|  |  |  |
| --- | --- | --- |
| UoP_Logo_Small | **ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ**  Σχολή Οικονομίας, Διοίκησης και Πληροφορικής  **Τμήμα Πληροφορικής και Τηλεπικοινωνιών** |  |

RASTA SEARCH ENGINE

Εργασία μαθήματος αΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

**Συγγραφέας:**

Γαλάτης Αθανάσιος, ΑΜ: 2022201500017, email: dit15017@uop.gr

Βακουφτσής Αθανάσιος, ΑΜ: 2022201500007, email: dit15007@uop.gr

**Διδάσκων:**

Τρυφωνόπουλος Χρήστος

Ιανουάριος 2019

Πίνακας περιεχομένων

1)Εισαγωγή ...4

2)Υλοποίηση ……………………………………………………………………………………………………………5

**3)Οδηγός γρήγορης εκμάθησης της μηχανής αναζήτησης Rasta .**……………………**………..6**

4)Βιβλιογραφία………………………………………………………………………………………………………. 8

# Εισαγωγή

Στην εργασία αυτή υλοποιήσαμε τη µηχανή αναζήτησης RASTA (Researcher Automated Search & Text Analytics), η οποία ευρετηριάζει και αναζητά ερευνητές µε βάση τα ερευνητικά τους ενδιαφέροντα και το δηµοσιευµένο τους έργο. Η υλοποίησή μας βασίζεται στη βιβλιοθήκη Apache Lucene, η οποία προσφέρει βασικές λειτουργίες Ανάκτησης Πληροφορίας, όπως οργάνωση της συλλογής κειµένων µε τη δηµιουργία ευρετήριου, αναζήτηση κειµένων µε ερωτήσεις εκφρασµένες σε διάφορα µοντέλα (όπως Boolean, Vector Space, φράσεων, κλπ), και κατάταξη των σχετικών αποτελεσµάτων. Επιλέξαμε να αναπτύξουμε το πρόγραμμά μας στη γλώσσα προγραμματισμού java, χρησιμοποιώντας το εργαλείο netbeans και λειτουργικό σύστημα windows κατά κύριο λόγο.

# Υλοποίηση

Έχουμε 8 κλάσεις:

* DocsCosineSimilarity
* FileInsertion
* Insertion
* Lucene
* MainFrame
* RemoveResearcher
* Search
* UrlInsertion

**DocsCosineSimilarity**

Στη συγκεκριμένη κλάση εκτελούμε την σύγκριση μεταξύ δύο ερευνητών για να βρεθεί ο βαθμός σχετικότητας μεταξύ τους, ο οποίος αναπαρίσταται σαν ένα πραγματικό αριθμό score.

**Insertion**

Στη συγκεκριμένη κλάση έχουμε τον parser των αρχείων .bib που δημιουργήσαμε με τον οποίο κρατάμε μόνο τα κατάλληλα πεδία title και booktitle για κάθε ερευνητή (κάνοντας και κατάλληλη αφαίρεση σημείων στίξης και μετατροπή των γραμμάτων σε μικρά), τα οποία εισάγουμε τόσο σαν κείμενα στο Lucene όσο και στο αντίστοιχο αρχείο .txt

**FileInsertion**

Η συγκεκριμένη κλάση αντιπροσωπεύει την εισαγωγή ερευνητών από αρχείο ανάλογα με αυτό που πληκτρολόγησε ο χρήστης. Η κλάση αποτελεί υποκλάση της Insertion από την οποία χρησιμοποιούμε τον parser (όπου επίσης γίνεται αφαίρεση σημείων στίξης και μετατροπή σε μικρά γράμματα) καθώς και τα κατάλληλα πεδία του Lucene για την εισαγωγή των κειμένων. Δημιουργείται επίσης και τοπικά στο φάκελο parsedFiles, ένα αρχείο txt με τον ερευνητή που εισήγαγε ο χρήστης προς εύρεση, όπου αποθηκεύονται κατάλληλα επεξεργασμένοι, μόνο οι τίτλοι και τα ονόματα των βιβλίων που μας ενδιαφέρουν.

**UrlInsertion**

Σε αυτή την κλάση έχουμε υλοποιήσει την αναζήτηση ενός ερευνητή από το διαδίκτυο και συγκεκριμένα από τον ιστοχώρο www.dplp.org. Η κλάση κληρονομεί επίσης την Insertion και με παρόμοιο τρόπο όπως και στη File Insertion εισάγονται τα κατάλληλα κείμενα των ερευνητών, με τη διαφορά ότι σε αυτή δημιουργούμε ένα url και βρίσκουμε διαδικτυακά το bib αρχείο με τα στοιχεία των ερευνητών.

**Lucene**

Στη συγκεκριμένη κλάση θέτουμε απλώς το MainFrame ορατό προς τον χρήστη, έτσι ώστε αυτός να μπορεί να δει τη διεπαφή και να χρησιμοποιήσει την εφαρμογή.

**MainFrame**

Σε αυτή την κλάση έχουμε την διεπαφή μας όπου μπορεί ο χρήστης να βρει όλες τις λειτουργίες της εφαρμογής μας και πατώντας στη κατάλληλη καρτέλα να αλληλεπιδράσει με όποια επιθυμεί.

**RemoveResearcher**

Στη συγκεκριμένη κλάση παίρνουμε το αλφαριθμητικό του προς διαγραφή ερευνητή που έχει πληκτρολογήσει ο χρήστης και ύστερα διαγράφουμε τον ερευνητή τόσο από το λεξικό του Lucene που έχει δημιουργηθεί όσο και από το αρχείο txt.

**Search**

Στη συγκεκριμένη κλάση έχουμε υλοποιήσει την αναζήτηση του ερωτήματος που εκτελεί ο χρήστης από τη διεπαφή.



Η διεπαφή της μηχανής αναζήτησης RASTA

# Οδηγός γρήγορης εκμάθησης της μηχανής αναζήτησης Rasta

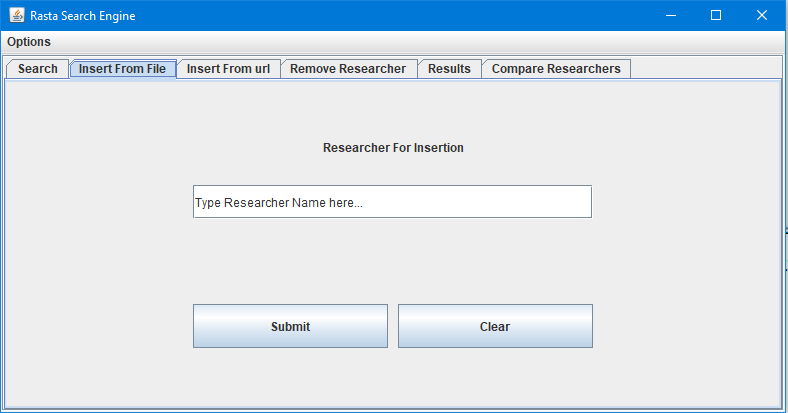


Αυτή είναι η αρχική σελίδα της μηχανής αναζήτησης όπου ο χρήστης μπορεί να ψάξει οποιοδήποτε κείμενο θέλει να βρει σε σχέση με τον ερευνητή που αναζητά.

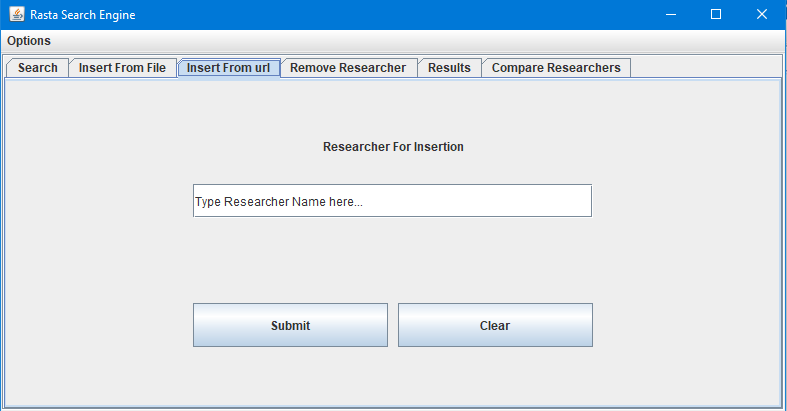
Πάνω δεξιά φαίνεται ένα μενού με όνομα Options με τις επιλογές:



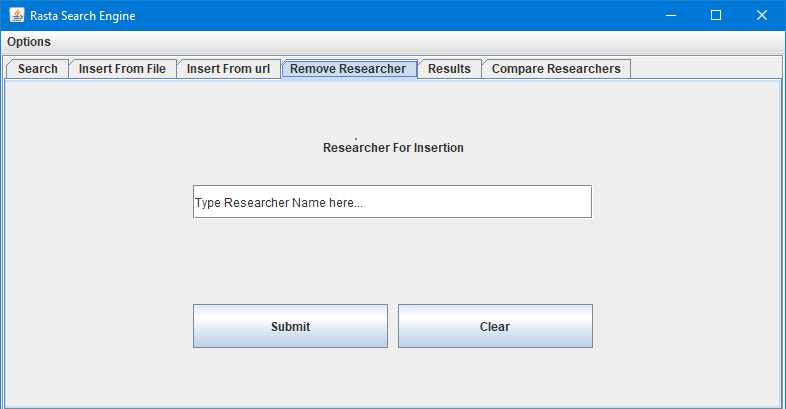
Info που εμφανίζει κάποιες βασικές πληροφορίες , credits που εμφανίζει τα ονόματα των προγραμματιστών και ένα κουμπί για τον τερματισμό του προγράμματος.



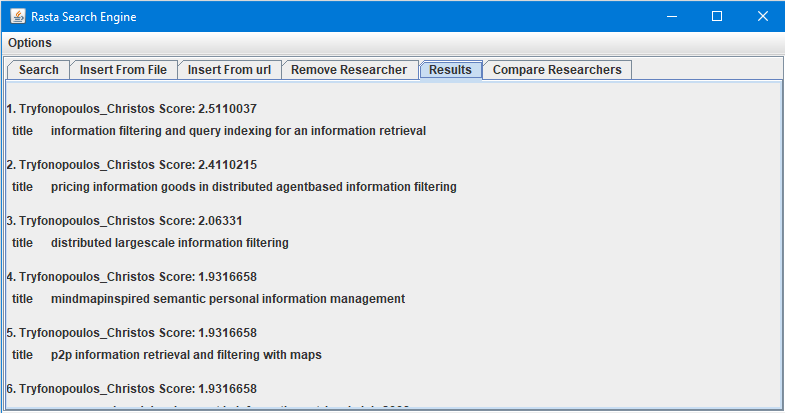
Έπειτα στην καρτέλα Insert From File βλέπουμε ένα πεδίο όπου μπορούμε να αναζητήσουμε ερευνητή με τη μορφή Επώνυμο\_Όνομα και μετά να πατήσουμε το πλήκτρο submit για να γίνει η εισαγωγή. Το κουμπί clear ’’καθαρίζει’’ το πεδίο της αναζήτησης.



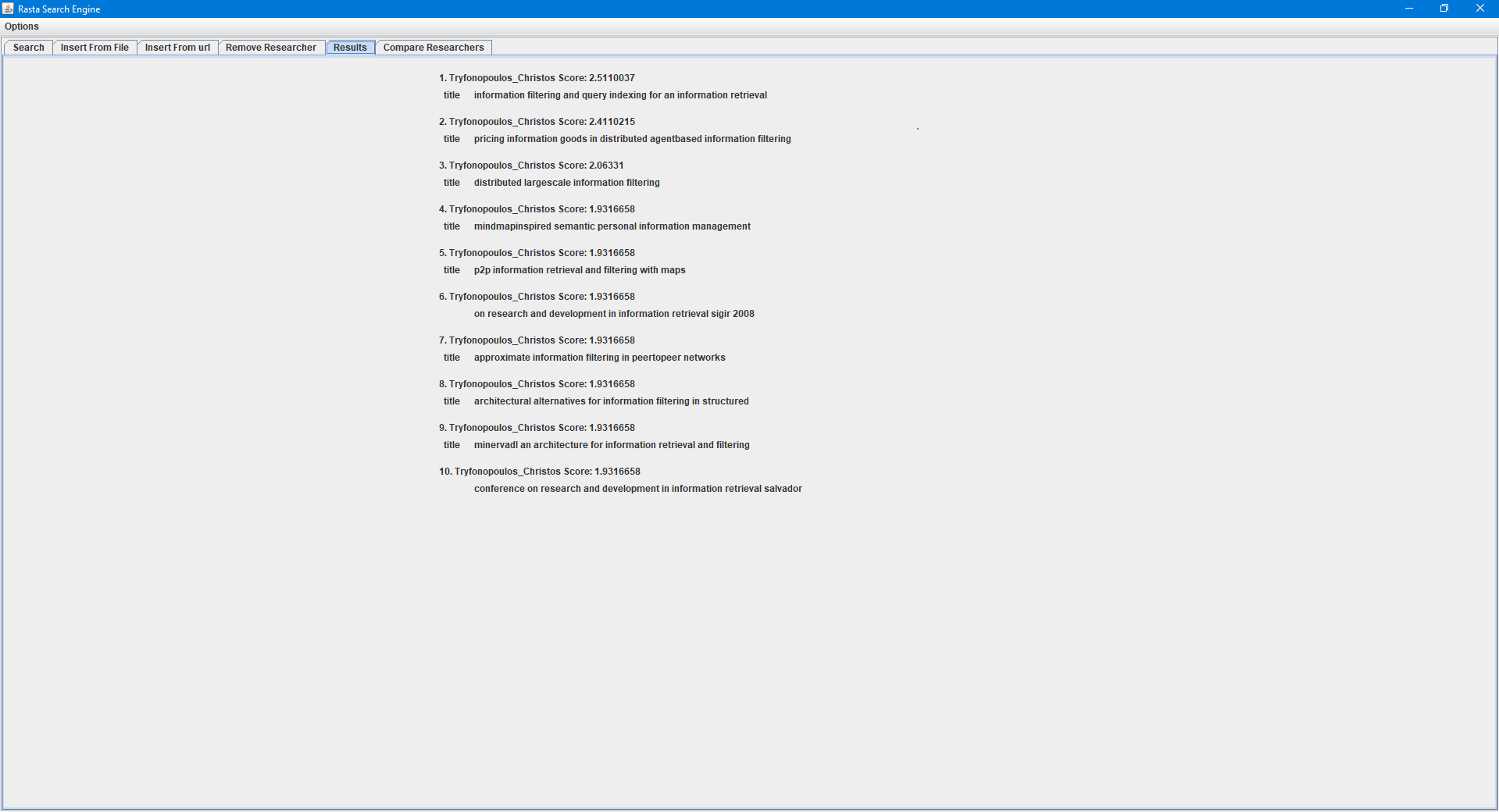
Στην καρτέλα Insert From url μπορούμε να αναζητήσουμε ερευνητή γράφοντας το όνομα του ερευνητή στο πεδίο με μορφή Επώνυμο κενό Όνομα και πατώντας το κουμπί submit γίνεται η αναζήτηση από το διαδίκτυο. Όπως και στην προηγούμενη καρτέλα του κουμπί clear ‘’καθαρίζει’’ το πεδίο αναζήτησης.

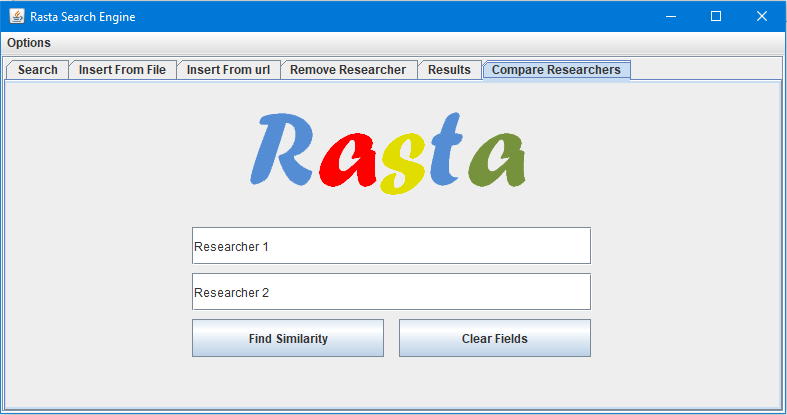


Στην καρτέλα Remove Researcher μπορούμε να πληκτρολογήσουμε στο πεδίο το όνομα του ερευνητή που θέλουμε να διαγράψουμε από την συλλογή του lucene με μορφή Επώνυμο\_Όνομα και έπειτα να πατήσουμε το κουμπί submit. Όπως και στην προηγούμενη καρτέλα του κουμπί clear καθαρίζει το πεδίο αναζήτησης.



Στην καρτέλα Results εκτυπώνονται τα δέκα πρώτα αποτελέσματα από την αναζήτηση μας.





Στην καρτέλα Compare Researchers μπορούμε να εισάγουμε τους δύο επιθυμητούς ερευνητές και να πάρουμε σε ξεχωριστό παράθυρο διαλόγου το μεταξύ τους βαθμό σχετικότητας ως αριθμό.

Τέλος η διεπαφή του προγράμματός μας ανταποκρίνεται άμεσα στην αλλαγή μεγέθους καθώς και στην οποιαδήποτε επιλογή του χρήστη.

Σημείωση!!

Επειδή δεν διαβάσαμε σωστά την εκφώνηση αντί να μπαίνουμε στον ιστοχώρο του ερευνητή και να κατεβάζουμε το .bib αρχείο εμείς κάναμε parse τον html κώδικα του αρχείου στο dblp και έπειτα συνεχίζουμε κανονικά στην προσθήκη του όπως έχουμε κάνει και στην εισαγωγή από ήδη κατεβασμένο αρχείο.

# Βιβλιογραφία

[01] https://stackoverflow.com/questions/9977221/open-url-in-java-to-get-the-content

[02] http://www.netinstructions.com/how-to-make-a-simple-web-crawler-in-java/

[03] http://www.lucenetutorial.com/lucene-in-5-minutes.html

[04] https://www.tutorialspoint.com/lucene/lucene\_booleanquery.htm

[05] https://dzone.com/articles/apache-lucene-a-high-performance-and-full-featured