

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Τελική Εργασία Εργαστηρίου

ΠΑΝΤΕΛΕΗΜΩΝ ΠΡΩΙΟΣ

ice18390023

7ο Εξάμηνο

ice18390023@uniwa.gr

Τμήμα Τρίτη 16:00-17:00



ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

UNIVERSITY OF WEST ATTICA

Υπεύθυνοι καθηγητές

ΠΑΝΑΓΙΩΤΑ ΤΣΕΛΕΝΤΗ

Τμήμα Μηχανικών και Πληροφορικής Υπολογιστών

23 Ιανουαρίου 2022

Περιεχόμενα

1 Γνωριμία με την Lucene	1
1.1 Τι είναι η Lucene και τι μπορεί να κάνει	1
1.2 Στοιχεία-συνιστώσες που απαρτίζουν μία εφαρμογή αναζήτησης	1
2 Εκτέλεση εντολών	2
2.1 Εντολή Indexer	2
2.2 Εντολή Searcher	2
2.3 Εντολή SortingExample	4
3 Μηχανή αναζήτησης σε python	8

Κώδικες

3.1 Μηχανή αναζήτησης	8
---------------------------------	---

Κατάλογος σχημάτων

2.1 Αποτέλεσμα Indexer	2
2.2 Αποτέλεσμα Searcher για patent	3
2.3 Αποτέλεσμα Searcher για copyright	3
2.4 Αποτέλεσμα Searcher για mozilla	4
2.5 Αποτέλεσμα Searcher για mozilla AND NOT thunderbird	4
2.6 Αποτέλεσμα SortingExample μέρος 1	5
2.7 Αποτέλεσμα SortingExample μέρος 2	6
2.8 Αποτέλεσμα SortingExample μέρος 3	7
2.9 Αποτέλεσμα SortingExample μέρος 4	7
3.1 Εμφάνιση ερωτήματος "What is Hitler's father name"	9

1 Γνωριμία με την Lucene

1.1 Τι είναι η Lucene και τι μπορεί να κάνει

Η Lucene, είναι μια βιβλιοθήκη ανάκτησης πληροφοριών, όπου μπορεί να προστεθεί σε οποιαδήποτε εφαρμογή με αποτέλεσμα να παρέχει την δυνατότητα αναζήτησης. Είναι από τις ποιο δημοφιλής διότι:

- είναι γρήγορη
- απλή, δωρεάν
- με ανοιχτό κώδικα

Η Lucene είναι μια εργαλειοθήκη. Την ενδιαφέρει η ευρετηρίαση και η αναζήτηση κειμένου. Ταυτόχρονα επιτρέπει την εφαρμογή των business rules που αφορούν τον τομέα αυτό. Μπορεί επίσης να αναζητήσει οποιαδήποτε δεδομένα από τα οποία μπορεί να βρει κάποιο κείμενο και αδιαφορεί για την προέλευση των δεδομένων, την δομή ή την γλώσσα με αποτέλεσμα να δίνει την δυνατότητα για αναζήτηση σε πάρα πολλά δεδομένα όπως:

- ιστοσελίδες σε απομακρυσμένους servers
- αποθηκευμένα έγγραφα στα τοπικά αρχεία
- απλά έγγραφα
- έγγραφα word
- XML
- HTML
- PDF

και γενικά οποιαδήποτε αρχείο στο οποίο υπάρχει κείμενο.

1.2 Στοιχεία-συνιστώσες που απαρτίζουν μία εφαρμογή αναζήτησης

Για τον έγκαιρο εντοπισμό μια λέξης ή φράσης σε ένα κείμενο, πρέπει πρώτα να γίνει ευρετηρίαση των κειμένων. Μετά, μετατρέπονται σε έτσι ώστε να μπορεί να γίνει γρήγορη αναζήτηση. Η ευρετηρίαση όταν τερματίσει παράγει το ευρετήριο. Αυτό είναι σαν μια δομή δεδομένων που επιτρέπει την πρόσβαση στις λέξεις που το αποτελούν. Τα βήματα για την ευρετηρίαση είναι τα εξής:

1. πρόσβαση στο περιεχόμενο που θα αναζητηθεί
2. δημιουργία επιμέρους έγγραφων από το ακατέργαστο συνολικό κείμενο
3. ανάλυση του εγγράφου
4. ευρετηρίαση του εγγράφου

Το Lucene μπορεί να υλοποιήσει την ανάλυση και την ευρετηρίαση του κειμένου.

Στην ανάλυση, το έγγραφο διασπάτε σε λεκτικές μονάδες (tokens). Επίσης, είναι σημαντική η μέθοδος διαχείρισης των σύνθετων λέξεων, του ενικού και πληθυντικού προσώπου, του έλεγχου ορθογραφίας και ο καθορισμός συνωνύμων.

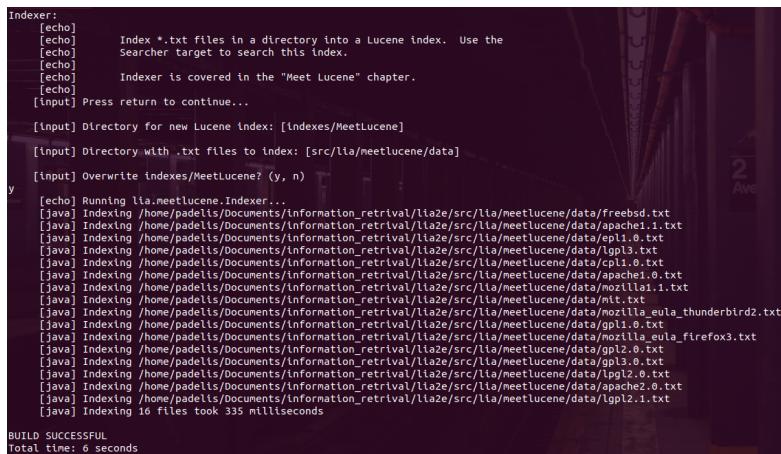
Στην ευρετηρίαση, προστίθεται το έγγραφο στο ευρετήριο και μέσο του API που παρέχει το Lucene γίνεται η ευρετηρίαση. Αναλυτικότερα, αναζητά λέξεις σε ένα ευρετήριο με σκοπό την εύρεση εγγράφων όπου τις περιέχουν. Επίσης, για την ποιότητα και την αξιοπιστία μιας αναζήτησης, υπολογίζονται μετρήσεις ακρίβειας και ανάκλησης.

Μετά το τέλος της ευρετηρίασης, ακολουθούν οι συνιστώσες αναζήτησης. Η αναζήτηση χαρακτηρίζει την διαδικασία εύρεσης εγγράφων στα οποία ανήκουν λέξεις ενός ευρετηρίου. Το πρώτο βήμα είναι η αλληλεπίδραση με τον χρήστη μέσο μιας διεπαφής. Αφού τεθεί το ερώτημα από τον χρήστη, μεταφράζεται σε Query Object το οποίο μπορεί να γίνει με την βοήθεια του QueryParser. Εν συνεχείᾳ, ακολουθεί το Search Query, μια διαδικασία αναζήτησης αποτελεσμάτων στο ευρετήριο. Το Lucene συνδυάζει τα θεωρητικά μοντέλα Pure Boolean και Vector Space Model και επιστρέφει τα ταξινομημένα έγγραφα που ταιριάζουν με το ερώτημα.

2 Εκτέλεση εντολών

2.1 Εντολή Indexer

Με την εντολή **ant Indexer**, ο Indexer εκτυπώνει τα ονόματα των αρχείων που έκανε index, με αποτέλεσμα να μπορούμε να διακρίνουμε ποια είναι αυτά τα αρχεία και πως όλα τα αρχεία είχαν .txt κατάληξη. Όταν τελειώσει, εμφανίζει των αριθμό των αρχείων και την ώρα.



```

Indexer:
[echo]
[echo] Index *.txt files in a directory into a Lucene index. Use the
[echo] Searcher target to search this index.
[echo] Indexer is covered in the "Meet Lucene" chapter.
[echo]
[echo] [input] Press return to continue...
[Input] Directory for new Lucene index: [indexes/MeetLucene]
[Input] Directory with .txt files to index: [src/lia/meetlucene/data]
[Input] Overwrite indexes/MeetLucene? (y, n)
y
[echo] Running lia.meetlucene.Indexer...
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/freebsd.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/apache1.1.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/gpl1.0.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/gpl3.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/gpl1.0.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/mozilla1.0.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/mozilla1.1.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/mozilla1.2.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/mozilla_eula_thunderbird2.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/gpl1.0.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/mozilla_eula_firefox3.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/gpl2.0.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/gpl1.0.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/apache2.0.txt
[Java] Indexing /home/padelis/Documents/information_retrival/liae/src/lia/meetlucene/data/apache2.0.txt
[Java] Indexing 16 files took 335 milliseconds
BUILD SUCCESSFUL
Total time: 6 seconds

```

Σχήμα 2.1: Αποτέλεσμα Indexer

2.2 Εντολή Searcher

Το αποτέλεσμα μετά την εκτέλεση της εντολής **ant Searcher** για την αναζήτηση των ερωτημάτων:

1. patent (default)

2. copyright
3. mozilla
4. mozilla AND NOT thunderbird

εμφανίζει ποία έγγραφα περιέχουν την αναζήτηση που κάναμε πόσα είναι αυτά και πόσο χρόνος χρειάστηκε. Ωστόσο, ο χρόνος είναι σχετικός διότι συμπεριλαμβάνεται και αυτός που πληκτρολογήθηκε το query.

```
Searcher:
[echo]
[echo]      Search an index built using Indexer.
[echo]
[echo]      Searcher is described in the "Meet Lucene" chapter.
[echo]
[input] Press return to continue...
[Input] Directory of existing Lucene index built by Indexer: [indexes/MeetLucene]

[input] Query: [patent]
[echo] Running lia.meetlucene.Searcher...

[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/cpl1.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/mozillal1.1.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/epl1.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/gpl3.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/apache2.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/gpl2.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/lpgl2.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/lGPL2.1.txt
[java] Found 8 document(s) (in 6 milliseconds) that matched query 'patent';

BUILD SUCCESSFUL
Total time: 1 second
```

Σχήμα 2.2: Αποτέλεσμα Searcher για patent

```
Searcher:
[echo]
[echo]      Search an index built using Indexer.
[echo]
[echo]      Searcher is described in the "Meet Lucene" chapter.
[echo]
[input] Press return to continue...
[Input] Directory of existing Lucene index built by Indexer: [indexes/MeetLucene]

[input] Query: [copyright]
copyright
[echo] Running lia.meetlucene.Searcher...
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/mit.txt
[java] Found 16 document(s) (in 8 milliseconds) that matched query 'copyright':
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/freebsd.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/apache1.1.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/apache1.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/apache2.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/apache2.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/gpl1.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/gpl3.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/gpl2.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/lpgl2.0.txt
[java] /home/padelis/Documents/information_retrival/liaze/src/lia/meetlucene/data/lGPL2.1.txt

BUILD SUCCESSFUL
Total time: 22 seconds
```

Σχήμα 2.3: Αποτέλεσμα Searcher για copyright

```

Searcher:
[echo]          Search an index built using Indexer.
[echo]          Searcher is described in the "Meet Lucene" chapter.
[echo]
[echo] [Input] Press return to continue...
[Input] Directory of existing Lucene index built by Indexer: [indexes/MeetLucene]
[Input] Query: [patent]
[echo] Running lia.meetlucene.Searcher...
[java] /home/padelis/Documents/information_retrieval/liae/src/lia/meetlucene/data/mozilla_eula_firefox3.txt
[java] /home/padelis/Documents/information_retrieval/liae/src/lia/meetlucene/data/mozilla_eula_thunderbird2.txt
[java] /home/padelis/Documents/information_retrieval/liae/src/lia/meetlucene/data/mozilla1.1.txt
[java] Found 3 document(s) (in 7 milliseconds) that matched query 'mozilla';

BUILD SUCCESSFUL
Total time: 6 seconds

```

Σχήμα 2.4: Αποτέλεσμα Searcher για mozilla

```

Searcher:
[echo]          Search an index built using Indexer.
[echo]          Searcher is described in the "Meet Lucene" chapter.
[echo]
[echo] [Input] Press return to continue...
[Input] Directory of existing Lucene index built by Indexer: [indexes/MeetLucene]
[Input] Query: [mozilla AND NOT thunderbird]
[echo] Running lia.meetlucene.Searcher...
[java] /home/padelis/Documents/information_retrieval/liae/src/lia/meetlucene/data/mozilla_eula_firefox3.txt
[java] /home/padelis/Documents/information_retrieval/liae/src/lia/meetlucene/data/mozilla1.1.txt
[java] Found 2 document(s) (in 13 milliseconds) that matched query 'mozilla AND NOT thunderbird';

BUILD SUCCESSFUL
Total time: 3 seconds

```

Σχήμα 2.5: Αποτέλεσμα Searcher για mozilla AND NOT thunderbird

2.3 Εντολή SortingExample

Μετά την εκτέλεση της εντολής **ant SortingExample**, επιστρέφονται με φθίνουσα σειρά score, τα έγγραφα τα οποία είναι ποιό κοντά σε αυτό που αναζητούμε.

```
SortingExample:  
[echo]  
[echo]      Lucene's sorting capabilities are demonstrated by sorting  
[echo]      the same search results in various ways.  
[echo]  
[echo]      Sorting is covered in the "Advanced searching" chapter.  
[echo]  
[input] Press return to continue...  
[Java] Running lia.advsearching.SortingExample...  
[Java] Results for: *:*(contents:java contents:action) sorted by <score>  
[Java] Title          pubmonth id   score  
[Java] Lucene in Action, Second E... 201005  3   1.052735  
[Java] /technology/computers/programming  
[Java] Ant in Action           200707  7   1.052735  
[Java] /technology/computers/programming  
[Java] Tapestry in Action       200403  2   0.447534  
[Java] /technology/computers/programming  
[Java] JUnit in Action, Second Ed.. 201005  1   0.429442  
[Java] /technology/computers/programming  
[Java] Gödel, Escher, Bach: an Et... 199905  0   0.151398  
[Java] /technology/computers/ai  
[Java] Extreme Programming Explained 200411  4   0.151398  
[Java] /technology/computers/programming/methodology  
[Java] Mindstorms: Children, Comp... 199307  5   0.151398  
[Java] /technology/computers/programming/education  
[Java] The Pragmatic Programmer    199910  6   0.151398  
[Java] /technology/computers/programming  
[Java] Imperial Secrets of Health... 199903  8   0.151398  
[Java] /health/alternative/chinese  
[Java] Lipinski, Thief of Memory   200611  9   0.151398  
[Java] /health  
[Java] Nudge : Improving Decisions... 200804  10  0.151398  
[Java] /health  
[Java] A Modern Art of Education   200403  11  0.151398  
[Java] /education/pedagogy  
[Java] Tao Te Ching 道德經           200609  12  0.151398  
[Java] /philosophy/eastern  
[Java]  
[Java] Results for: *:*(contents:java contents:action) sorted by <doc>  
[Java] Title          pubmonth id   score  
[Java] Gödel, Escher, Bach: an Et... 199905  0   0.151398  
[Java] /technology/computers/ai  
[Java] JUnit in Action, Second Ed.. 201005  1   0.429442  
[Java] /technology/computers/programming  
[Java] Tapestry in Action       200403  2   0.447534  
[Java] /technology/computers/programming  
[Java] Lucene in Action, Second E... 201005  3   1.052735  
[Java] /technology/computers/programming  
[Java] Extreme Programming Explained 200411  4   0.151398  
[Java] /technology/computers/programming/methodology  
[Java] Mindstorms: Children, Comp... 199307  5   0.151398
```

Σχήμα 2.6: Αποτέλεσμα SortingExample μέρος 1

```
[java]   /technology/computers/programming/education
[java] The Pragmatic Programmer 199910 6 0.151398
[java] /technology/computers/programming
[java] Ant in Action 208707 7 1.052735
[java] /technology/computers/programming
[java] Imperial Secrets of Health... 199903 8 0.151398
[java] /health/alternative/chinese
[java] Lipitor, Thief of Memory 200611 9 0.151398
[java] /health
[java] Nudge: Improving Decisions... 200804 10 0.151398
[java] /health
[java] A Modern Art of Education 200403 11 0.151398
[java] /education/pedagogy
[java] Tao Te Ching 道德經 200609 12 0.151398
[java] /philosophy/eastern
[java]
[java] Results for: ** (contents:java contents:action) sorted by <string: "category">
[java] Title pubmonth id score
[java] A Modern Art of Education 200403 11 0.151398
[java] /education/pedagogy
[java] Lipitor, Thief of Memory 200611 9 0.151398
[java] /health
[java] Nudge: Improving Decisions... 200804 10 0.151398
[java] /health
[java] Imperial Secrets of Health... 199903 8 0.151398
[java] /health/alternative/chinese
[java] Tao Te Ching 道德經 200609 12 0.151398
[java] /philosophy/eastern
[java] Gödel, Escher, Bach: an Et... 199905 6 0.151398
[java] /technology/computers/ai
[java] JUnit in Action, Second Ed... 201005 1 0.429442
[java] /technology/computers/programming
[java] Tapestry in Action 200403 2 0.447534
[java] /technology/computers/programming
[java] Lucene in Action, Second E... 201005 3 1.052735
[java] /technology/computers/programming
[java] The Pragmatic Programmer 199910 6 0.151398
[java] /technology/computers/programming
[java] Ant in Action 200707 7 1.052735
[java] /technology/computers/programming
[java] Mindstorms: Children, Comp... 199307 5 0.151398
[java] /technology/computers/programming/education
[java] Extreme Programming Explained 200411 4 0.151398
[java] /technology/computers/programming/methodology
[java]
[java] Results for: ** (contents:java contents:action) sorted by <int: "pubmonth">!
[java] Title pubmonth id score
[java] JUnit in Action, Second Ed... 201005 1 0.429442
[java] /technology/computers/programming
[java] Lucene in Action, Second E... 201005 3 1.052735
[java] /technology/computers/programming
[java] Nudge: Improving Decisions... 200804 10 0.151398
[java] /health
[java] Ant in Action 200707 7 1.052735
```

Σχήμα 2.7: Αποτέλεσμα SortingExample μέρος 2

```
[java] /technology/computers/programming
[java] Liptor, Thief of Memory 200611 9 0.151398
[java] /health
[java] Tao Te Ching 道德經 200609 12 0.151398
[java] /philosophy/eastern
[java] Extreme Programming Explained 200411 4 0.151398
[java] /technology/computers/programming/methodology
[java] Tapestry in Action 200403 2 0.447534
[java] /technology/computers/programming
[java] A Modern Art of Education 200403 11 0.151398
[java] /education/pedagogy
[java] The Pragmatic Programmer 199910 6 0.151398
[java] /technology/computers/programming
[java] Gödel, Escher, Bach: an Et... 199905 0 0.151398
[java] /technology/computers/ai
[java] Imperial Secrets of Health... 199903 8 0.151398
[java] /health/alternative/chinese
[java] Mindstorms: Children, Comp... 199307 5 0.151398
[java] /technology/computers/programming/education
[java]
[java] Results for: *:(contents:java contents:action) sorted by <string: "category">,<score>,<int: "pubmonth">!
[java] Title pubmonth id score
[java] A Modern Art of Education 200403 11 0.151398
[java] /education/pedagogy
[java] Nudge: Improving Decisions... 200804 10 0.151398
[java] /health
[java] Liptor, Thief of Memory 200611 9 0.151398
[java] /health
[java] Imperial Secrets of Health... 199903 8 0.151398
[java] /health/alternative/chinese
[java] Tao Te Ching 道德經 200609 12 0.151398
[java] /philosophy/eastern
[java] Gödel, Escher, Bach: an Et... 199905 0 0.151398
[java] /technology/computers/ai
[java] Lucene in Action, Second E... 201005 3 1.052735
[java] /technology/computers/programming
[java] Ant in Action 200707 7 1.052735
[java] /technology/computers/programming
[java] Tapestry in Action 200403 2 0.447534
[java] /technology/computers/programming
[java] JUnit in Action, Second Ed... 201005 1 0.429442
[java] /technology/computers/programming
[java] The Pragmatic Programmer 199910 6 0.151398
[java] /technology/computers/programming
[java] Mindstorms: Children, Comp... 199307 5 0.151398
[java] /technology/computers/programming/education
[java] Extreme Programming Explained 200411 4 0.151398
[java] /technology/computers/programming/methodology
[java]
[java] Results for: *:(contents:java contents:action) sorted by <score>,<string: "category">
[java] Title pubmonth id score
[java] Lucene in Action, Second E... 201005 3 1.052735
[java] /technology/computers/programming
[java] Ant in Action 200707 7 1.052735
```

Σχήμα 2.8: Αποτέλεσμα SortingExample μέρος 3

```
[java] /technology/computers/programming
[java] Tapestry in Action 200403 2 0.447534
[java] /technology/computers/programming
[java] JUnit in Action, Second Ed... 201005 1 0.429442
[java] /technology/computers/programming
[java] A Modern Art of Education 200403 11 0.151398
[java] /education/pedagogy
[java] Liptor, Thief of Memory 200611 9 0.151398
[java] /health
[java] Nudge: Improving Decisions... 200804 10 0.151398
[java] /health
[java] Imperial Secrets of Health... 199903 8 0.151398
[java] /health/alternative/chinese
[java] Tao Te Ching 道德經 200609 12 0.151398
[java] /philosophy/eastern
[java] Gödel, Escher, Bach: an Et... 199905 0 0.151398
[java] /technology/computers/ai
[java] The Pragmatic Programmer 199910 6 0.151398
[java] /technology/computers/programming
[java] Mindstorms: Children, Comp... 199307 5 0.151398
[java] /technology/computers/programming/education
[java] Extreme Programming Explained 200411 4 0.151398
[java] /technology/computers/programming/methodology
[java]
BUILD SUCCESSFUL
total time: 2 seconds
```

Σχήμα 2.9: Αποτέλεσμα SortingExample μέρος 4

3 Μηχανή αναζήτησης σε python

Ο κώδικας 3.1, κάνει ευρετηρίαση στα αρχεία από το path που του δίνεται (στην προκειμένη περίπτωση είναι docs) τα οποία έχουν txt κατάληξη. Κάθε ένα αρχείο, προστίθεται στο dataframe στην στήλη Text και στην στήλη Title εισάγει το όνομα του αρχείου. Έπειτα, η στήλη Text γίνεται tokenized για κάθε γραμμή και στην συνέχεια χρησιμοποιείται ο αλγόριθμος BM25 για ranking. Τέλος, ζητάει από τον χρήστη να κάνει αναζητήσει και εμφανίζει το κείμενο και το όνομα του αρχείου από τις 2 καλύτερες προσεγγίσεις όπως και των χρόνο αλλά και το πλήθος των αρχείων που εξετάστηκαν.

Ως παράδειγμα, στον κατάλογο docs, υπάρχουν 15 έγγραφα, εκ των οποίων, τα πρώτα 11 αφορούν των Χίτλερ και είναι από την wikipedia, ενώ τα υπόλοιπα είναι επίσης από την wikipedia και αφορούν την αρχαία Αίγυπτο. Το ερώτημα που δόθηκε στο παράδειγμα ήταν "What is Hitler's father name" και τα αποτελέσματα που εμφανίστηκαν, περιείχαν ακριβώς αυτό που αναζητήσαμε (εικ. 3.1).

```

1 import pandas as pd
2 import spacy
3 from tqdm import tqdm
4 from rank_bm25 import BM25Okapi
5 import time
6 import os
7
8
9 # Η συνάρτηση αυτή, διαβάζει ένα αρχείο και το επιστρέφει σε συμβολοσειρά.
10 def read_file(file_name:str):
11     text_file = open(file_name, "r", encoding="utf8")
12     data = text_file.read()
13     text_file.close()
14     return data
15
16 # Μονοπάτι που υπάρχουν τα αρχεία
17 path = "docs"
18
19 # Αποτέλεσμα αναζήτησεις αρχείων με κατάληξη txt εντός του path
20 rslt = [ x for x in os.listdir(path) if x.endswith(".txt")]
21
22 # Δημιουργία κενού dataframe
23 df = pd.DataFrame(columns = ["Title","Text"])
24
25 # Για κάθε ένα έγγραφο διάβασε το περιεχόμενο του
26 # και πρόσθεσε το στο dataframe μαζί με το όνομα του αρχείου ως τίτλο
27 for file in rslt:
28     text = read_file(path + '/' + file)
29     title = file.split('.')
30     row = {"Title": title[0], "Text": text}
31     df = df.append(row, ignore_index = True)
32
33 pd.set_option('display.max_colwidth', None)
34
35 nlp = spacy.load("en_core_web_sm")
36 tok_text=[]
37 # Κάνε tokenize
38 for doc in tqdm(nlp.pipe(df.Text.str.lower().values, disable=[ "tagger", "parser", "ner"])):
39     tok = [t.text for t in doc if t.is_alpha]
40     tok_text.append(tok)
41
42 # Χρησιμοποίησε τον αλγόριθμο BM25

```

```

43 bm25 = BM25Okapi(tok_text)
44
45 #query = "hitler's father's name"
46 # Αναζήτησε το εξής query
47 query = input('Search: ')
48
49 # Κάνε tokenize ανα κενό το query και μετετρεψέτα σε πεζά
50 tokenized_query = query.lower().split(" ")
51
52 t0 = time.time()
53 # Εμφάνισε τις δύο καλύτερες προσεγγίσεις
54 results = bm25.get_top_n(tokenized_query, df.Text.values, n=2)
55 fileName = bm25.get_top_n(tokenized_query, df.Title.values, n=2)
56 t1 = time.time()
57 print(f'Searched {len(results)} records in {round(t1-t0,3)} seconds \n')
58 for i,j in zip(results,fileName):
59     print("File name: " + path + '/' + j +'.txt')
60     print(i)

```

Κώδικας 3.1: Μηχανή αναζήτησης

```

Search: What is hitler's father name
Searched 15 records in 0.001 seconds

File name: docs/doc6.txt
Hitler's father, Alois Hitler Sr. (1837-1903), was the illegitimate child of Maria Anna Schicklgruber.[5] The baptismal register did not show the name of his father, and Alois initially bore his mother's surname, 'Schicklgruber'. In 1842, Johann Georg Hiedler married Alois's mother. Alois was brought up in the family of Hiedler's brother, Johann Nepomuk Hiedler.[6] In 1876, Alois was made legitimate and his baptismal record annotated by a priest to register Johann Georg Hiedler as Alois's father (recorded as "Georg Hitler").[7][8] Alois then assumed the surname "Hitler", [8] also spelled 'Hiedler', 'Höttler', or 'Huettler'. The name is probably based on the German word 'hütte' (lit., "hut"), and likely has the meaning "one who lives in a hut".[9]

File name: docs/doc10.txt
Alois had made a successful career in the customs bureau and wanted his son to follow in his footsteps.[29] Hitler later dramatised an episode from this period when his father took him to visit a customs office, depicting it as an event that gave rise to an unforgiving antagonism between father and son, who were both strong-willed.[30][31][32] Ignoring his son's desire to attend a classical high school and become an artist, Alois sent Hitler to the Realschule in Linz in September 1900.[c][33] Hitler rebelled against this decision, and in Mein Kampf states that he intentionally did poorly in school, hoping that once his father saw "what little progress I was making at the technical school he would let me devote myself to my dream".[34]

```

Σχήμα 3.1: Εμφάνιση ερωτήματος "What is Hitler's father name"

Οι διαφορές μεταξύ των δύο προσεγγίσεων, είναι πως η Lucene είναι γνωστή, έχει μελετηθεί καλά από πολλά άτομα έτσι ώστε να είναι γρήγορη και αποδοτική. Επίσης, παρότι υπάρχει σε java και έχει υλοποιηθεί σε αυτή, μπορεί να χρησιμοποιηθεί μέσο API και σε πολλές άλλες γλώσσες.

Το κάθε project φέρει βελτίωσης. Συγκεκριμένα η Lucene θα βελτιωνόταν αν υποστήριζε:

- Ευριστικούς αλγορίθμους αναζήτησης με αποτέλεσμα να υπάρχουν ταχύτερες αποδόσεις
- Απομακρυσμένη σύνδεση σε εξυπηρετητή για την αναζήτηση
- Ποικίλους αλγορίθμους βαθμολόγησης αποτελεσμάτων