Συστήματα Ανάκτησης Πληροφορίας
2η σειρά ασκήσεων

Φοιτητής:
Ιωάννης Κανακαράκης p3060190

Άσκηση 1

→ *Μέρος 1*

- `TF → Term Frequency`
  Αριθμός εμφάνισης του όρου στο κείμενο.
- `DF → Document Frequency`
  Αριθμός κειμένων που εμφανίζεται ο όρος.
- `IDF → Inverted Document Frequency =` $\log_{10}(\frac{Documents}{DF})$
- `TF-IDF →` $TF \times IDF$
- Θεωρούμε `StopWords` το σύνολο `{ το }`
- Εφαρμόζουμε `Stemming` *αφαιρώντας τις καταλήξεις* και *αγνοώντας τους τόνους*

Σύμφωνα με τα παραπάνω έχουμε:

Term	#Document	TF	DF	IDF	TF-IDF
Δρόμ-οι	1	1	4	0	0
δρόμ-ο	2	2	4	0	0
δρόμ-ο	3	1	4	0	0
δρόμ-ο	4	1	4	0	0
μεγάλ-ο	4	1	1	0.60206	0.60206
Πάρ-ε	3	1	1	0.60206	0.60206
Πήγαιν-ε	2	1	1	0.60206	0.60206
Πήρ-ες	4	1	1	0.60206	0.60206
παλι-οί	1	1	1	0.60206	0.60206

Η διανυσματική παράσταση είναι:

#Document	δρόμ-ο	μεγάλ-ο	Πάρ-ε	Πήγαιν-ε	Πήρ-ες	παλι-οί
1	0	0	0	0	0	0.60206
2	0	0	0	0.60206	0	0
3	0	0	0.60206	0	0	0
4	0	0.60206	0	0	0.60206	0

→ *Μέρος 2*

query = «Μεγάλοι δρόμοι της Αθήνας»

#Document	δρόμ-ο	μεγάλ-ο	Πάρ-ε	Πήγαιν-ε	Πήρ-ες	παλι-οί
query	0	0.60206	0	0	0	0

Αξιολόγηση:

#Document	q*d	||d||	q*d/||d||
1	0	0	0
2	0	0	0
3	0	0	0
4	0.36247	0.85114	0.42586

Η λίστα που θα επιστραφεί είναι:

#Document	Text	Score
1	Πήρες το μεγάλο δρόμο	0.42586
2	Πήγαινε, δρόμο-δρόμο	0
3	Πάρε δρόμο	0
4	Δρόμοι παλιοί	0

Το ανεστραμμένο αρχείο/ευρετήριο είναι {documentId, termPosition}:

Term	DF	{#Doc, Pos}				
δρόμ	4	{1, 0}	{2, 1}	{2, 2}	{3 1}	{4, 2}
μεγάλ	1	{4, 1}				
Πάρ	1	{3, 0}				
Πήγαιν	1	{2, 0}				
Πήρ	1	{4, 0}				
παλι	1	{1, 1}				

Άσκηση 2

Για το σκοπό της άσκησης δημιουργήθηκε μια απλή εφαρμογή σε Java, χρησιμοποιώντας την βιβλιοθήκη Lucene.
Η εφαρμογή δέχεται όρισμα τη τοποθεσία ενός xml αρχείου, το οποίο διαβάζει και δημιουργεί ένα ευρετήριο. Στη συνέχεια εκτελεί τις επερωτήσεις πάνω σε αυτό.
Εκτυπώνει τα κείμενα που διάβασε, πληροφορίες για το ευρετήριο, το ευρετήριο, τα διανύσματα των κειμένων και τα αποτελέσματα των επερωτήσεων.
Οι επερωτήσεις έγιναν πάνω στον τίτλο των κειμένων και πάνω στο σώμα – περιεχόμενο των κειμένων σε δύο περάσματα.

Ο κώδικας της εφαρμογής βρίσκεται εδώ.

Παρατήρηση:
Για τον υπολογισμό του αποτελέσματος του TF-IDF διανύσματος υπάρχουν πολλές μέθοδοι. Η βιβλιοθήκη Lucene παρέχει έναν default μηχανισμό υπολογισμού του διανύσματος και είναι αυτός που χρησιμοποιήθηκε στην υλοποίηση.

```
      float idf = Similarity.getDefault().idf(df, docsnum);
     float tfidf = idf * tf;
```

→ Παρακάτω φαίνεται το output που παράγει ο κώδικας

Το output του προγράμματος είναι:

```
------------------------------------------------------------------------
The Documents:
------------------------------------------------------------------------
docID:0
title: Visual retrieval engine
body:Search for images by content is a challenge
docID:1
title: Semantic retrieval
body: Use of semantic retrieval as a technique to retrieve images
docID:2
title: Object Oriented Programming
body: C++ is a object oriented programming language
docID:3
title: Natural Language processing
body:NLP and programming techniques
docID:4
title: Language models
body: Semantic retrieval using the language model
docID:5
title: Multimedia retrieval
body: Combined visual and semantic retrieval of images
docID:6
title: Semantic retrieval of images
body: impact of semantic retrieval to image retrieval
docID:7
title: Java
body: Programming tools with java
------------------------------------------------------------------------
The Index Info:
------------------------------------------------------------------------
Index is current
Index version is 1290390267379
Index is optimized
Index is segmented into 3 file
Index files are [_0.cfs, _0.cfx, segments_2]
Index currently operates on segments_2
Index generation is 2
Index holds 8 documents
Index has modified 0 documents
Index doesn't have any deletions
```

```
----------------------------------------------------------------------
The Index:
----------------------------------------------------------------------
DocId        DF        TF        IDF        TF-IDF        Term
0        3        1        1.693147        1.693147        a
1        3        1        1.693147        1.693147        a
2        3        1        1.693147        1.693147        a
3        2        1        1.980829        1.980829        and
5        2        1        1.980829        1.980829        and
1        1        1        2.386294        2.386294        as
0        1        1        2.386294        2.386294        by
2        1        1        2.386294        2.386294        c
0        1        1        2.386294        2.386294        challenge
5        1        1        2.386294        2.386294        combined
0        1        1        2.386294        2.386294        content
0        1        1        2.386294        2.386294        for
6        1        1        2.386294        2.386294        image
0        3        1        1.693147        1.693147        images
1        3        1        1.693147        1.693147        images
5        3        1        1.693147        1.693147        images
6        1        1        2.386294        2.386294        impact
0        2        1        1.980829        1.980829        is
2        2        1        1.980829        1.980829        is
7        1        1        2.386294        2.386294        java
2        2        1        1.980829        1.980829        language
4        2        1        1.980829        1.980829        language
4        1        1        2.386294        2.386294        model
3        1        1        2.386294        2.386294        nlp
2        1        1        2.386294        2.386294        object
1        3        1        1.693147        1.693147        of
5        3        1        1.693147        1.693147        of
6        3        1        1.693147        1.693147        of
2        1        1        2.386294        2.386294        oriented
2        3        1        1.693147        1.693147        programming
3        3        1        1.693147        1.693147        programming
7        3        1        1.693147        1.693147        programming
1        4        1        1.470004        1.470004        retrieval
4        4        1        1.470004        1.470004        retrieval
5        4        1        1.470004        1.470004        retrieval
6        4        2        1.470004        2.940007        retrieval
1        1        1        2.386294        2.386294        retrieve
0        1        1        2.386294        2.386294        search
1        4        1        1.470004        1.470004        semantic
4        4        1        1.470004        1.470004        semantic
5        4        1        1.470004        1.470004        semantic
6        4        1        1.470004        1.470004        semantic
1        1        1        2.386294        2.386294        technique
3        1        1        2.386294        2.386294        techniques
4        1        1        2.386294        2.386294        the
1        2        1        1.980829        1.980829        to
6        2        1        1.980829        1.980829        to
7        1        1        2.386294        2.386294        tools
1        1        1        2.386294        2.386294        use
```

```
4       1       1       2.386294        2.386294        using
5       1       1       2.386294        2.386294        visual
7       1       1       2.386294        2.386294        with
0       1       1       2.386294        2.386294        engine
6       1       1       2.386294        2.386294        images
7       1       1       2.386294        2.386294        java
3       2       1       1.980829        1.980829        language
4       2       1       1.980829        1.980829        language
4       1       1       2.386294        2.386294        models
5       1       1       2.386294        2.386294        multimedia
3       1       1       2.386294        2.386294        natural
2       1       1       2.386294        2.386294        object
6       1       1       2.386294        2.386294        of
2       1       1       2.386294        2.386294        oriented
3       1       1       2.386294        2.386294        processing
2       1       1       2.386294        2.386294        programming
0       4       1       1.470004        1.470004        retrieval
1       4       1       1.470004        1.470004        retrieval
5       4       1       1.470004        1.470004        retrieval
6       4       1       1.470004        1.470004        retrieval
1       2       1       1.980829        1.980829        semantic
6       2       1       1.980829        1.980829        semantic
0       1       1       2.386294        2.386294        visual
```

```
---------------------------------------------------------------------
The Document Vectors:
---------------------------------------------------------------------
docID        Term: a
0            1.693147
1            1.693147
2            1.693147
docID        Term: and
3            1.980829
5            1.980829
docID        Term: as
1            2.386294
docID        Term: by
0            2.386294
docID        Term: c
2            2.386294
docID        Term: challenge
0            2.386294
docID        Term: combined
5            2.386294
docID        Term: content
0            2.386294
docID        Term: for
0            2.386294
docID        Term: image
6            2.386294
docID        Term: images
0            1.693147
1            1.693147
5            1.693147
docID        Term: impact
6            2.386294
docID        Term: is
0            1.980829
2            1.980829
docID        Term: java
7            2.386294
docID        Term: language
2            1.980829
4            1.980829
docID        Term: model
4            2.386294
docID        Term: nlp
3            2.386294
docID        Term: object
2            2.386294
docID        Term: of
1            1.693147
5            1.693147
6            1.693147
docID        Term: oriented
2            2.386294
```

```
docID        Term: programming
2            1.693147
3            1.693147
7            1.693147
docID        Term: retrieval
1            1.470004
4            1.470004
5            1.470004
6            2.940007
docID        Term: retrieve
1            2.386294
docID        Term: search
0            2.386294
docID        Term: semantic
1            1.470004
4            1.470004
5            1.470004
6            1.470004
docID        Term: technique
1            2.386294
docID        Term: techniques
3            2.386294
docID        Term: the
4            2.386294
docID        Term: to
1            1.980829
6            1.980829
docID        Term: tools
7            2.386294
docID        Term: use
1            2.386294
docID        Term: using
4            2.386294
docID        Term: visual
5            2.386294
docID        Term: with
7            2.386294
docID        Term: engine
0            2.386294
docID        Term: images
6            2.386294
docID        Term: java
7            2.386294
docID        Term: language
3            1.980829
4            1.980829
docID        Term: models
4            2.386294
docID        Term: multimedia
5            2.386294
docID        Term: natural
3            2.386294
```

```
docID        Term: object
2            2.386294
docID        Term: of
6            2.386294
docID        Term: oriented
2            2.386294
docID        Term: processing
3            2.386294
docID        Term: programming
2            2.386294
docID        Term: retrieval
0            1.470004
1            1.470004
5            1.470004
6            1.470004
docID        Term: semantic
1            1.980829
6            1.980829
docID        Term: visual
0            2.386294

----------------------------------------------------------------------
Searching title for: image retrieval engines
----------------------------------------------------------------------
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  0: Visual retrieval engine
        Search for images by content is a challenge
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
----------------------------------------------------------------------
Searching title for: image retrieval
----------------------------------------------------------------------
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  0: Visual retrieval engine
        Search for images by content is a challenge
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
```

```
------------------------------------------------------------------
Searching title for: image retrieval image
------------------------------------------------------------------
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  0: Visual retrieval engine
        Search for images by content is a challenge
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
------------------------------------------------------------------
Searching title for: processing with programming languages processes
------------------------------------------------------------------
  2: Object Oriented Programming
         C++ is a object oriented programming language
  3: Natural Language processing
        NLP and programming techniques
------------------------------------------------------------------
Searching title for: Visual multimedia
------------------------------------------------------------------
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  0: Visual retrieval engine
        Search for images by content is a challenge
------------------------------------------------------------------
Searching title for: java
------------------------------------------------------------------
  7: Java
         Programming tools with java
------------------------------------------------------------------
Searching title for: Visual and semantic multimedia retrieval
------------------------------------------------------------------
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  0: Visual retrieval engine
        Search for images by content is a challenge
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
------------------------------------------------------------------
Searching title for: models
------------------------------------------------------------------
  4: Language models
         Semantic retrieval using the language model
------------------------------------------------------------------
Searching body for: image retrieval engines
------------------------------------------------------------------
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
  4: Language models
         Semantic retrieval using the language model
```

```
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
----------------------------------------------------------------------
Searching body for: image retrieval
----------------------------------------------------------------------
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
  4: Language models
         Semantic retrieval using the language model
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
----------------------------------------------------------------------
Searching body for: image retrieval image
----------------------------------------------------------------------
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
  4: Language models
         Semantic retrieval using the language model
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
----------------------------------------------------------------------
Searching body for: processing with programming languages processes
----------------------------------------------------------------------
  7: Java
         Programming tools with java
  3: Natural Language processing
        NLP and programming techniques
  2: Object Oriented Programming
         C++ is a object oriented programming language
----------------------------------------------------------------------
Searching body for: Visual multimedia
----------------------------------------------------------------------
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
----------------------------------------------------------------------
Searching body for: java
----------------------------------------------------------------------
  7: Java
         Programming tools with java
```

```
------------------------------------------------------------------
Searching body for: Visual and semantic multimedia retrieval
------------------------------------------------------------------
  5: Multimedia retrieval
         Combined visual and semantic retrieval of images
  6: Semantic retrieval of images
         impact of semantic retrieval to image retrieval
  4: Language models
         Semantic retrieval using the language model
  1: Semantic retrieval
         Use of semantic retrieval as a technique to retrieve images
  3: Natural Language processing
        NLP and programming techniques
------------------------------------------------------------------
Searching body for: models
------------------------------------------------------------------
```

Τέλος του output του προγράμματος.

Άσκηση 3

Για το σκοπό της άσκησης δημιουργήθηκε μια νέα γενική βιβλιοθήκη πάνω από τη βιβλιοθήκη Lucene με δομές και αντικείμενα κατάλληλα για την ανάλυση, επεξεργασία, αποθήκευση και εκτύπωση των επιθυμητών στοιχείων. Η βιβλιοθήκη αυτή χρησιμοποιήθηκε για τη δημιουργία ενός εκτελέσιμου που φέρει εις πέρας το σκοπό της άσκησης.
Ο κώδικας της εργασίας βρίσκεται εδώ, μαζί με τα αποτελέσματα και τα πηγαία αρχεία. Ο κώδικας διατίθεται ελεύθερα με την άδεια χρήσης GNU GPLv3.

Όπως φαίνεται και στο output (όπως δίδεται παρακάτω) αυτό που κάνει ο κώδικας του εκτελέσιμου αρχείου είναι:

1. Διαβάζει το αρχείο της συλλογής cacm και το φορτώνει σε κατάλληλη δομή.
2. Γράφει την συλλογή σε xml αρχείο.
3. Επιβεβαιώνει την ακεραιότητα της συλλογής διαβάζοντας την ξανά από το xml αρχείο.
4. Διαβάζει τις επερωτήσεις για τη συλλογή cacm και τις φορτώνει σε κατάλληλη δομή.
5. Δημιουργεί ένα ευρετήριο και ρωτά τις επερωτήσεις επιστρέφοντας και αποθηκεύοντας τα αποτελέσματα σε κατάλληλη δομή.
6. Αποθηκεύει τα αποτελέσματα σε αρχείο επεξεργασμένα κατάλληλα ώστε να διαβάζονται από το 'trec_eval'
7. Διαβάζει τις πληροφορίες σχετικότητας επερωτήσεων-κειμένων της συλλογής και τις φορτώνει σε κατάλληλη δομή
8. Γράφει τις πληροφορίες σχετικότητας επερωτήσεων-κειμένων σε αρχείο επεξεργασμένες κατάλληλα ώστε να διαβάζονται από το 'trec_eval'
9. Καλεί το 'trec_eval' αξιολογώντας τα αποτελέσματα των επερωτήσεων.
10. Αποθηκεύει τα αποτελέσματα της αξιολόγησης σε αρχείο

```
Parsing cacm documents from file: data/cacm/cacm.all
Writing cacm documents to xml file: data/results/cacm.all.xml
Loading cacm documents from xml file: data/results/cacm.all.xml
Parsing cacm queries from file: data/cacm/query.text
Searching cacm documents with cacm queries
Writing trec-formated results to file: data/results/trec_searchresults
Parsing cacm qrels from file: data/cacm/qrels.text
Writing treq-formated qrels to file: data/results/trec_qrels
Evaluating results with trec_eval
Evaluation results are in file: data/results/trec_results
```

Επιπλέον έχω προσθέσει δύο γραμμές σε σχόλια που εκτυπώνουν τα αποτελέσματα των επερωτήσεων στην οθόνη. Αν αφαιρέσετε τα σχόλια μπορείτε να δείτε το αποτέλεσμα σχετικότητας, το κωδικό του κειμένου και το τίτλο του, για τα πρώτα 20 αποτελέσματα για κάθε επερώτηση.

```
//         System.out.println("Printing results to output");
//         printSearchResults(results);
```

→ *παράδειγμα:*

```
[...]
-----------------------------------
Searching for: 55 - Anything dealing with star height of regular languages
or regular expressions or regular events.
-----------------------------------
  2650 - 1.760452      Order-n Correction for Regular Languages
  1739 - 1.118815      Regular Expression Search Algorithm
  2921 - 1.118815      Regular Right Part Grammars and Their Parsers
  1355 - 0.839111      Regular Coulomb Wave Functions (Algorithm 292 )
  1896 - 0.839111      Regular Coulomb Wave Functions (Algorithm 292 S22])
  1898 - 0.839111      Regular Coulomb Wave Functions (Algorithm 292 [S22])
  1993 - 0.839111      Regular Coulomb Wave Functions (Algorithm 292 $S22))
  2889 - 0.134752      Performance of Height-Balanced Trees
  1694 - 0.101064      An Algorithm for the Probability of the Union of a
Large Number of Events
  1846 - 0.101064      On Simulating Networks of Parallel Processes in
Which Simultaneous Events May Occur
  2839 - 0.101064      An Insertion Technique for One-Sided Height-Balanced
Trees
  3009 - 0.101064      Insertions and Deletions In One-Sided Height-
Balanced Trees
  3056 - 0.101064      Counting Large Numbers of Events in Small Registers
[...]
```

Όπως παρατηρείτε όλη η διαδικασία γίνεται μέσα από το κώδικα της Java.
Το 'trec_eval' είναι εκτελέσιμο αρχείο:

```
/usr/bin/trec_eval: ELF 64-bit LSB executable, x86-64, version 1 (SYSV),
dynamically linked (uses shared libs), for GNU/Linux 2.6.18, stripped
```

Η Java μέσω της βιβλιοθήκης 'Runtime' μας δίνει τη δυνατότητα να τρέξουμε άλλα εκτελέσιμα αρχεία. Έτσι το 'trec_eval' καλείται μέσα από το εκτελέσιμο αρχείο της Java. Δημιουργούνται δύο νέα streams που παρακολουθούν τα std::err και std::out streams του 'trec_eval'. Το std::out παράγει τα αποτελέσματα και τα αποθηκεύει σε αρχείο, ενώ το std::err παράγει μηνύματα λάθους και τα εκτυπώνει στην οθόνη.

→ Τα αποτελέσματα του 'trec_eval' φαίνονται παρακάτω. Έγιναν δύο δοκιμές:

- Με χρήση stemming/stopwords

```
runid                   all   RUN
num_q                   all   43
num_ret                 all   860
num_rel                 all   719
num_rel_ret             all   143
map                     all   0.1786
gm_map                  all   0.0471
Rprec                   all   0.2191
bpref                   all   0.2998
recip_rank              all   0.5786
iprec_at_recall_0.00    all   0.6102
iprec_at_recall_0.10    all   0.4399
iprec_at_recall_0.20    all   0.3152
iprec_at_recall_0.30    all   0.2397
iprec_at_recall_0.40    all   0.1567
iprec_at_recall_0.50    all   0.1024
iprec_at_recall_0.60    all   0.0754
iprec_at_recall_0.70    all   0.0638
iprec_at_recall_0.80    all   0.0631
iprec_at_recall_0.90    all   0.0631
iprec_at_recall_1.00    all   0.0631
P_5                     all   0.3023
P_10                    all   0.2488
P_15                    all   0.1984
P_20                    all   0.1663
P_30                    all   0.1109
P_100                   all   0.0333
P_200                   all   0.0166
P_500                   all   0.0067
P_1000                  all   0.0033
```

- Χωρίς stemming/stopwords

```
runid                   all   RUN
num_q                   all   43
num_ret                 all   860
num_rel                 all   719
num_rel_ret             all   116
map                     all   0.1323
gm_map                  all   0.0237
Rprec                   all   0.1765
bpref                   all   0.2281
recip_rank              all   0.5040
iprec_at_recall_0.00    all   0.5299
iprec_at_recall_0.10    all   0.3896
iprec_at_recall_0.20    all   0.2547
iprec_at_recall_0.30    all   0.1307
iprec_at_recall_0.40    all   0.0729
iprec_at_recall_0.50    all   0.0588
iprec_at_recall_0.60    all   0.0465
iprec_at_recall_0.70    all   0.0465
iprec_at_recall_0.80    all   0.0465
iprec_at_recall_0.90    all   0.0465
iprec_at_recall_1.00    all   0.0465
P_5                     all   0.2233
P_10                    all   0.1860
P_15                    all   0.1504
P_20                    all   0.1349
P_30                    all   0.0899
P_100                   all   0.0270
P_200                   all   0.0135
P_500                   all   0.0054
P_1000                  all   0.0027
```

Δείτε όλα τα αρχεία ολόκληρα στον ιστότοπο της εργασίας.
Ενδεικτικά παραδείγματα υπάρχουν στον φάκελο `output`.