Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
HY463 – Συστήματα Ανάκτησης Πληροφοριών
Άνοιξη 2022
Διδάσκων: Γιάννης Τζίτζικας
Υπεύθυνοι Βοηθοί: Όλοι

Β. Το Θέμα Έργου 2022

Στόχος. Καλείστε να εφαρμόσετε και να δοκιμάσετε το σύστημα που θα φτιάξετε σε μια πραγματική συλλογή βιοϊατρικών άρθρων και να αξιολογήσετε την αποτελεσματικότητα του συστήματός σας, ακολουθώντας την επιστημονικά καθιερωμένη μεθοδολογία.

Συλλογή Εγγράφων. Η συλλογή που θα χρησιμοποιήσετε περιέχει άρθρα σχετικά με την βιοϊατρική τα οποία έχουν εξαχθεί από τη ψηφιακή βάση δεδομένων «PubMed Central» (PMC). Κάθε άρθρο της συλλογής αναπαριστάται ως ένα NXML αρχείο (XML αρχείο κωδικοποιημένο με χρήση της βιβλιοθήκης «NLM Journal Archiving and Interchange Tag Library» και αναγνωρίζεται μοναδικά από τον αριθμό PMCID (το όνομα κάθε αρχείου είναι στην ουσία ο αριθμός PMCID του αντίστοιχου άρθρου).

Κάθε NXML αρχείο περιλαμβάνει πολλές ετικέτες όπως: αναγνωριστικά του άρθρου, αναγνωριστικό περιοδικού, τίτλος περιοδικού, εκδότης περιοδικού, τίτλος άρθρου, περίληψη άρθρου, συγγραφείς, κυρίως σώμα άρθρου, ημερομηνία δημοσίευσης, αναφορές, και πολλά άλλα.

Στα πλαίσια αυτής της εργασίας μας ενδιαφέρουν οι παρακάτω 8 ετικέτες:

- *Αναγνωριστικό PMCID*
- *Τίτλος άρθρου*
- *Συγγραφείς του άρθρου*
- *Περίληψη άρθρου*
- *Κυρίως σώμα άρθρου*
- *Κατηγορίες άρθρου*
- *Περιοδικό που δημοσιεύτηκε*
- *Εκδότης περιοδικού*

Στο **Παράρτημα Β-1** θα βρείτε κώδικα για την ανάγνωση των παραπάνω ετικετών. Φυσικά, αν το επιθυμείτε, είστε ελεύθεροι να χρησιμοποιήσετε όποιες άλλες ετικέτες θέλετε.

Συλλογή Θεμάτων Αναζήτησης/Αξιολόγησης. Σας δίνεται μια συλλογή από 30 **ιατρικά θέματα** (αρχείο **topics.xml**). Κάθε ιατρικό θέμα αναπαριστά τα πρακτικά μιας επίσκεψης ασθενή σε γιατρό και περιέχει πληροφορίες όπως το ιστορικό του ασθενή, τα συμπτώματά του, τυχόν εξετάσεις που έκανε, κτλ. Τα θέματα έχουν επισημειωθεί σύμφωνα με τους 3 πιο συνηθισμένους **τύπους** κλινικών ερωτήσεων:

- **Διάγνωση** – Ποια είναι η διάγνωση του ασθενή;
- **Εξέταση** – Τι εξετάσεις πρέπει να κάνει ο ασθενής;
- **Θεραπεία** – Ποια αγωγή πρέπει να ακολουθήσει ο ασθενής;

Παραδείγματος χάριν, για ένα θέμα «Διάγνωση», το σύστημά σας πρέπει να ανακτήσει άρθρα που θα φανούν χρήσιμα στον γιατρό για να διαγνώσει την αρρώστια του ασθενή. Αντίστοιχα, για ένα θέμα τύπου «Εξέταση», το σύστημά σας πρέπει να ανακτήσει σχετικά άρθρα που προτείνουν εξετάσεις για την διάγνωση του ασθενή. Τέλος, για ένα θέμα τύπου «Θεραπεία», το σύστημά σας πρέπει να ανακτήσει άρθρα που προτείνουν στον γιατρό την καλύτερη θεραπεία που πρέπει να ακολουθήσει ο ασθενής. Στη φόρμα αναζήτησης ο χρήστης θα επιλέγει τι θέλει να ψάξει (Διάγνωση, Εξέταση ή Θεραπεία).
Για κάθε θέμα, δίνεται μια ***περιγραφή*** και μια (μικρότερη) ***σύνοψη***. Για τη δημιουργία της αντίστοιχης επερώτησης μπορείτε να χρησιμοποιήσετε είτε τη περιγραφή, είτε τη σύνοψη, όμως **όχι** και τα 2 μαζί ταυτόχρονα (αυτός ήταν κανόνας του διαγωνισμού TREC[1] το 2015). Επίσης, μπορείτε να **επεξεργαστείτε** τα θέματα όπως εσείς νομίζετε (ώστε να αποδίδει καλύτερα το σύστημά σας), **είτε με αυτόματο τρόπο είτε με παρέμβαση χρήστη**. Σε κάθε περίπτωση **πρέπει να αναφέρετε** αν ο κάθε τρόπος που προτείνετε χρειάζεται παρέμβαση χρήστη ή είναι πλήρως αυτόματος.

[1] http://trec.nist.gov/pubs/call2016.html, http://trec-cds.appspot.com/

Σας ζητείται να **επεκτείνεται το πρόγραμμά σας** (το διανυσματικό μοντέλο, τον επεξεργαστή επερωτήσεων, κτλ.) ώστε να ζητάει από το χρήστη να δώσει τα στοιχεία ενός ιατρικού θέματος (τον **τύπο** του και την **περιγραφή ή τη σύνοψη** του) και το σύστημα εν συνεχεία να ανακτά τα **σχετικά άρθρα**. Για κάθε στοιχείο της απάντησης πρέπει να επιστρέφεται το **μονοπάτι του αρχείου** (path) και ο **βαθμός ομοιότητας** (score) που υπολόγισε το σύστημά σας. Στην αναφορά σας, εκτός των άλλων, πρέπει **να αναφέρετε** τις αλλαγές που κάνατε στο σύστημά σας (για τις ανάγκες αυτής της εφαρμογής) καθώς και τον λόγο που τις κάνατε.

Συλλογή Αξιολόγησης

Η συλλογή αξιολόγησης που θα χρησιμοποιήσετε προέρχεται από το TREC 2014 (θυμηθείτε αυτά που λέγαμε στην ενότητα Αξιολόγηση Αποτελεσματικότητας Ανάκτησης), και όποιος θέλει μπορεί (δεν είναι υποχρεωτικό) να ενημερωθεί για το σκεπτικό αυτού του Clinical Decision Support Track Track του TREC, διαβάζοντας το άρθρο

> Matthew S. Simpson, Ellen M. Voorhees, and William Hersh, Overview of the TREC 2014 Clinical Decision Support Track, URL: https://pdfs.semanticscholar.org/fcf8/1b7641c0cd7be089051018a53fabfa685da0.pdf)

(Επίσης στην διεύθυνση http://trec.nist.gov/pubs/trec23/trec2014.html υπάρχουν papers από ομάδες οι οποίες συμμετείχαν στο διαγωνισμό εκείνης της χρονιάς).

Για την αξιολόγηση της αποτελεσματικότητας ανάκτησης (retrieval effectiveness) του συστήματός σας θα χρησιμοποιήσετε τα συνολικά 30 ιατρικά θέματα του αρχείου **topics.xml** (10 από κάθε τύπο κλινικής ερώτησης). Κώδικας για την ανάγνωσή τους δίνεται στο **Παράρτημα Β-2.**

Επίσης για κάθε ιατρικό θέμα στο αρχείο **topics.xml**, σας δίνεται:

- ένα σύνολο εγγράφων της συλλογής που είναι πολύ σχετικά για την απάντηση του αντίστοιχου θέματος
- ένα σύνολο εγγράφων της συλλογής που είναι σχετικά για την απάντηση του αντίστοιχου θέματος
- ένα σύνολο εγγράφων της συλλογής που δεν είναι σχετικά για την απάντηση του αντίστοιχου θέματος

Οι παραπάνω πληροφορίες δίνονται στο TSV (Tab-Separated Values) αρχείο **qrels.txt**. Κάθε γραμμή αυτού του αρχείου περιέχει 4 στοιχεία:

1. *topic number:* αριθμός από 1 έως 30 που αναπαριστά το αντίστοιχο ιατρικό θέμα του αρχείου **topics.xml**)
2. *αριθμός 0* (δεν χρησιμοποιείται)
3. *document PMCID*: αναγνωριστικό PMC βιοϊατρικού άρθρου από τη συλλογή **Medical Collection**
4. *relevance score*: η σχετικότητα του βιοϊατρικού άρθρου για την απάντηση του ιατρικού θέματος (0 = μη σχετικό, 1 = σχετικό, 2 = πολύ σχετικό)

Για παράδειγμα, η γραμμή «**1 0 1033658 0**» σημαίνει ότι το έγγραφο της συλλογής με PMCID "1033658 " δεν είναι σχετικό για το ιατρικό θέμα με αριθμό 1.

Διαδικασία Αυτόματης Αξιολόγησης

Στη συγκεκριμένη φάση καλείστε να αξιολογήσετε την **αποτελεσματικότητα** του συστήματός σας.

Συγκεκριμένα **δημιουργήστε** ένα πρόγραμμα (μπορείτε να το πείτε **IRQualityEvaluator**) που θα αυτοματοποιεί τον υπολογισμό των μέτρων αξιολόγησης. Το πρόγραμμα πρέπει να μπορεί να διαβάζει το αρχείο των θεμάτων (topics), να στέλνει επερωτήσεις στο σύστημα που φτιάξατε στη Φάση Α, και κατόπιν να αποθηκεύει τα 1000 κορυφαία αποτελέσματα (όπως αυτά επιστρέφονται από το σύστημά σας) σε ένα αρχείο με όνομα «**results.txt**», στη μορφή:

```
TOPIC_NO Q0 PMCID RANK SCORE RUN_NAME
```

όπου **TOPIC_NO** είναι ο αριθμός του ιατρικού θέματος, **Q0** μια σταθερά (βάλτε τον αριθμό 0), **PMCID** είναι το PMC ID του εγγράφου που ανακτήθηκε, **RANK** είναι η κατάταξη (αριθμός από 1 έως 1000) του εγγράφου που ανακτήθηκε (το έγγραφο με RANK 1 είναι αυτό με τον υψηλότερο βαθμό ομοιότητας, κ.ο.κ), **SCORE** είναι ο βαθμός ομοιότητας που έδωσε το σύστημά σας στο συγκεκριμένο έγγραφο, και **RUN_NAME** είναι ένα αναγνωριστικό για το σύστημά σας (αυτό είναι χρήσιμο σε περίπτωση που θέλετε να δοκιμάσετε διαφορετικές παραμετροποιήσεις του συστήματός σας). Το αρχείο πρέπει να είναι **ταξινομημένο** ως προς το **RANK** (δηλαδή στην πρώτη γραμμή θα είναι το έγγραφο με RANK 1, στην δεύτερη το έγγραφο με RANK 2, κ.ο.κ.).

Εν συνεχεία το πρόγραμμά σας πρέπει να διαβάζει το αρχείο με τα μερικά αποτελέσματα συνάφειας (**qrels.txt**) και συγκρίνοντας το με τα αποτελέσματα του συστήματος σας (στο results.txt), να υπολογίζει τις τιμές των μέτρων που περιγράφονται παρακάτω **για κάθε ιατρικό θέμα**, και θα τις αποθηκεύει σε ένα TSV αρχείο με όνομα «**eval_results.txt**»στην μορφή:

```
TOPIC_NO BPREF_VALUE AVEP_VALUE NDCG_VALUE
```

Επειδή για πολλά από τα έγγραφα τις συλλογής δεν γνωρίζουμε αν είναι σχετικά ή όχι για την απάντηση ενός ή περισσοτέρων θεμάτων, πρέπει να χρησιμοποιήσετε κάποιες μετρικές αξιολόγησης που είναι σχεδιασμένες για τέτοιες περιπτώσεις. Οι μετρικές που πρέπει να χρησιμοποιήσετε είναι οι παρακάτω:

- bpref
- AveP'
- NDCG']

Λίγα λόγια για τις μετρικές

- bpref:
  - Την είχαμε δει στο μάθημα (ενότητα Αξιολόγηση)
  - Περιγράφεται και στο https://trec.nist.gov/pubs/trec16/appendices/measures.pdf καθώς κα στα άρθρα [1] και [2]
- AveP' and NDCG' είναι εκδοχές του Average Precision και NDCG για συλλογές αξιολόγησης που δεν είναι πλήρεις. Στην ουσία κατά τον υπολογισμό τους αφαιρούμε από τις απαντήσεις του συστήματος εκείνα τα έγγραφα που δεν είναι judged. Αναλυτική περιγραφή (για όποιον ενδιαφέρεται) υπάρχει στο άρθρο [2]. Όποιος υπολογίσει και αυτές τις δύο έχει **5μ bonus**.

Τα δύο σχετικά άρθρα:

- [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.

Χρησιμοποιώντας τις παραπάνω μετρικές, θα αξιολογήσετε την αποτελεσματικότητα του συστήματός σας. Παρατηρώντας τα αποτελέσματα της αξιολόγησης αποτελεσματικότητας μπορείτε να κάνετε ό,τι παραλλαγή νομίζετε στο σύστημά σας (π.χ. στη βάρυνση του ευρετηρίου, στον επεξεργαστή επερωτήσεων, στη συνάρτηση υπολογισμού του βαθμού συνάφειας, κλπ.) **ώστε να μεγιστοποιήσετε την αποτελεσματικότητα του συστήματός σας (που είναι το τελικό ζητούμενο) και να αυξήσετε την πιθανότητα να βγείτε νικητές!**

Καλή εργασία!

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΡΤΗΜΑ Β-1 – ΑΝΑΓΝΩΣΗ ΒΙΟΪΑΤΡΙΚΩΝ ΑΡΘΡΩΝ

Για την ανάγνωση ενός βιοϊατρικού άρθρου μπορείτε να χρησιμοποιήσετε την βιβλιοθήκη «**BioReader.jar**». Αφού προσθέσετε την βιβλιοθήκη στο πρόγραμμά σας, με τον παρακάτω κώδικα μπορείτε να διαβάσετε τις ετικέτες ενός άρθρου:

```
import gr.uoc.csd.hy463.NXMLFileReader;
import java.io.File;
import java.io.IOException;
import java.io.UnsupportedEncodingException;
import java.util.ArrayList;
import java.util.HashSet;

public class MYEXAMPLE {

    public static void main(String[] args) throws UnsupportedEncodingException, IOException {

        File example = new File("C:\\dataset\\clinic\\3536594.nxml");

        NXMLFileReader xmlFile = new NXMLFileReader(example);
        String pmcid = xmlFile.getPMCID();
        String title = xmlFile.getTitle();
        String abstr = xmlFile.getAbstr();
        String body = xmlFile.getBody();
        String journal = xmlFile.getJournal();
        String publisher = xmlFile.getPublisher();
        ArrayList<String> authors = xmlFile.getAuthors();
        HashSet<String> categories =xmlFile.getCategories();

        System.out.println("- PMC ID: " + pmcid);
        System.out.println("- Title: " + title);
        System.out.println("- Abstract: " + abstr);
        System.out.println("- Body: " + body);
        System.out.println("- Journal: " + journal);
        System.out.println("- Publisher: " + publisher);
        System.out.println("- Authors: " + authors);
        System.out.println("- Categories: " + categories);
    }
}
```

ΠΑΡΑΡΤΗΜΑ Β2 – ΑΝΑΓΝΩΣΗ ΙΑΤΡΙΚΩΝ Θεμάτων

Για την ανάγνωση ενός του αρχείου με τις ιατρικές αναφορές μπορείτε να χρησιμοποιήσετε την βιβλιοθήκη «**BioReader.jar**». Αφού προσθέσετε την βιβλιοθήκη στο πρόγραμμά σας, με τον παρακάτω κώδικα μπορείτε να διαβάσετε τις όλες τις ιατρικές αναφορές:

```
import gr.uoc.csd.hy463.Topic;
import gr.uoc.csd.hy463.TopicsReader;
import java.util.ArrayList;

public class MYEXAMPLE {
    public static void main(String[] args) throws Exception {
        ArrayList<Topic> topics = TopicsReader.readTopics("C:\\dataset\\topics.xml");
        for (Topic topic : topics) {
            System.out.println(topic.getNumber());
            System.out.println(topic.getType());
            System.out.println(topic.getSummary());
            System.out.println(topic.getDescription());
            System.out.println("---------");
        }

    }
}
```