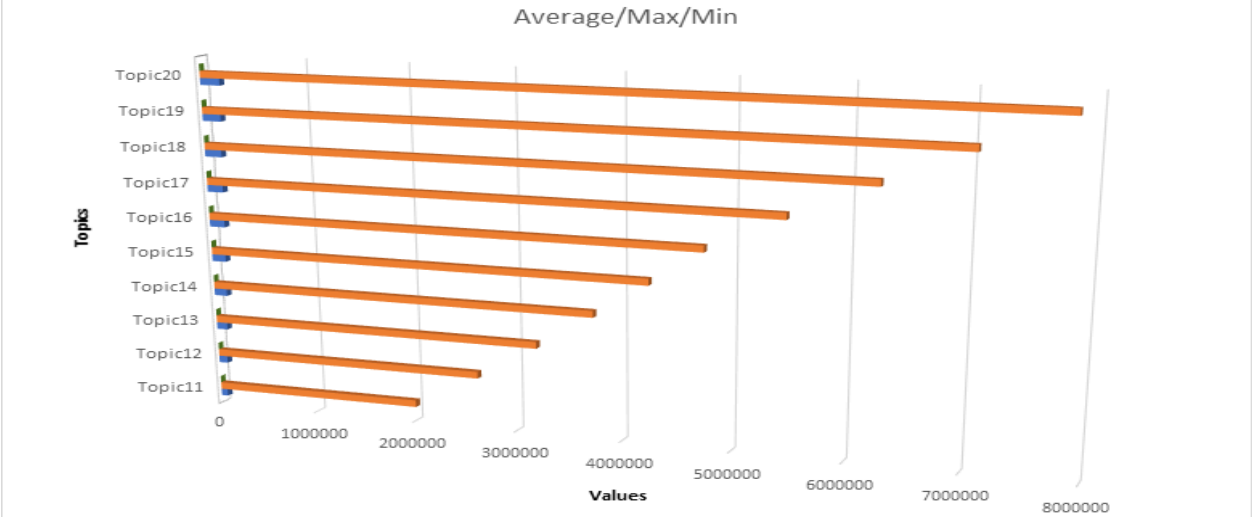
όπου E είναι τα ευρεθέντα έγγραφα, Σ είναι τα συναφή έγγραφα, και rel(i) είναι μία συνάρτηση η οποία μας επιστρέφει 1 αν το έγγραφο i είναι συναφές, διαφορετικά 0.

- *Μετρήσεις*

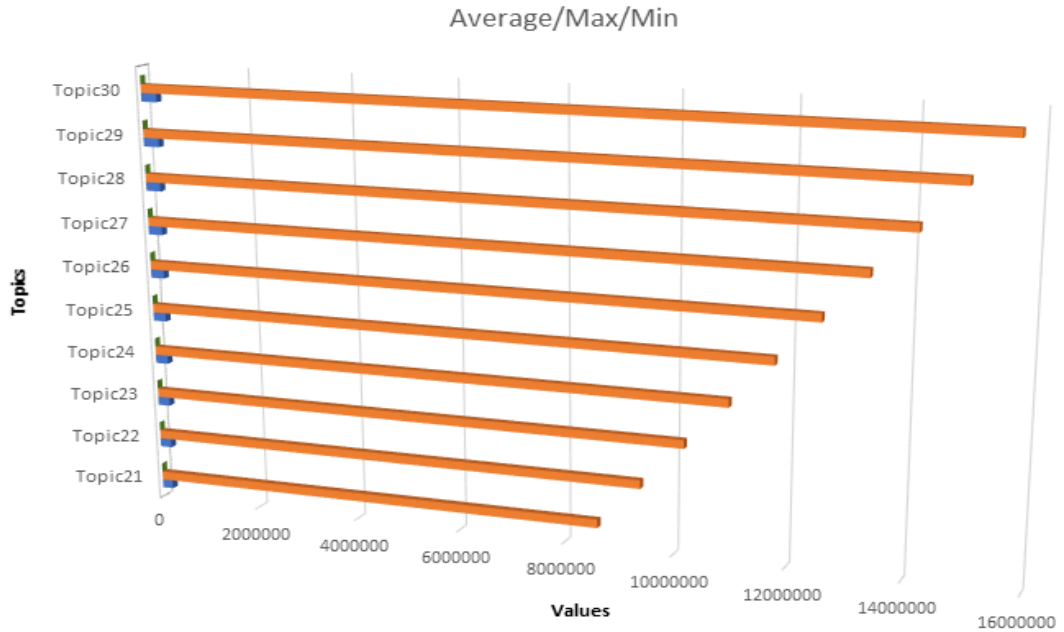
Υπολογίσαμε τις min, max και average τιμές για κάθε topic από το results.txt αρχείο. Παρακάτω παρουσιάζονται τα δεδομένα που εξάγαμε διαγραμματικά:



	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
Min	0.00010465	0.00000114	0.0000001	0	0.00000194	0.00095851	0.00110361	0.00877915	0.02626315	0.04945234
Max	0.02124125	0.03213204	0.00901301	0.00115184	3.40149363	1051.589794	30174.15151	240033.1565	718068.1148	1352090.276
Average	0.004617289	0.003817263	0.000704229	6.56833E-05	0.22513382	77.84636599	1940.705841	9745.603123	25835.11997	42489.6288



	Topic11	Topic12	Topic13	Topic14	Topic15	Topic16	Topic17	Topic18	Topic19	Topic20
Min	0.07345709	0.09609354	0.11705255	0.13668928	0.15543831	0.13515212	0.08133258	0.03043006	0.06090312	0.01920317
Max	2008410.94	2627320.57	3200366.7	3737259.9	4249882.46	4747430.66	5465501.58	6254781.97	7041600.94	7826686.85
Average	64400.3956	84076.7003	105757.956	125653.179	138049.793	151309.098	158417.636	167749.291	183599.464	195233.032

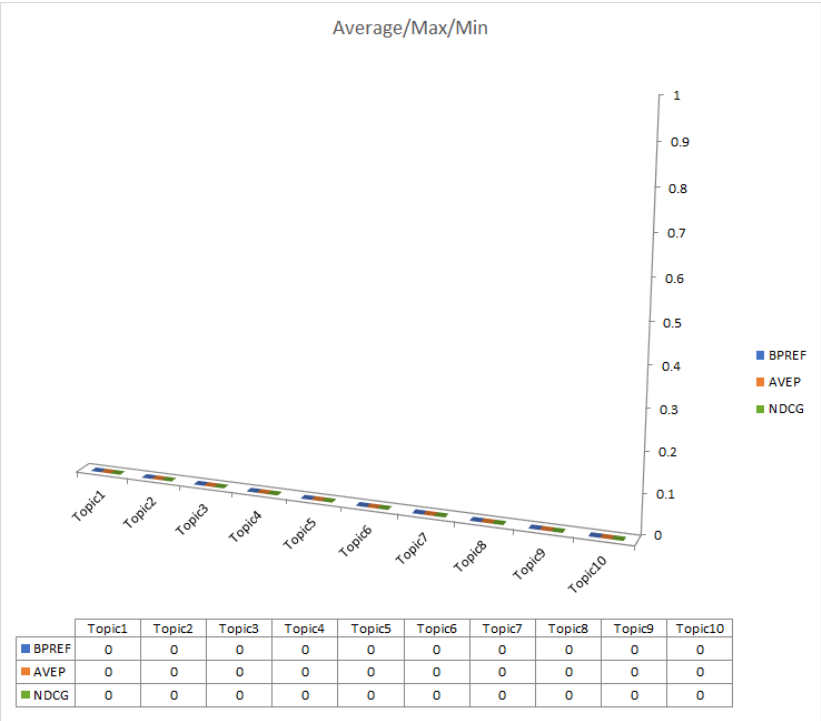


	Topic21	Topic22	Topic23	Topic24	Topic25	Topic26	Topic27	Topic28	Topic29	Topic30
Min	0.0384123	0.0576245	0.016834	0.0336683	0.0505028	0.0345301	0.0690605	0.1010084	0.1178442	0.0027104
Max	8610680.4	9394471.6	10177848	10960899	11743895	12527009	13309986	14093053	14876140	15658722
Average	206087.71	219912.14	232108.47	242512.25	250682.13	261876.71	271027.25	285908.58	303025.72	305945.74

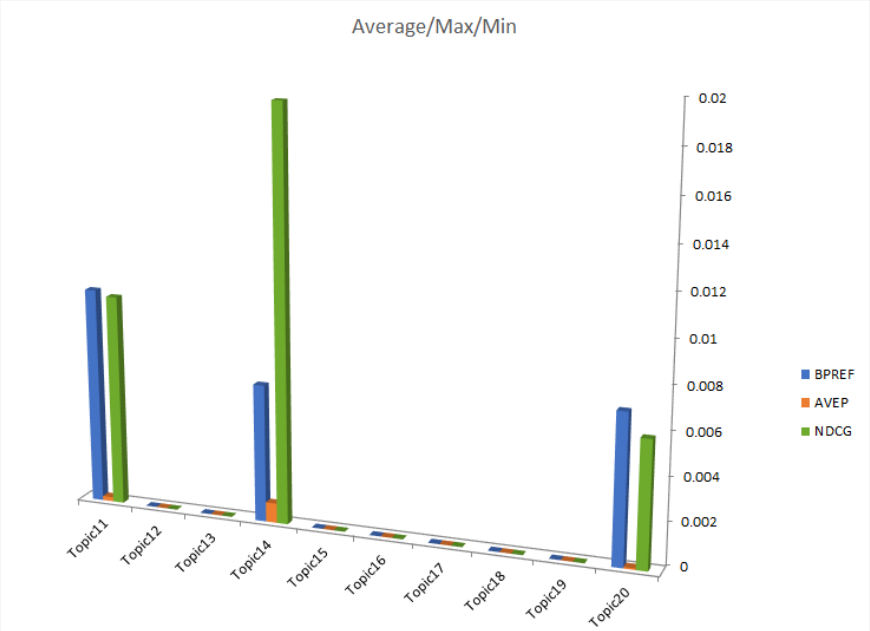
Παρατηρούμε ότι όσο προχωράμε στην λίστα των topics, τόσο αυξάνεται η μέση τιμή των scores όπως και η μέγιστη τιμή αντίστοιχα.

Σχετικά με τα μέτρα bpref, AveP και nDCG, παρατηρήσαμε ότι έχουμε πολλές μηδενικές τιμές όσον αφορά τον βαθμό συνάφειας. Μετά από δοκιμές στον κώδικα μας, είδαμε ότι σε ορισμένα ιατρικά θέματα, ενώ τα ευρεθέντα έγγραφα που μας επιστρέφει το σύστημα, εντοπίζονται στα συναφή του αρχείου qrels.txt, έχουν όλα βαθμό συνάφειας 0, με αποτέλεσμα αρκετές φορές οι τιμές των μέτρων να μηδενίζονται. Επίσης, σε άλλες περιπτώσεις, η τομή των ευρεθέντων εγγράφων με τα συναφή, είναι το κενό σύνολο και αυτός είναι άλλος ένας λόγος που τα μέτρα είναι μηδενικά.

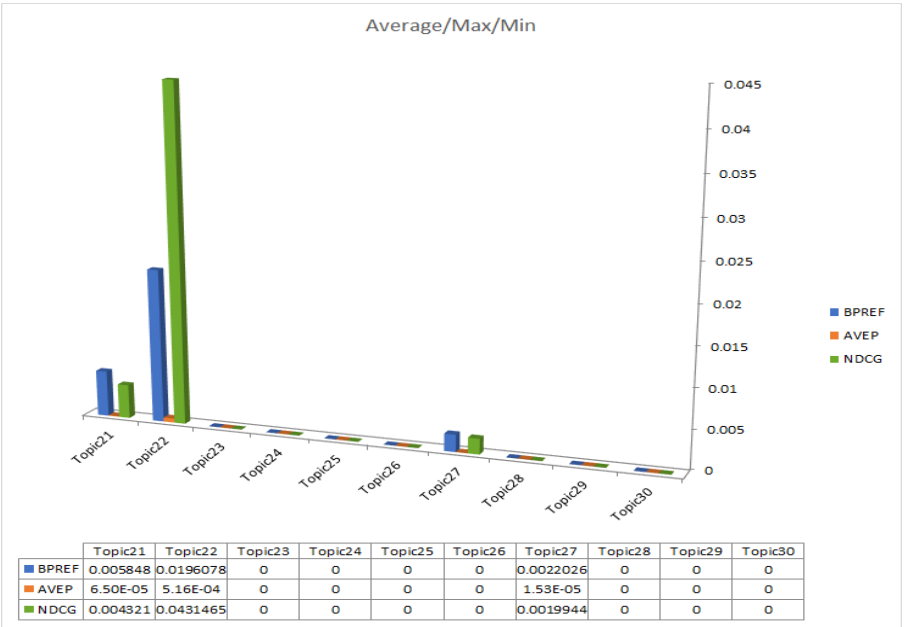
Παρακάτω παρουσιάζονται διαγράμματα για τα μέτρα bpref, AveP και nDCG, για κάθε ένα από τα 30 ιατρικά θέματα:



	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
BPREF	0	0	0	0	0	0	0	0	0	0
AVEP	0	0	0	0	0	0	0	0	0	0
NDCG	0	0	0	0	0	0	0	0	0	0



	Topic11	Topic12	Topic13	Topic14	Topic15	Topic16	Topic17	Topic18	Topic19	Topic20
BPREF	0.01	0	0	0.0063694	0	0	0	0	0	0.0069444
AVEP	1.96E-04	0	0	9.10E-04	0	0	0	0	0	7.16E-05
NDCG	0.0097649	0	0	0.0190911	0	0	0	0	0	0.0058688



	Topic21	Topic22	Topic23	Topic24	Topic25	Topic26	Topic27	Topic28	Topic29	Topic30
BPREF	0.005848	0.0196078	0	0	0	0	0.0022026	0	0	0
AVEP	6.50E-05	5.16E-04	0	0	0	0	1.53E-05	0	0	0
NDCG	0.004321	0.0431465	0	0	0	0	0.0019944	0	0	0

- *Απαιτούμενα αρχεία για την λειτουργία του συστήματος*

Σχετικά με το indexing και το queryevaluator , θα πρέπει να υπάρχει ένας φάκελος με το όνομα stoplists , ο οποίος θα περιέχει τα αρχεία stopwordsEn.txt , stopwordsGr.txt και θα βρίσκεται στον ίδιο φάκελο με τα indexer.jar queryevaluator.jar καθώς χρειάζονται για την αφαίρεση των stopwords. Για το queryevaluator.jar χρειαζόμαστε επίσης το αρχείο qrels.txt και τον έναν φάκελο με όνομα topics ο οποίος θα περιέχει το αρχείο topics.xml καθώς και τον φάκελο CollectionIndex ο οποίο περιέχει τα αρχεία VocabularyFile.txt, PostingFile.txt και DocumentsFile.txt τα οποία παράχθηκαν από τον Indexer. Όλα τα παραπάνω αρχεία θα πρέπει να βρίσκονται στο ίδιο path με τα jar αρχεία.

- *Επίλογος*

Συνοψίζοντας, όπως μπορεί κανείς να δει και στα διαγράμματα, το σύστημα μας είχε αρκετά καλή απόδοση, στις περιπτώσεις όπου η τομή των ευρεθέντων εγγράφων με τα συναφή δεν είναι το κενό σύνολο ή ο βαθμός συνάφειας των ευρεθέντων εγγράφων δεν είναι παντού μηδενικός. Η αποτυχία του συστήματος στις περιπτώσεις μηδενισμού, είναι πιθανόν να οφείλεται στον λάθος υπολογισμό της νόρμας και αυτό να έχει ως συνέπεια το σύστημα μας να επιστρέφει λάθος έγγραφα ως ευρεθέντα για κάποια επερώτηση.

- *Αναφορές*

1. Chris Buckley and Ellen M. Voorhees, "Retrieval evaluation with incomplete information.", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.

2. Tetsuya Sakai, "Alternatives to bpref.", Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.