\* Σημείωση προς τον καθηγητή: Προσπάθησα να αναφέρω στο κείμενο όλα τα ζητούμενα της εργασίας 1, αλλά δεν κατάφερα να καλύψω το όριο των τεσσάρων σελίδων. Παρόλα αυτά πιστεύω πως έχω περιγράψει επαρκώς τη διαδικασία της δημιουργίας των ευρετηρίων και τη περιγραφή των τροποποιήσεών τους. Ελπίζω λοιπόν το μειωμένο πλήθος των σελίδων να μην είναι πρόβλημα.

Περίοδος: Χειμερινό Εξάμηνο 2020 - 2021

Μάθημα: Ανάκτηση Πληροφοριών - Μηχανές Αναζήτησης

Συντάκτης: Μάρτσης Βασίλειος

ΑΜ: 154489

Θέμα:

Μηχανή Αναζήτησης Terrier (Εργασία 1)

# περιγραφη του περιβαλλοντος

## Η μηχανή αναζήτησης Terrier

Για την υλοποίηση της εργασίας γίνεται χρήση της μηχανής αναζήτησης Terrier (Version Number: 5.2). Το Terrier είναι μία πολύ ευέλικτη, αποτελεσματική και αποδοτική μηχανή αναζήτησης, Ανοιχτού κώδικα, που αναπτύσσεται γρήγορα για μεγάλες συλλογές εγγράφων. Υλοποιεί προηγμένες λειτουργίες ευρετηρίασης και ανάκτησης δεδομένων, και αποτελεί ιδανική πλατφόρμα για γρήγορη ανάπτυξη και αξιολόγηση αιτημάτων ανάκτησης μεγάλης κλίμακας. Είναι λογισμικό ανοιχτού κώδικα (ΕΛ/ΛΑΚ) και είναι μια περιεκτική και ευέλικτη πλατφόρμα για έρευνα και πειραματισμό στην ανάκτηση κειμένου. Η έρευνα διεξάγεται εύκολα σε standard TREC (Text Retrieval Conference) και CLEF συλλογές δοκιμής. Είναι υλοποιημένο σε Java, και έχει αναπτυχθεί στο School of Computing Science, University of Glasgow.

# προετοιμασία για τη δημιουργία των ευρετηρίων (index)

Μετά την εγκατάσταση της μηχανής στο σύστημα, μέσο του command line των windows, που είναι και το λειτουργικό σύστημα που χρησιμοποιείται για την υλοποίηση της εργασίας, γίνεται χρήση της εντολής “trec\_setup.bat” η οποία παίρνει ως όρισμα τη διαδρομή του φακέλου προορισμού των αρχείων από τα οποία θέλουμε να δημιουργήσουμε το ευρετήριο. Η εντολή αυτή θα δημιουργήσει ορισμένα αρχεία απαραίτητα για προχωρήσουμε στην κατασκευή του ευρετηρίου. Τα αρχεία που θα χρησιμοποιηθούν για την υλοποίηση της εργασίας αυτής είναι μία συλλογή 53.006 εγγράφων (documents), που ακολουθούν τη μορφή TREC (trec format). Για τον σκοπό αυτής της εργασίας θα δημιουργηθούν βασισμένα στη συλλογή εγγράφων αυτή, πέντε ξεχωριστά ευρετήρια.

# δημιουργία ευρετηρίων

Για τη δημιουργία των 5 ευρετηρίων, πραγματοποιήθηκαν για το κάθε ένα από αυτά, διαφορετικές παραμετροποιήσεις στα αρχεία του Terrier. Αρχικά θα χρησιμοποιηθεί η εντολή “terrier bi’’ (Batch Indexing) η οποία θα κατασκευάσει ένα ευρετήριο βασισμένο στα πρώτα δέκα χιλιάδες αρχεία από τη συλλογή που προαναφέρθηκε.

Ακολουθούν η περιγραφή των 5 ευρετηρίων, οι παραμετροποιήσεις που πραγματοποιήθηκαν πάνω σε αυτά και τα αποτελέσματά τους:

## Δημιουργία του πρώτου ευρετηρίου

Για τη δημιουργία του πρώτου ευρετηρίου χρησιμοποιήθηκαν οι προκαθορισμένες ρυθμίσεις του Terrier οι ρυθμίσεις αυτές είναι οι εξής:

### Αρχείο terrier.properites

#### TrecDocTags.doctag=DOC

Η ετικέτα που είναι ορισμένη ως η κύρια ετικέτα των αρχείων.

#### TrecDocTags.idtag=DOCNO

Η ετικέτα που περιέχει το νούμερο το οποίο αντιστοιχεί στο κάθε αρχείο.

#### TrecDocTags.skip=DOCHDR

Οι ετικέτες που θα παραλειφθούν κατά τη διαδικασία της ευρετηρίασης.

#### TrecDocTags.casesensitive=false

false σημαίνει ότι δεν παίζει ρόλο το αν στις ετικέτες χρησιμοποιούνται κεφαλαία ή πεζά γράμματα.

#### TrecQueryTags.doctag=TOP

Το όνομα της ετικέτας που ορίζει την αρχή του TREC αρχείου.

#### TrecQueryTags.idtag=NUM

Το όνομα της ετικέτας που σημαδεύει το id του TREC αρχείο.

#### TrecQueryTags.process=TOP,NUM,TITLE

Οι ετικέτες μέσα στο TREC αρχείο που θα υποστούν επεξεργασία.

#### TrecQueryTags.skip=DESC,NARR

Οι ετικέτες μέσα στο TREC αρχείο που θα παραληφθούνε κατά τη διάρκεια των ερωτημάτων.

Χρησιμοποιώντας τις παραπάνω default ρυθμίσεις, το Terrier θα επεξεργαστεί τους τίτλους των TREC αρχείων, καθώς θα παραλείψει τα descriptions και τα narratives.

#### termpipelines=Stopwords,PorterStemmer

Χρησιμοποιώντας τη ρύθμιση αυτή το σύστημα αγνοεί τις λέξεις των ερωτημάτων που περιέχονται στο αρχείο Stopwords και μετά την ανάγει σε μία ποιο βασική της μορφή χρησιμοποιώντας τον προκαθορισμένο Stemmer, ο οποίος ακολουθεί τον αλγόριθμο αναγωγής λέξεων της αγγλικής γλώσσας του Porter (PorterStemmer).

### Χρόνος εκτέλεσης της εντολής «terrier bi»

Ο χρόνος που χρειάστηκε για να δημιουργηθεί το ευρετήριο στο σύστημα που πραγματοποιείται η εργασία 359 δευτερόλεπτα.

Για τη δημιουργία των υπόλοιπων ευρετηρίων θα αναφερθούν μόνο οι ρυθμίσεις που τροποποιήθηκαν έτσι ώστε να διαφέρουν από τις προκαθορισμένες.

## Δημιουργία του δεύτερου ευρετηρίου

### Αρχείο terrier.properites

#### TrecDocTags.skip=DOCHDR,CLASSIFICATIONS

Επειδή η ετικέτα CLASSIFICATIONS περιέχει μόνο αριθμούς των διαφόρων πατεντών, πιθανών η παράληψή της να ελαφρύνει το σύστημα, επηρεάζοντάς το ελάχιστα.

#### terrier.index.path=C:/Terrier/terrier-project-5.2/var/index2

Η διαδρομή του φακέλου στον οποίο θα αποθηκευτεί το ευρετήριο. Ο λόγος χρήσης αυτής της παραμέτρου είναι η δημιουργία 5 διαφορετικών ευρετηρίων, έτσι ώστε να μην χρειαστεί να διαγραφεί το προηγούμενο προκειμένου να δημιουργηθεί το καινούριο. Για να χρησιμοποιηθεί αυτή η μεταβλητή, θα πρέπει να τεθεί ως όρισμα στην εντολή “terrier bi” (π.χ. terrier bi terrier.index.path).

### Χρόνος εκτέλεσης της εντολής «terrier bi»

Ο χρόνος που χρειάστηκε για να δημιουργηθεί το ευρετήριο στο σύστημα που πραγματοποιείται η εργασία 81 δευτερόλεπτα.

## Δημιουργία του τρίτου ευρετηρίου

### Αρχείο terrier.properites

#### TrecDocTags.skip=DOCHDR,CLASSIFICATIONS

#### stopwords.filename=stopword-list2.txt

Η διαδρομή του φακέλου στον οποίο βρίσκεται το αρχείο με τα stopwords (Λέξεις που θα παραληφθούν κατά τη διάρκεια των ερωτημάτων). Η διαδρομή αυτή που εισήχθη είναι ένα προσαρμοσμένο το οποίο περιέχει ως stopword μόνο τη λέξη «the».

#### termpipelines=WeakPorterStemmer

#### terrier.index.path=C:/Terrier/terrier-project-5.2/var/index3

Θα χρησιμοποιηθεί μία ποιο αδύναμη έκδοση του PorterStemmer

### Χρόνος εκτέλεσης της εντολής «terrier bi»

Ο χρόνος που χρειάστηκε για να δημιουργηθεί το ευρετήριο στο σύστημα που πραγματοποιείται η εργασία 359 δευτερόλεπτα.

## Δημιουργία του τέταρτου ευρετηρίου

### Αρχείο terrier.properites

#### TrecDocTags.skip=DOCHDR,CLASSIFICATIONS

#### Η μεταβλητή «TrecQueryTags.skip» αφαιρέθηκε ώστε να μην παραληφθεί καμία ετικέτα

#### Η μεταβλητή «TrecQueryTags.process» αφερέθηκε ώστε να επεξεργαστούν όλες οι ετικέτες

#### Η μεταβλητή «TrecQueryTags.skip» αφαιρέθηκε ώστε να μην παραληφθεί καμία ετικέτα

#### Η μεταβλητή «termpipelines» αφαιρέθηκε ώστε να μην χρησημοποιηθούν ούτε stopwords ούτε stemmer.

#### terrier.index.path=C:/Terrier/terrier-project-5.2/var/index4

### Χρόνος εκτέλεσης της εντολής «terrier bi»

Ο χρόνος που χρειάστηκε για να δημιουργηθεί το ευρετήριο στο σύστημα που πραγματοποιείται η εργασία 359 δευτερόλεπτα.

## Δημιουργία του πέμπτου ευρετηρίου

### Αρχείο terrier.properites

#### FieldTags.process=TITLE

Οι ετικέτες που θα συμπεριληφθούν στην ευρετηρίαση.

### Χρόνος εκτέλεσης της εντολής «terrier bi»

Ο χρόνος που χρειάστηκε για να δημιουργηθεί το ευρετήριο στο σύστημα που πραγματοποιείται η εργασία 81 δευτερόλεπτα.