Εργασία 1η

Υπέυθυνος διδάσκων, ευσταθιοσ σταματατοσ | Πανεπιστήμιο Αιγαίου

Ανάκτηση πληροφοριας

νικολαοσ μπουσιοσ

2015

# Εισαγωγή

Το συγκεκριμένο σύστημα ανάκτησης υλοποιήθηκε πάνω σε Java έκδοση 8 και αφορά την δημιουργία και αξιολόγηση διαφορετικών παραλλαγών ενός απλού συστήματος ανάκτησης πληροφορίας lemurcgi. Το σύστημα μπορεί να δέχεται ερωτήματα, τα συγκρίνει με ένα σύνολο κειμένων και επιστρέφει αποτελέσματα σε κατάταξη σχετικότητας. Η πρόσβαση και το crawling της σελίδας γίνεται με την βιβλιοθήκη jsoup (*μια από τις διασημότερες opensource αυτήν την στιγμή για crawling, άλλες είναι το framework που προσφέρει το rapidminer*).

Η εφαρμογή που υλοποιήθηκε κατασκευάστηκε με γνώμονα την ακρίβεια και την ταχύτητα. Ποιο συγκεκριμένα το crawling της βάσης γίνετε σε περίπου **4 λεπτά** (κατέβασμα όλων των συλλογών των λέξεων που αφορούν 800 queries, 8 βάσεων, 2 διαφορετικών παραλλαγών, αυτό που δόθηκε για την εργασία και η παραλλαγή αυτού που αφορά έως 5 λέξεις.

Αφού παραχθούν αυτά τα δεδομένα ξεκινά η διαδικασία την παραγωγής αποτελεσμάτων και εξαγωγής στατιστικών βάση 4 μοντέλων εξαγωγής και 4 μοντέλων αξιολόγησης. Ποιο συγκεκριμένα με χρήση συνδεμένων λιστών καθ’ όλη την διάρκεια εκτέλεσης του προγράμματος γίνεται παραγωγή αποτελεσμάτων σε μόλις **1 λεπτό και 25 περίπου δευτερόλεπτα** (αφορά εκτέλεση σε i7 Mobile 4ης γενιάς). Σημαίνει πρόχειρα ότι ο υπολογισμός του κάθε query σε μόλις περίπου **200** **millisecond**.

Τα αποτελέσματα εξάγονται σε αρχείο csv ουτοσώστε να επεξεργαστούν σε excel για παραγωγή γραφημάτων.

# Αποτελέσματα

## Ερώτημα 1ο: Ποιο μοντέλο ανάκτησης είναι το πιο αποτελεσματικό;

Με βάση τα αποτελέσματα προκύπτει ότι καλύτερο είναι το TF-IDF καθώς σε όλες τις βάσεις για τις 2 παραλλαγές ερωτημάτων Full και έως 5 λέξεων παρουσιάζει το μεγαλύτερο score και στα 3 μοντέλα αξιολόγησης. Για περισσότερες πληροφορίες συμβουλευτείτε το excel αποτελεσμάτων.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Average of Recall** | **Average of Precision** | **Average of R-Precision** |
| BM25 | 0,248920377 | 0,1123 | 0,11518346 |
| Boolean | 0,025461083 | 0,082652624 | 0,083860912 |
| TF | 0,213005058 | 0,113 | 0,113890425 |
| TF-IDF | 0,368328809 | 0,1701 | 0,174512915 |

## Ερώτημα 2ο: Η χρήση stemming (εύρεση ρίζας λέξεων) βελτιώνει τα αποτελέσματα;

Με βάση τα αποτελέσματα σύγκρισης της βάσης 0 και της 1(stemmed only) προκύπτει ότι το stemming βελτιώνει τα αποτελέσματα καθώς παρουσιάζετε αισθητά μεγαλύτερο recall και ελαφρώς αλλά με προβάδισμα μεγαλύτερο precision και R-Precision. Για περισσότερες πληροφορίες συμβουλευτείτε το excel αποτελεσμάτων.

## Ερώτημα 3ο: Η χρήση stopword removal (απομάκρυνση πολύ συχνών λέξεων) βελτιώνει τα αποτελέσματα;

Βάση των αποτελεσμάτων σύγκρισης της βάσης 0 και 1, 1 και 2 προκύπτει ότι η απομάκρυνση των stopword βελτιώνει τα αποτελέσματα αλλά σε ελαφρώς μικρότερο βαθμό σε σχέση με τη χρήση stemming. Για περισσότερες πληροφορίες συμβουλευτείτε το excel αποτελεσμάτων.

## Ερώτημα 4ο: Με ποια από τις 2 παραλλαγές των ερωτημάτων προκύπτουν τα καλύτερα αποτελέσματα;

Βάση των αποτελεσμάτων φαίνεται ξεκάθαρα ότι στην δεύτερη (δική μου) παραλλαγή των ερωτημάτων προκύπτουν καλύτερα αποτελέσματα. Φυσικά αυτό εξαρτάται και από την φύση των ερωτημάτων καθώς στην παραλλαγή με τις έως 5 λέξεις επιλέχθηκαν λέξεις «κλειδιά».

# Οδηγίες Εκτέλεσης

Το πρόγραμμα απαιτεί προ εγκατεστημένη έκδοση της Java, JVM 8+ για να τρέξει. Είναι console based εφαρμογή και έχει δυο τρόπους εκτέλεσης ο 1ος αφορά την εκτέλεση και του crawling και της ανάκτησης. Ενώ η δεύτερη μόνο της ανάκτησης

Για την εκτέλεση του προγράμματος γράφουμε

java –jar InfoRetrieval.jar

Και αν έχουμε κατεβασμένες ήδη τις βάσεις στο ίδιο path τότε μπορούμε να τρέξουμε

java –jar InfoRetrieval.jar models

Επίσης τοπικά θα δημιουργηθεί και αρχείο **log.txt.**

*Διαφορετικά για την εκτέλεση του σε περιβάλλον μόνο* ***windows*** *μπορείτε να τρέχετε το* ***executable.bat*** *που υπάρχει εσωτερικά!*

**Για οδηγίες εκτέλεσης ή άλλες πληροφορίες επικοινωνήστε μαζί μου στο icsd10124@icsd.aegean.gr**