ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

1η Προγραμματιστική Εργασία στο μάθημα «Συστήματα Ανάκτησης Πληροφορίας»

Η παρούσα εργασία εκπονήθηκε ατομικά.

Ονοματεπώνυμο	Αριθμός Μητρώου	E-mail επικοινωνίας
Γεώργιος Μοσχόβης	3150113	p3150113@aueb.gr

Χρήσιμες πληροφορίες για την εργασία:

- ✓ Όνομα εκτελέσιμου αρχείου: **Tests.java**, στο φάκελο src/main/java (ή Tests.class, στο φάκελο target\classes).
- ✓ Λόγω εξαρτήσεων με τις βιβλιοθήκες σχετικές με την Elastic Search παραδίδεται **IntelliJ Idea Maven Project**.
- ✓ Για την εκτέλεση αξιολόγησης trec_eval παραδίδεται Linux Bash Script, ακολουθούν οδηγίες εκτέλεσης.

Georgios M. Moschovis
Undergraduate Computer Science Student
Athens University of Economics & Business
76, 28is Oktovriou Str., 10434 Athens, Greece
website: www.linkedin.com/in/georgios-moschovis-96428029
mail: p3150113@aueb.gr, georgiosmoshovis@hotmail.gr

1. Εισαγωγή

Στην παρούσα εργασία, ανέπτυξα ένα μηχανισμό ευρετηρίασης και διενέργειας ερωτημάτων, δημιουργώντας Elastic Search Client, μέσω του ολοκληρωμένου περιβάλλοντος ανάπτυξης (IDE: Integrated Development Environment) του **IDE IntelliJ Idea Community**, η ίδια διαδικασία μπορεί όμως να πραγματοποιηθεί μέσω οποιουδήποτε IDE υποστηρίζει Maven Projects με εξαρτήσεις (Maven Dependencies).

Η χρήση εξαρτήσεων ήταν απαραίτητη, λόγω χρήσης πολλαπλών διεπαφών πρόσβασης/χρήσης (APIs: Application Programming Interfaces) για τις απαιτούμενες υπηρεσίες **Lucene Indexes** (ευρετηρίαση, αναζήτηση, κ.λπ) και την δημιουργία συλλογής **.json ισοδύναμων τροποποιημένων** αρχείων με τα δοθέντα, επί των οποίων εφαρμόστηκαν οι παραπάνω αναφερόμενες λειτουργείες.

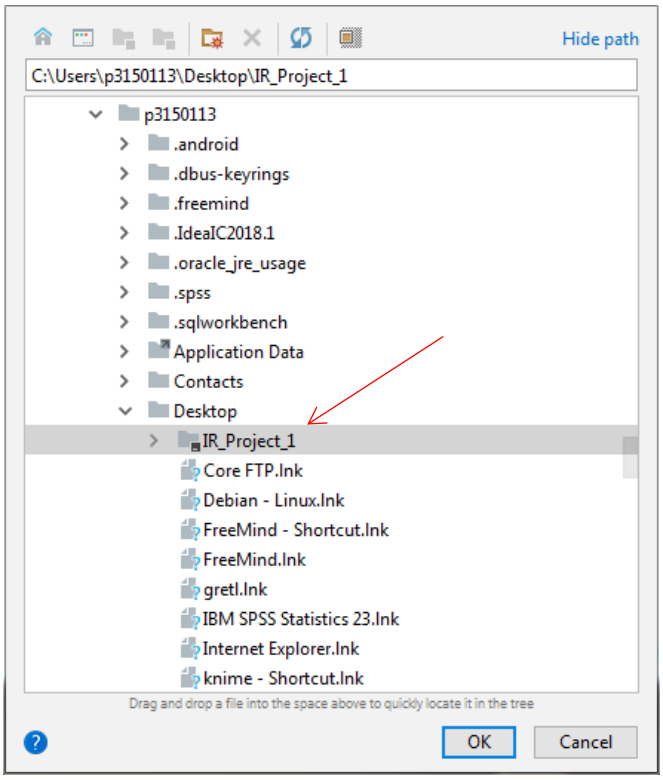
Κατόπιν ολοκλήρωσης της ανωτέρω διαδικασίας, προέβην σε αξιολόγηση για το σύστημα ανάκτησης, μέσω σεναρίου επί του τερματικού (bash script, εκτελούμενο σε ΛΣ Linux Mint), που ενεργοποιεί το trec_eval με κατάλληλες παραμέτρους. Η ίδια διαδικασία επαναλήφθηκε από γραμμή εντολών των Windows.

2. Οδηγίες εκτέλεσης

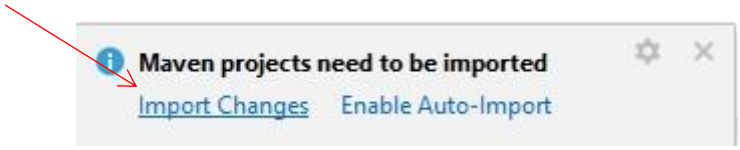
Για να εκτελέσετε την εργασία, θα πρέπει να επιλέξετε άνοιγμα υφιστάμενου Project, στον κατάλογο επιλογών του **IDE IntelliJ Idea Community**, ως φαίνεται στην Εικόνα 2.1. Στη συνέχεια οφείλουν να γίνουν φορτώσεις των εξαρτήσεων και των εξωτερικών χρησιμοποιούμενων βιβλιοθηκών, για την επιτυχή μεταγλώττιση:

i. για τη **φόρτωση βιβλιοθηκών online**, που προκύπτουν απ'το pom.xml αρχείο επιλέγοντας «Import Changes» στο παράθυρο διαλόγου του IntelliJ Idea (βλ. Εικόνα 2.2).
ii. για τη **φόρτωση προεγκατεστημένων βιβλιοθηκών** μέσω ρυθμίσεων, που εμφανίζονται με τη συντόμευση Ctrl+Alt+Shift+S, επιλέγοντας το .jar αρχείο που υπάρχει αναρτημένο στον υποκατάλογο libs, όπως και στην Εικόνα 2.3.

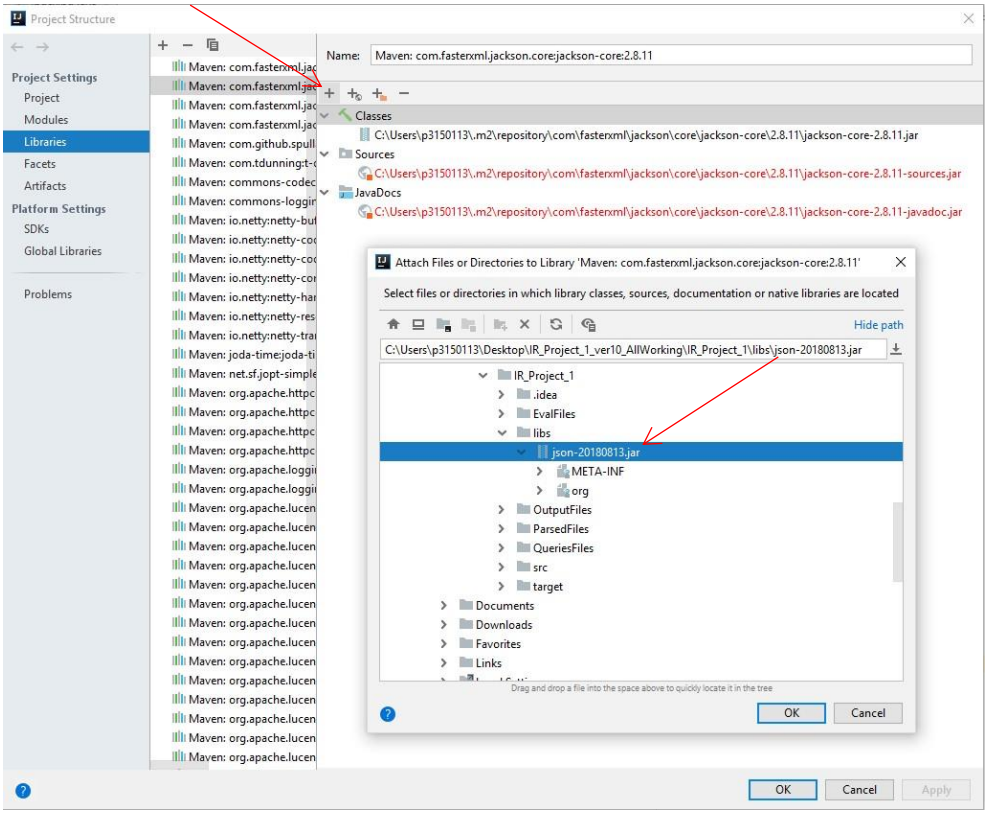
Για την περίπτωση που χρησιμοποιήσετε διαφορετικό περιβάλλον ολοκληρωμένης ανάπτυξης, εκτελέστε τις απολύτως ανάλογες ενέργειες. Εν τούτοις συνίσταται η χρήση IntelliJ Idea Community, επί του οποίου θα γίνει πιθανή εξέταση.



Εικόνα 2.1 Άνοιγμα υφιστάμενου project στο περιβάλλον ανάπτυξης IntelliJ Idea Community



Εικόνα 2.2 Φόρτωση εξαρτώμενων βιβλιοθηκών (Maven Dependencies) βάσει του pom.xml



Εικόνα 2.3 Φόρτωση προεγκατεστημένων βιβλιοθηκών (αρχεία .jar) μέσω ρυθμίσεων

Κατόπιν πραγματοποίησης αυτών των βημάτων, όπως εκκινήσατε την Elastic Search, διαγράφοντας τυχόν προηγούμενα ευρετήρια με την ονομασία «documents» και **εκτελώντας το Tests.java**, που περιλαμβάνει τη δημιουργία JSON συλλογής και το σύνολο των εργασιών πρόσβασης στις υπηρεσίες Lucene.

Αναλυτικά, η μετατροπή των αρχείων .xml στο κατάλληλο μορφότυπο .json για τη ευρετηρίαση, γίνεται από το αρχείο Conversions.java, η δημιουργία ευρετηρίου με όνομα "documents", η ανάλυση των αρχείων και η ευρετηρίασή τους από το αρχείο Indexing.java, ενώ η **πανομοιότυπη ανάλυση** των ερωτημάτων και η αναζήτηση σχετικών κειμένων από το αρχείο Searching.java.

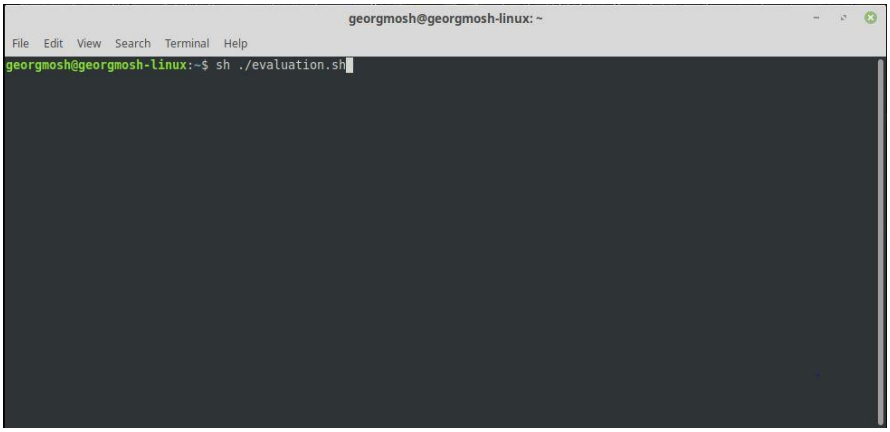
Κατά την ανάλυση, χρησιμοποιήθηκε ο **english analyzer** που παρέχει το αντίστοιχο API (Elastic Search), καθώς περιελάμβανε τις απαιτούμενες μετατροπές. Από τα 21 ανακτηθέντα σχετικά κείμενα, **διαγράφουμε και το 1ο**, διότι είναι «ο εαυτός» του υποβαλλόμενου κειμένου-ερωτήματος από τη συλλογή.

Εικόνα 2.4 Εκτέλεση αρχείου Tests.java, για την ευρετηρίαση των κειμένων και αναζήτηση σχετικών

3. Αξιολόγηση

Για την αξιολόγηση του συστήματος ανάκτησης, χρησιμοποιήθηκε το εργαλείο trec_eval, σε περιβάλλοντα Linux Mint και Windows 10 Education. Τα αποτελέσματα ήταν προφανώς πανομοιότυπα. Η εκτέλεση αξιολόγησης των ανακτηθέντων εγγράφων, όπως έγινε από το σύστημα με παραμέτρους όπως περιγράφονται στην Εικόνα 2.5, μπορεί να επαναληφθεί μέσω δοθέντος bash script.

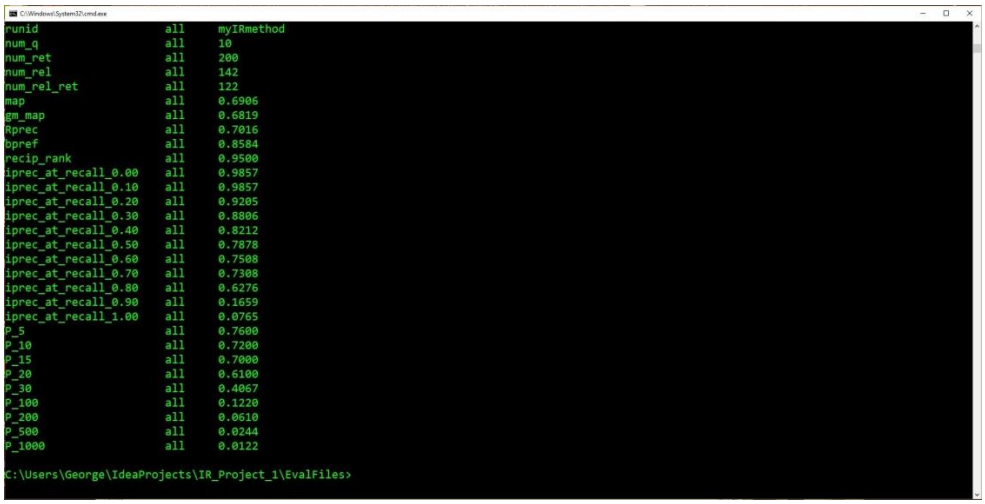
Από τον υποκατάλογο src/EvalFiles, όπου το σύστημα ανάκτησής μου ενέγραψε τα αποτελέσματά του στο αρχείο results.txt, για να κρίνουμε ως προς την απόδοση μπορούμε να τρέξουμε σε οποιοδήποτε περιβάλλον Linux, το δοθέν σενάριο φλοιού (bash script) evaluation.sh, που περιλαμβάνει συνεταγμένη την εντολή εκτέλεσης του trec_eval, που έχει γίνει εκ των προτέρων make και παρατίθεται επίσης στον αναφερόμενο υποκατάλογο, όπως στην Εικόνα 3.1.



Εικόνα 3.2 Εκτέλεση φλοιού αξιολόγησης evaluation.sh σε Linux Mint

Στιγμιότυπο εμφάνισης τον αποτελεσμάτων ακολουθεί στην Εικόνα 3.2, από ισοδύναμη εκτέλεση της αναγραφόμενης εντολής στο αρχείο evaluation.sh σε Windows 10, μέσω της έκδοσης trec_eval.exe που επίσης περιλαμβάνεται συνοδευόμενο από την αναλυτική έξοδο του trec_eval στη γραμμή εντολών.

Τα αναλυτικά αποτελέσματα, όπως προέκυψαν από την εκτέλεση του συστήματος ανάκτησης στον υπολογιστή μου, παρατίθενται επίσης σε .docx μορφή στον ίδιο υποκατάλογο src/EvalFiles, εκτός από παρακάτω στην τρέχουσα αναφορά. Πιστεύω το παρόν κείμενο είναι σχετικά πλήρες. Σε περίπτωση όμως που έχω παραλείψει να περιγράψουμε κάτι που σας προβληματίσει, θα δεχθώ ευχαρίστως να δώσω περαιτέρω διευκρινήσεις, με mail ή εξέταση.



Εικόνα 3.3 Αποτελέσματα μετρήσεων για όλα τα κείμενα στην γραμμή εντολών Windows 10

Με ιδιαίτερη εκτίμηση,

Γεώργιος Μ. Μοσχόβης
Προπτυχιακός φοιτητής πληροφορικής, ειδεικ. σε Βάσεις Δεδομένων και Δίκτυα Υπολογιστών, Οικονομικό Πανεπιστήμιο Αθηνών & Συμμετέχων Huawei Seeds For the Future 2018
Αριστεία προόδου, Ιωάννη Κάβουρα (2016-17) & Μαρίας Δημοπούλου (2015-16), Athens University of Economics & Business & The Real Fake Competition: European Unit & Comission – The European Observatory on Counterfeiting & Piracy
Πατησίων 76, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών –
Mail-to: p3150113@aueb.gr

Evaluation Output:

```
num_ret             Q01     20
num_rel             Q01     15
num_rel_ret         Q01     12
map                 Q01     0.5826
Rprec               Q01     0.7333
bpref               Q01     0.8000
recip_rank          Q01     0.5000
iprec_at_recall_0.00    Q01     0.8571
iprec_at_recall_0.10    Q01     0.8571
iprec_at_recall_0.20    Q01     0.8571
iprec_at_recall_0.30    Q01     0.8571
iprec_at_recall_0.40    Q01     0.8571
iprec_at_recall_0.50    Q01     0.7857
iprec_at_recall_0.60    Q01     0.7857
iprec_at_recall_0.70    Q01     0.7857
iprec_at_recall_0.80    Q01     0.6000
iprec_at_recall_0.90    Q01     0.0000
iprec_at_recall_1.00    Q01     0.0000
P_5                 Q01     0.8000
P_10                Q01     0.7000
P_15                Q01     0.7333
P_20                Q01     0.6000
P_30                Q01     0.4000
P_100               Q01     0.1200
P_200               Q01     0.0600
P_500               Q01     0.0240
P_1000              Q01     0.0120
num_ret             Q02     20
num_rel             Q02     11
num_rel_ret         Q02     9
map                 Q02     0.5661
Rprec               Q02     0.4545
bpref               Q02     0.8182
recip_rank          Q02     1.0000
iprec_at_recall_0.00    Q02     1.0000
iprec_at_recall_0.10    Q02     1.0000
iprec_at_recall_0.20    Q02     0.7500
iprec_at_recall_0.30    Q02     0.7143
iprec_at_recall_0.40    Q02     0.7143
iprec_at_recall_0.50    Q02     0.5625
iprec_at_recall_0.60    Q02     0.5625
iprec_at_recall_0.70    Q02     0.5625
iprec_at_recall_0.80    Q02     0.5625
```

iprec_at_recall_0.90 Q02 0.0000
iprec_at_recall_1.00 Q02 0.0000
P_5 Q02 0.6000
P_10 Q02 0.5000
P_15 Q02 0.5333
P_20 Q02 0.4500
P_30 Q02 0.3000
P_100 Q02 0.0900
P_200 Q02 0.0450
P_500 Q02 0.0180
P_1000 Q02 0.0090
num_ret Q03 20
num_rel Q03 13
num_rel_ret Q03 13
map Q03 0.7219
Rprec Q03 0.6923
bpref Q03 1.0000
recip_rank Q03 1.0000
iprec_at_recall_0.00 Q03 1.0000
iprec_at_recall_0.10 Q03 1.0000
iprec_at_recall_0.20 Q03 0.7647
iprec_at_recall_0.30 Q03 0.7647
iprec_at_recall_0.40 Q03 0.7647
iprec_at_recall_0.50 Q03 0.7647
iprec_at_recall_0.60 Q03 0.7647
iprec_at_recall_0.70 Q03 0.7647
iprec_at_recall_0.80 Q03 0.7647
iprec_at_recall_0.90 Q03 0.7647
iprec_at_recall_1.00 Q03 0.7647
P_5 Q03 0.4000
P_10 Q03 0.7000
P_15 Q03 0.7333
P_20 Q03 0.6500
P_30 Q03 0.4333
P_100 Q03 0.1300
P_200 Q03 0.0650
P_500 Q03 0.0260
P_1000 Q03 0.0130
num_ret Q04 20
num_rel Q04 13
num_rel_ret Q04 10
map Q04 0.5487
Rprec Q04 0.6154

```
bpref                   Q04     0.7692
recip_rank              Q04     1.0000
iprec_at_recall_0.00    Q04     1.0000
iprec_at_recall_0.10    Q04     1.0000
iprec_at_recall_0.20    Q04     1.0000
iprec_at_recall_0.30    Q04     0.6364
iprec_at_recall_0.40    Q04     0.6364
iprec_at_recall_0.50    Q04     0.6364
iprec_at_recall_0.60    Q04     0.6154
iprec_at_recall_0.70    Q04     0.5556
iprec_at_recall_0.80    Q04     0.0000
iprec_at_recall_0.90    Q04     0.0000
iprec_at_recall_1.00    Q04     0.0000
P_5                     Q04     0.6000
P_10                    Q04     0.6000
P_15                    Q04     0.5333
P_20                    Q04     0.5000
P_30                    Q04     0.3333
P_100                   Q04     0.1000
P_200                   Q04     0.0500
P_500                   Q04     0.0200
P_1000                  Q04     0.0100
num_ret                 Q05     20
num_rel                 Q05     15
num_rel_ret             Q05     13
map                     Q05     0.7718
Rprec                   Q05     0.7333
bpref                   Q05     0.8667
recip_rank              Q05     1.0000
iprec_at_recall_0.00    Q05     1.0000
iprec_at_recall_0.10    Q05     1.0000
iprec_at_recall_0.20    Q05     1.0000
iprec_at_recall_0.30    Q05     1.0000
iprec_at_recall_0.40    Q05     1.0000
iprec_at_recall_0.50    Q05     0.8889
iprec_at_recall_0.60    Q05     0.8333
iprec_at_recall_0.70    Q05     0.7333
iprec_at_recall_0.80    Q05     0.7222
iprec_at_recall_0.90    Q05     0.0000
iprec_at_recall_1.00    Q05     0.0000
P_5                     Q05     1.0000
P_10                    Q05     0.8000
P_15                    Q05     0.7333
```

```
P_20                    Q05     0.6500
P_30                    Q05     0.4333
P_100                   Q05     0.1300
P_200                   Q05     0.0650
P_500                   Q05     0.0260
P_1000                  Q05     0.0130
num_ret                 Q06     20
num_rel                 Q06     18
num_rel_ret             Q06     17
map                     Q06     0.9124
Rprec                   Q06     0.8889
bpref                   Q06     0.9444
recip_rank              Q06     1.0000
iprec_at_recall_0.00    Q06     1.0000
iprec_at_recall_0.10    Q06     1.0000
iprec_at_recall_0.20    Q06     1.0000
iprec_at_recall_0.30    Q06     1.0000
iprec_at_recall_0.40    Q06     1.0000
iprec_at_recall_0.50    Q06     1.0000
iprec_at_recall_0.60    Q06     0.9375
iprec_at_recall_0.70    Q06     0.9375
iprec_at_recall_0.80    Q06     0.9375
iprec_at_recall_0.90    Q06     0.8947
iprec_at_recall_1.00    Q06     0.0000
P_5                     Q06     1.0000
P_10                    Q06     1.0000
P_15                    Q06     0.9333
P_20                    Q06     0.8500
P_30                    Q06     0.5667
P_100                   Q06     0.1700
P_200                   Q06     0.0850
P_500                   Q06     0.0340
P_1000                  Q06     0.0170
num_ret                 Q07     20
num_rel                 Q07     15
num_rel_ret             Q07     12
map                     Q07     0.6464
Rprec                   Q07     0.8000
bpref                   Q07     0.8000
recip_rank              Q07     1.0000
iprec_at_recall_0.00    Q07     1.0000
iprec_at_recall_0.10    Q07     1.0000
iprec_at_recall_0.20    Q07     0.8333
```

iprec_at_recall_0.30 Q07 0.8333
iprec_at_recall_0.40 Q07 0.8333
iprec_at_recall_0.50 Q07 0.8333
iprec_at_recall_0.60 Q07 0.8333
iprec_at_recall_0.70 Q07 0.8000
iprec_at_recall_0.80 Q07 0.8000
iprec_at_recall_0.90 Q07 0.0000
iprec_at_recall_1.00 Q07 0.0000
P_5 Q07 0.6000
P_10 Q07 0.8000
P_15 Q07 0.8000
P_20 Q07 0.6000
P_30 Q07 0.4000
P_100 Q07 0.1200
P_200 Q07 0.0600
P_500 Q07 0.0240
P_1000 Q07 0.0120
num_ret Q08 20
num_rel Q08 13
num_rel_ret Q08 11
map Q08 0.6746
Rprec Q08 0.6923
bpref Q08 0.8462
recip_rank Q08 1.0000
iprec_at_recall_0.00 Q08 1.0000
iprec_at_recall_0.10 Q08 1.0000
iprec_at_recall_0.20 Q08 1.0000
iprec_at_recall_0.30 Q08 1.0000
iprec_at_recall_0.40 Q08 0.6923
iprec_at_recall_0.50 Q08 0.6923
iprec_at_recall_0.60 Q08 0.6923
iprec_at_recall_0.70 Q08 0.6923
iprec_at_recall_0.80 Q08 0.5500
iprec_at_recall_0.90 Q08 0.0000
iprec_at_recall_1.00 Q08 0.0000
P_5 Q08 1.0000
P_10 Q08 0.6000
P_15 Q08 0.6000
P_20 Q08 0.5500
P_30 Q08 0.3667
P_100 Q08 0.1100
P_200 Q08 0.0550
P_500 Q08 0.0220

P_1000 Q08 0.0110
num_ret Q09 20
num_rel Q09 20
num_rel_ret Q09 17
map Q09 0.8253
Rprec Q09 0.8500
bpref Q09 0.8500
recip_rank Q09 1.0000
iprec_at_recall_0.00 Q09 1.0000
iprec_at_recall_0.10 Q09 1.0000
iprec_at_recall_0.20 Q09 1.0000
iprec_at_recall_0.30 Q09 1.0000
iprec_at_recall_0.40 Q09 1.0000
iprec_at_recall_0.50 Q09 1.0000
iprec_at_recall_0.60 Q09 0.9375
iprec_at_recall_0.70 Q09 0.9375
iprec_at_recall_0.80 Q09 0.8947
iprec_at_recall_0.90 Q09 0.0000
iprec_at_recall_1.00 Q09 0.0000
P_5 Q09 1.0000
P_10 Q09 1.0000
P_15 Q09 0.9333
P_20 Q09 0.8500
P_30 Q09 0.5667
P_100 Q09 0.1700
P_200 Q09 0.0850
P_500 Q09 0.0340
P_1000 Q09 0.0170
num_ret Q10 20
num_rel Q10 9
num_rel_ret Q10 8
map Q10 0.6566
Rprec Q10 0.5556
bpref Q10 0.8889
recip_rank Q10 1.0000
iprec_at_recall_0.00 Q10 1.0000
iprec_at_recall_0.10 Q10 1.0000
iprec_at_recall_0.20 Q10 1.0000
iprec_at_recall_0.30 Q10 1.0000
iprec_at_recall_0.40 Q10 0.7143
iprec_at_recall_0.50 Q10 0.7143
iprec_at_recall_0.60 Q10 0.5455
iprec_at_recall_0.70 Q10 0.5385

iprec_at_recall_0.80 Q10 0.4444
iprec_at_recall_0.90 Q10 0.0000
iprec_at_recall_1.00 Q10 0.0000
P_5 Q10 0.6000
P_10 Q10 0.5000
P_15 Q10 0.4667
P_20 Q10 0.4000
P_30 Q10 0.2667
P_100 Q10 0.0800
P_200 Q10 0.0400
P_500 Q10 0.0160
P_1000 Q10 0.0080
runid all myIRmethod
num_q all 10
num_ret all 200
num_rel all 142
num_rel_ret all 122
map all 0.6906
gm_map all 0.6819
Rprec all 0.7016
bpref all 0.8584
recip_rank all 0.9500
iprec_at_recall_0.00 all 0.9857
iprec_at_recall_0.10 all 0.9857
iprec_at_recall_0.20 all 0.9205
iprec_at_recall_0.30 all 0.8806
iprec_at_recall_0.40 all 0.8212
iprec_at_recall_0.50 all 0.7878
iprec_at_recall_0.60 all 0.7508
iprec_at_recall_0.70 all 0.7308
iprec_at_recall_0.80 all 0.6276
iprec_at_recall_0.90 all 0.1659
iprec_at_recall_1.00 all 0.0765
P_5 all 0.7600
P_10 all 0.7200
P_15 all 0.7000
P_20 all 0.6100
P_30 all 0.4067
P_100 all 0.1220
P_200 all 0.0610
P_500 all 0.0244
P_1000 all 0.0122