2019-2020

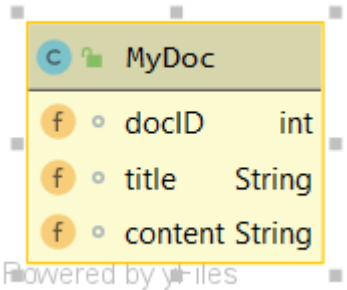
Συστήματα Ανάκτησης Πληροφοριών

Προγραμματιστική Εργασία

Επέκταση Ερωτημάτων με Συνώνυμους Όρους για τη Βελτίωση των Αποτελεσμάτων της Ανάκτησης

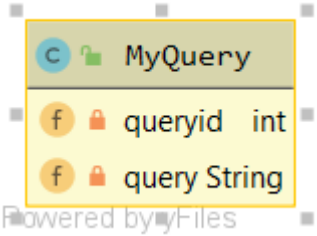
1η ΦΑΣΗ-Baseline

Ξεκινώντας με το 1ο βήμα και το βήμα 2ο τα οποία αφορούσαν την προεπεξεργασία της συλλογής και την δημιουργία του ευρετηρίου ,δημιούργησα μία κλάση Indexer στην οποία χρησιμοποιώντας τον English Analyzer και τον BM25 για το similarity δημιούργησα το index. Η μέθοδος createIndex κάνει parse τα documents μέσω της κλάσης TXTparsing και της μεθόδου parse όπου παίρνει σαν όρισμα το αρχείο docs/documents.txt(Πρέπει το documents.txt να είναι στο φάκελο docs). Μέσα στην parse χωρίζω το αρχείο όπου υπάρχουν οι χαρακτήρες /// και έτσι κάνοντας τους trim ,χωρίζω τον τίτλο(title) ,το docID και το περιεχόμενο του document(content) και τα εισάγω σε ένα αντικείμενο τύπου myDoc. Επιστρέφοντας στην createIndex καλώ για κάθε document την μέθοδο indexDoc όπου εκεί έχω 3 StoreFields που περνάω τα πεδία του myDoc και 1 TextField όπου βάζω τον *τίτλο* και το *περιεχόμενο* του document. Έτσι, προσθέτω κάθε document στον indexWriter ώστε να ολοκληρωθεί η διαδικασία και να αποθηκευτεί το index.Το ευρετήριο θα δημιουργηθεί στο *φάκελο index*.



Εικόνα 1:Η κλάση MyDoc

Στο 3ο βήμα το οποίο αφορούσε την εκτέλεση των ερωτημάτων έχω υλοποιήσει την κλάση Searcher η οποία ανοίγει το index και ψάχνει και εμφανίζει για κάθε query τα TopDocs και τα Scores. Για να μπορέσω να διαβάσω τα queries από το αρχείο queries.txt δημιούργησα την μέθοδο parseQueries στην κλάση TXTparsing η οποία παίρνει ως όρισμα τη θέση του queries.txt (το οποίο πρέπει να βρίσκεται στον φάκελο docs) και



Εικόνα 2.Η κλάση MyQuery

όπως και πριν με την μέθοδο parse χωρίζει με το κάθε query ανά /// και δημιουργεί αντικείμενα τύπου MyQuery το οποίο έχει ως πεδία το queryID και το ίδιο query. Έτσι, για να πάρω αποτελέσματα για διαφορετικά κ=20,30 και 50 άλλαξα στις γραμμές 59 και 67 n ώστε να έχω τα διαφορετικά αποτελέσματα για κάθε κ(τα αποτελέσματα αποθηκεύονται στον φάκελο docs με όνομα resultsk.txt,όπου κ=20,30,50)

```
58          List<MyQuery> queries=TXTparsing.parseQueries(queriestxt);
59          FileWriter fileWriter=new FileWriter( fileName: "docs/results20.txt", append: true);
60          BufferedWriter bufferedWriter=new BufferedWriter(fileWriter);
61          for(MyQuery q:queries){
62              // parse the query according to QueryParser
63              Query query = parser.parse(q.getQuery());
64              System.out.println("Searching for: " + query.toString(field));
65
66              // search the index using the indexSearcher
67              TopDocs results = indexSearcher.search(query, n: 20);
```

Εικόνα 3. Οι αλλαγές στις γραμμές για την δημιουργία αποτελεσμάτων για κάθε κ

Στο 4ο βήμα παίρνοντας τα αποτελέσματα results.txt χρησιμοποίησα το εργαλείο trec_eval (το οποίο το έχω βάλει στο φάκελο docs)δίνοντας ως command τις εξής εντολές:

- trec_eval -q -M 5 -m map -m num_rel_ret qrels.txt results20.txt>answers/answers5.txt
- trec_eval -q -M 10 -m map -m num_rel_ret qrels.txt results20.txt>answers/answers10.txt
- trec_eval -q -M 15 -m map -m num_rel_ret qrels.txt results20.txt>answers/answers15.txt
- trec_eval -q -m map -m num_rel_ret qrels.txt results20.txt>answers/answers20.txt
- trec_eval -q -m map -m num_rel_ret qrels.txt results30.txt>answers/answers30.txt
- trec_eval -q -m map -m num_rel_ret qrels.txt results50.txt>answers/answers50.txt

Η παράμετρος -M ,διαβάζοντας και στο documentation του trec_eval ,βλέπουμε πως είναι :

```
-Max_retrieved_per_topic num:
 -M <num>: Max number of docs per topic to use in evaluation (discard rest).
      Default is MAX_LONG.
```

Τα αποτελέσματα του trec_eval αποθηκεύονται στο φάκελο /docs/answers/ και αναλόγως με το k=5,10,15,20,30,50 έχω answersk.txt

Παρακάτω παραθέτω τον ολοκληρωμένο πίνακα για κάθε ερώτημα με τις απαντήσεις του trec_eval:

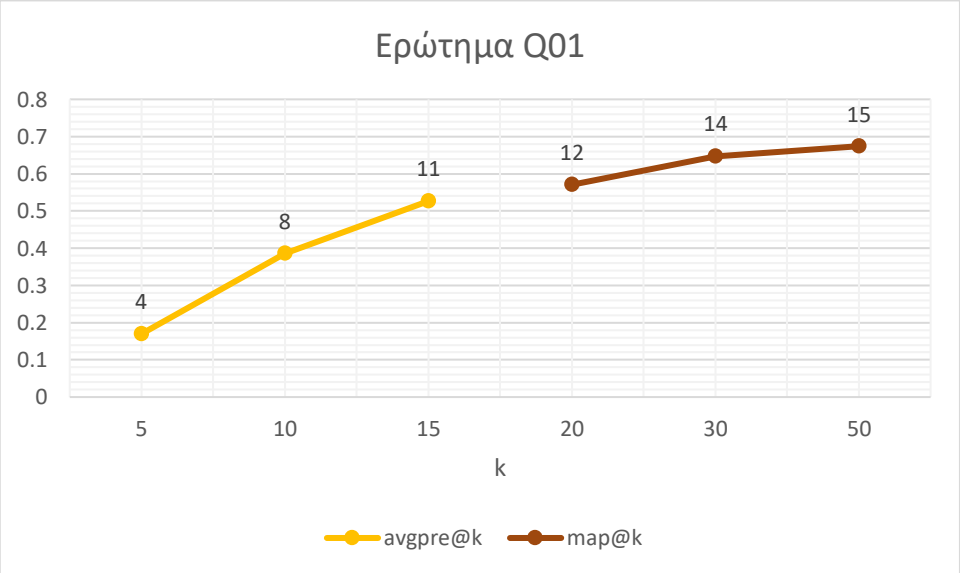
Query	k	avgpre@k	num_rel_ret	map@k
Q01	5	0.1698	4	
	10	0.3857	8	
	15	0.5265	11	
	20		12	0.5706
	30		14	0.6473
	50		15	0.6741
Q02	5	0.1389	2	
	10	0.1746	3	
	15	0.1746	3	
	20		3	0.1746
	30		3	0.1746
	50		3	0.1746
Q03	5	0.2536	4	
	10	0.3743	6	
	15	0.3743	6	
	20		8	0.4322
	30		10	0.4798
	50		14	0.5689
Q04	5	0.0464	2	
	10	0.0464	2	
	15	0.0607	3	
	20		3	0.0607
	30		3	0.0607
	50		4	0.0694
Q05	5	0.1	3	
	10	0.1	3	
	15	0.1	3	

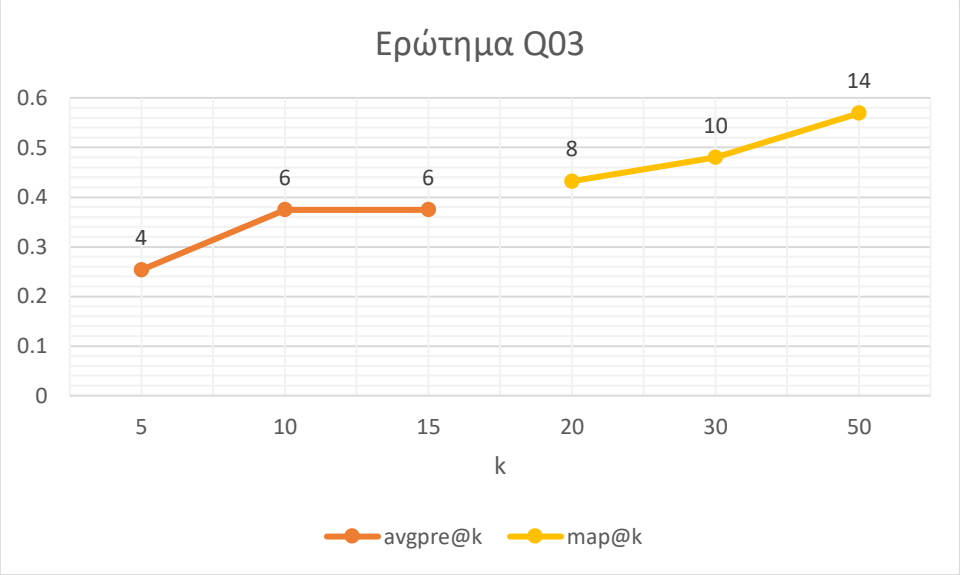
	20		5	0.1295
	30		9	0.198
	50		12	0.2489
Q06	5	0.0263	1	
	10	0.0263	1	
	15	0.0263	1	
	20		1	0.0263
	30		1	0.0263
	50		6	0.0504
Q07	5	0.0625	1	
	10	0.0764	2	
	15	0.0934	3	
	20		3	0.0934
	30		9	0.1868
	50		12	0.2379
Q08	5	0.3571	5	
	10	0.5714	8	
	15	0.5714	8	
	20		11	0.6929
	30		11	0.6929
	50		11	0.6929
Q09	5	0.131	3	
	10	0.15	4	
	15	0.1698	5	
	20		5	0.1698
	30		7	0.1927
	50		9	0.2107
Q10	5	0.05	1	

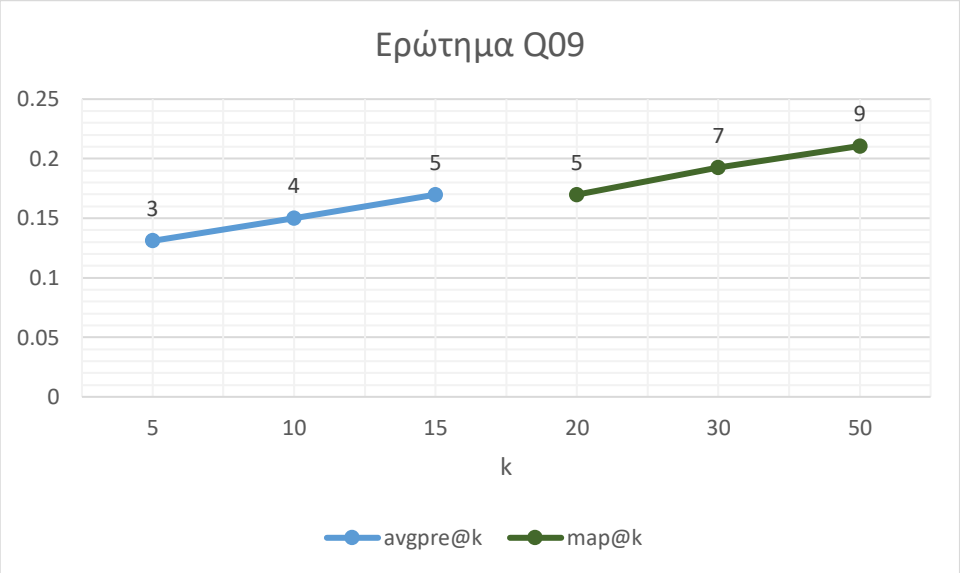
	10	0.0722	2	
	15	0.0722	2	
	20		2	0.0722
	30		3	0.0826
	50		3	0.0826
ALL	5	0.1336	26	
	10	0.1977	39	
	15	0.2169	45	
	20		53	0.2422
	30		70	0.2742
	50		89	0.301

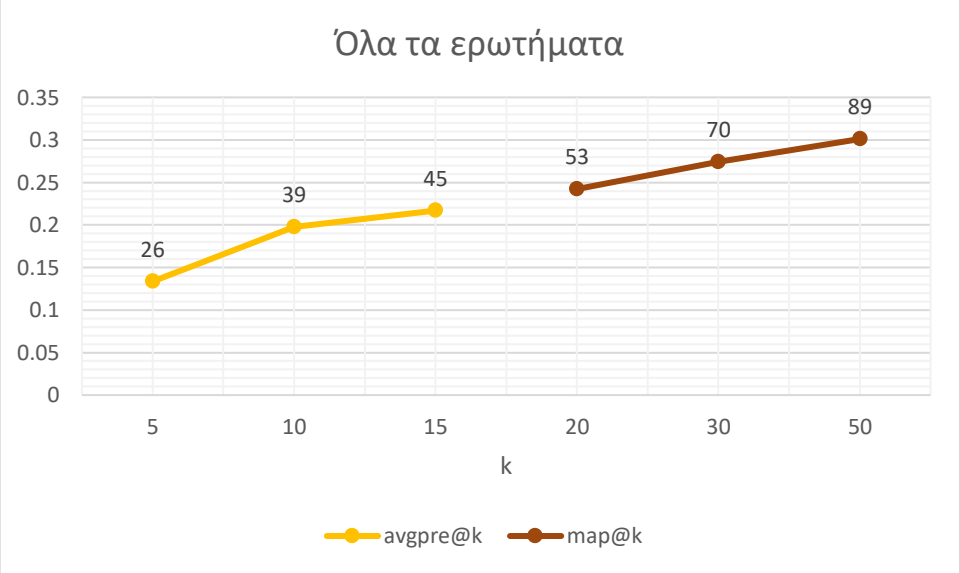
Τα πεδία όπου είναι κενά είναι ίδια με avgpre@k=map@k.

Παρακάτω παραθέτω μερικά γραφήματα με τα πιο σημαντικά ερωτήματα όπου βλέπουμε κάποια εμφανή διαφορά καθώς το κ μεταβάλλεται. Πάνω από κάθε σημείο εμφανίζονται τα σχετικά επιστρεφόμενα κείμενα.




Ερώτημα Q03
0.6
0.5
0.4
0.3
0.2
0.1
0
4
6
6
8
10
14
5
10
15
20
30
50
k
avgpre@k
map@k


Ερώτημα Q09
0.25
0.2
0.15
0.1
0.05
0
3
4
5
5
7
9
5
10
15
20
30
50
k
avgpre@k
map@k

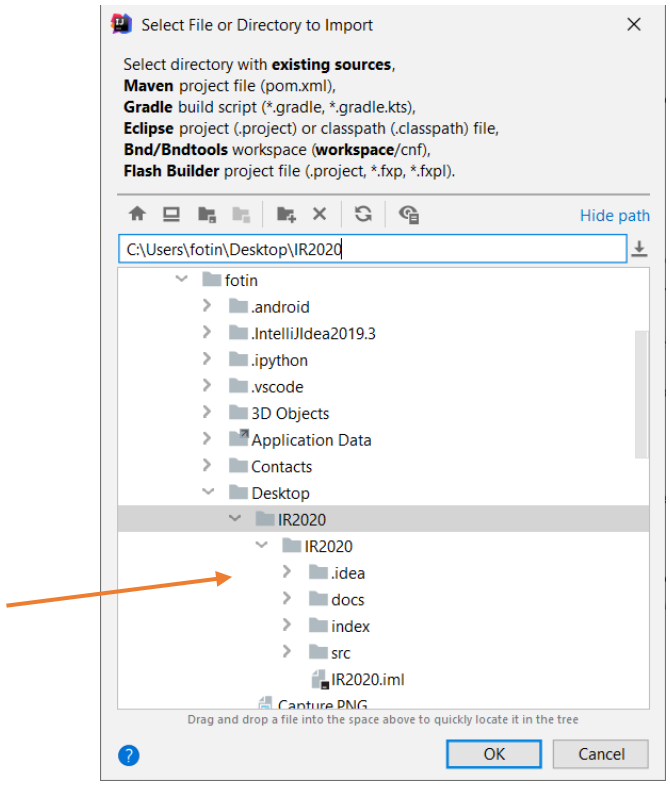

Όλα τα ερωτήματα
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
26
39
45
53
70
89
5
10
15
20
30
50
k
avgpre@k
map@k

Ο λόγος που δεν τα έκανα σαν ενιαία γραμμή τα avgpre@k και map@k είναι απλά για να φαίνεται η διαφορά τους για ποια κ είναι υπεύθυνα.

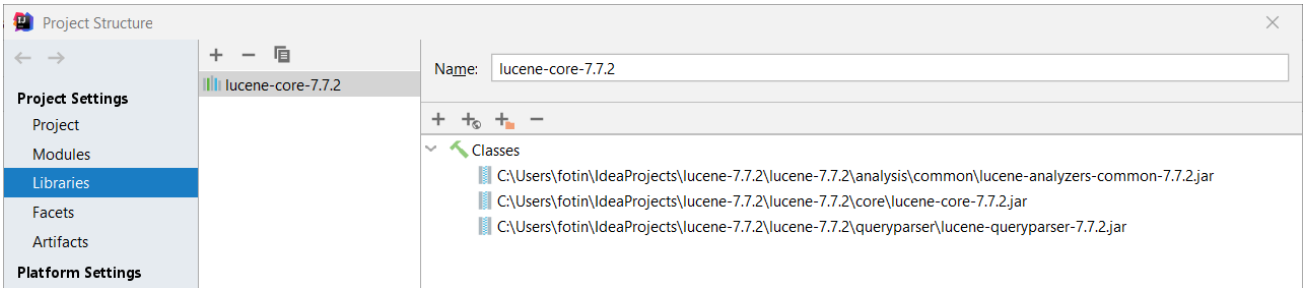
Οδηγίες Φόρτωσης

Προκειμένου να φορτώσετε την εργασία :

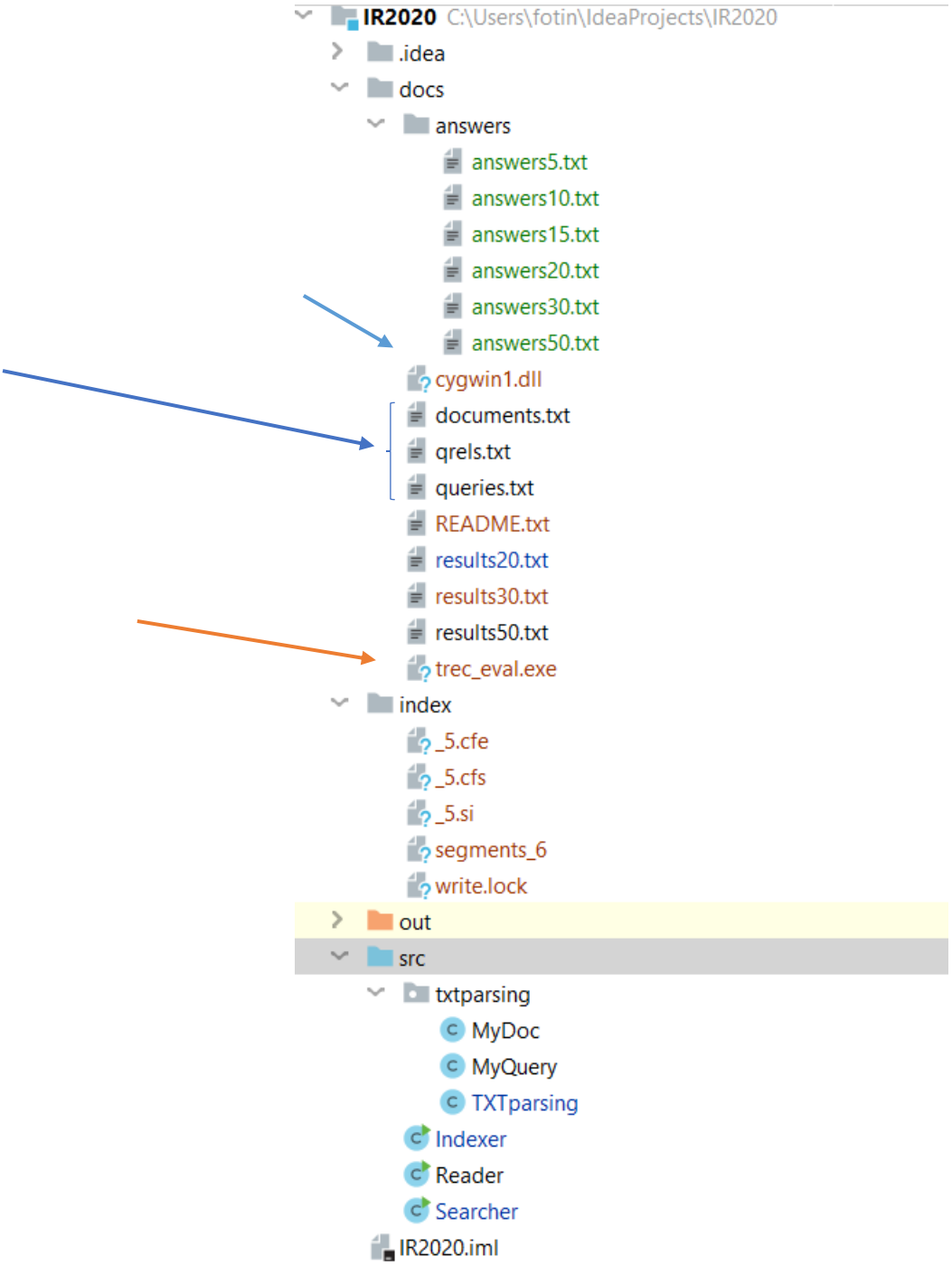
1. κάνετε unzip το αρχείο και μέσω του Intellij επιλέγετε να φορτωθεί File>New>Project from Existing Sources όπου από εκεί επιλέγετε το φάκελο που έγινε unzip.



Εάν δεν έχουν φορτωθεί οι βιβλιοθήκες αυτόματα , επιλέγετε CTLR+ALT+SHIFT+S και μέσω του project structure επιλέγετε libraries και εισάγετε την lucene όπως στο 2^{ο} εργαστήριο.



2. Το ευρετήριο δεν έχει παραμείνει επομένως πρέπει να τρέξετε τον Indexer αφού πρώτα στο φάκελο docs εισαχθούν τα αρχεία qrels.txt ,queries.txt ,documents.txt καθώς και αν θέλετε να τρέξετε και το εργαλείο trec_eval πρέπει να μπει στο φάκελο docs.
3. Η τελική διάταξη του project πρέπει να είναι όπως στην παρακάτω εικόνα:



Η κλάση Reader υπάρχει ώστε να διαβάζω το ευρετήριο μου , δεν χρειάζεται σε κάποιο μέρος της εκτέλεσης.

Πηγές:

- 2ο εργαστήριο ,Συστήματα Ανακτησης Πληροφοριών,2019-2020
- Readme, trec_eval, https://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/A.README