

Ανάκτηση και Εξόρυξη Πληροφορίας

Μηχανή αναζήτησης
TReSA

Ευράφης Γιώργος
dit18138

Τζανερά Στεφανία
dit18201



12/01/2022

Υλοποίηση

Προεπεξεργασία

Η μηχανή αναζήτησης TReSA υλοποιήθηκε σε Java με την βοήθεια της βιβλιοθήκης Apache Lucene, ενώ το γραφικό περιβάλλον για την εφαρμογή αυτή υλοποιήθηκε με χρήση JavaFX. Για να να μπορέσουμε να εξηγήσουμε το πως υλοποιήθηκε η εργασία αυτή, θα αναλύσουμε μια τις λειτουργίες της. Με το που ανοίξει ο χρήστης την εφαρμογή, του δίνεται η επιλογή είτε να επεξεργαστεί την λίστα των άρθρων που συμμετέχουν στο αντεστραμμένο ευρετήριο, είτε να κάνει αναζήτηση. Εάν διαλέξει να επεξεργαστεί την λίστα των άρθρων, μεταφέρεται σε μια καινούρια οθόνη όπου υπάρχουν δύο λίστες. Η αριστερή λίστα είναι διάφορα άρθρα (αρχεία txt) που υπάρχουν στο directory “Reuters_articles” ενώ η δεξιά αποτελείται από άρθρα που έχουν ήδη γίνει indexed. Για να το καταφέρουμε αυτό, με τις κατάλληλες μεθόδους όπως φαίνεται στην μέθοδο “getFileNames” της κλάσης “functionality.java”, σκανάρουμε τον φάκελο και κρατάμε όλα τα αρχεία που τελειώνουν σε txt. Όσο για την δεξιά λίστα, αντίστοιχα με την βοήθεια ενός IndexReader διαβάζουμε το αντεστραμμένο ευρετήριο και κρατάμε το όνομα κάθε Document. Ο χρήστης έχει επίσης την δυνατότητα να αλλάξει τον φάκελο εισαγωγής αρχείων προκειμένου να προσθέσει δικά του αρχεία.

Κάθε νέο αρχείο που προστίθεται στο ευρετήριο, είτε είναι του χρήστη είτε από την συλλογή άρθρων, περνάει από λεκτική ανάλυση όπως φαίνεται στην κλάση PreProcessing.java. Στην μέθοδο articlesEditor διαβάζεται το αρχείο γραμμή προς γραμμή. Σε κάθε γραμμή ελέγχουμε αν περιέχεται κάποια από τις ετικέτες μας (<TITLE>,<BODY>,<PLACE>,<PEOPLE>). Αν υπάρχει ανανεώνουμε την μεταβλητή category με την νέα κατηγορία και αποθηκεύουμε το υπόλοιπο κείμενο στην μεταβλητή content μέχρι να βρούμε την ετικέτα κλεισίματος π.χ. </TITLE>. Όταν την βρούμε περνάμε στο επόμενο στάδιο όπου ανάλογα την ετικέτα στέλνουμε το κείμενο για την κατάλληλη προεπεξεργασία. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να διαβαστεί όλο το κείμενο.

Στην προεπεξεργασία, με την βοήθεια του CustomAnalyzer μαζί με τις παροχές που έχει από μόνη της η Τζάβα, γίνονται τα ακόλουθα. Στον τίτλο του άρθρου κάνουμε όλα τα γράμματα μικρά (case folding) και αφαιρούμε σημεία στίξης (punctuation removal), καθώς και τελείες από ακρωνύμια, εκτός αν πρόκειται για δεκαδικό αριθμό. Επίσης τον περνάμε από αλγόριθμο PorterStemmer για να κρατήσουμε τις ρίζες (περίπου) των λέξεων. Αφού ο τίτλος είναι ήδη μικρός και ορισμένα stopwords μπορεί να είναι μέρος του (π.χ. υπάρχει συγκρότημα που λέγεται “The the”) δεν περνάει άλλη προεπεξεργασία.

Τα μέρη και τα άτομα περνάνε την ίδια προεπεξεργασία, συγκεκριμένα case folding και punctuation removal. Τα stopwords τα αφήνουμε καθώς μπορεί να αποτελούν μέρος του ονόματος μιας χώρας ή ενός ατόμου και το stemming θα ήταν ανούσιο.

Τέλος, όσον αφορά το κύριο σώμα του άρθρου, αφού εκεί βρίσκεται ο κύριος όγκος πληροφορίας, το περνάμε από case folding, punctuation removal, stemming και stop-word removal.

Αναζήτηση

Όταν ο χρήστης πατήσει το κουμπί της αναζήτησης, μεταφέρεται σε μια νέα οθόνη όπου μπορεί να επιλέξει να κάνει μια “Γρήγορη αναζήτηση”, εισάγοντας απλώς κάποια φράση στο πεδίο αναζήτησης και διαλέγοντας τον επιθυμητό αριθμό αποτελεσμάτων, ή μπορεί να δοκιμάσει για κάποιου είδους “Advanced αναζήτησης”.

Στην γρήγορη αναζήτηση, δημιουργείται ένα περίπλοκο query, το οποίο αν σπαστεί σε κομμάτια μπορούμε να δούμε πως πρακτικά κάνει τα εξής: Ελέγχει αν ο τίτλος, τα άτομα ή τα μέρη περιέχουν κάποια από τις λέξεις που έδωσε ο χρήστης, αφότου αυτές περάσουν κατάλληλη προεπεξεργασία για να ταιριάζουν με τα αντίστοιχα πεδία. Έπειτα ελέγχει αν ο τίτλος ή το κύριο σώμα περιέχουν κάποια από τις λέξεις και δίνει μεγαλύτερη βαρύτητα σε αυτόν τον έλεγχο απ ότι πριν, καθώς αν το σώμα περιέχει τις λέξεις, το άρθρο θα είναι πιο σχετικό. Τέλος ελέγχει αν το σώμα και ο τίτλος περιέχουν όλες τις λέξεις που έδωσε ο χρήστης και εκεί δίνει ακόμα μεγαλύτερη βαρύτητα στο score του document. Με βάση το query αυτό επιστρέφουν τα αποτελέσματα και ο χρήστης οδηγείται σε νέα οθόνη, όπου παρουσιάζονται τα άρθρα κατά σειρά σχετικότητας μαζί με τις λέξεις από την αναζήτηση του που βρέθηκαν σε αυτά, ή ενημερώνεται πως δεν βρέθηκαν αποτελέσματα από το ευρετήριο αν δεν έχουν βρεθεί. Προφανώς από αυτό το μενού μπορεί να επιλέξει πιο άρθρο θα διαβάσει.

Επιλέγοντας “Advanced” αναζήτηση, παρέχονται στον χρήστη οι εξής επιλογές: Αναζήτηση παραπλήσιων άρθρων με κάποιο που θα επιλέξει, Boolean αναζήτηση, Αναζήτηση με βάση κάποιο πεδίο.

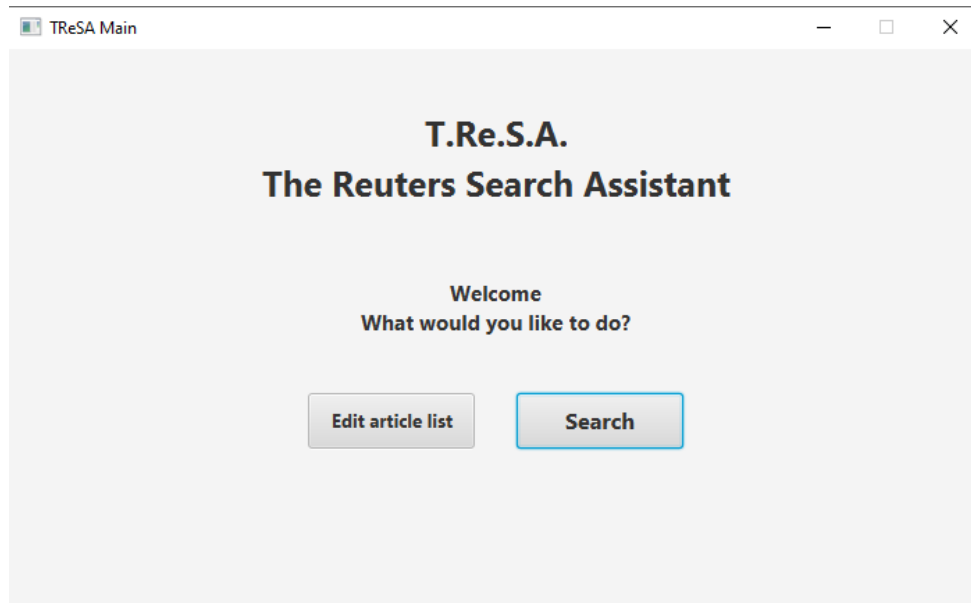
Στην αναζήτηση παραπλήσιου άρθρου, στον χρήστη παρέχεται μια λίστα με άρθρα που είναι περασμένα στο ευρετήριο και οι επιλογές είτε να τα διαβάσει, είτε να κάνει αναζήτηση με βάση αυτά, καθώς και το πόσα αποτελέσματα θέλει να πάρει. Άμα επιλέξει την αναζήτηση, από το άρθρο κρατάμε τον τίτλο του και το κυρίως σώμα του και με βάση αυτά δημιουργούμε αντίστοιχα ένα query το οποίο δίνει παραπάνω βάρος αν και ο τίτλος και το σώμα μοιάζουν, ενώ λιγότερο αν ισχύει μόνο ένα από τα δύο. Με το query αυτό γίνεται αναζήτηση και παρέχονται στον χρήστη τα

αποτελέσματα προς ανάγνωση, μαζί με τα σκορ και ορισμένες (μαξ 6) κοινές λέξεις.

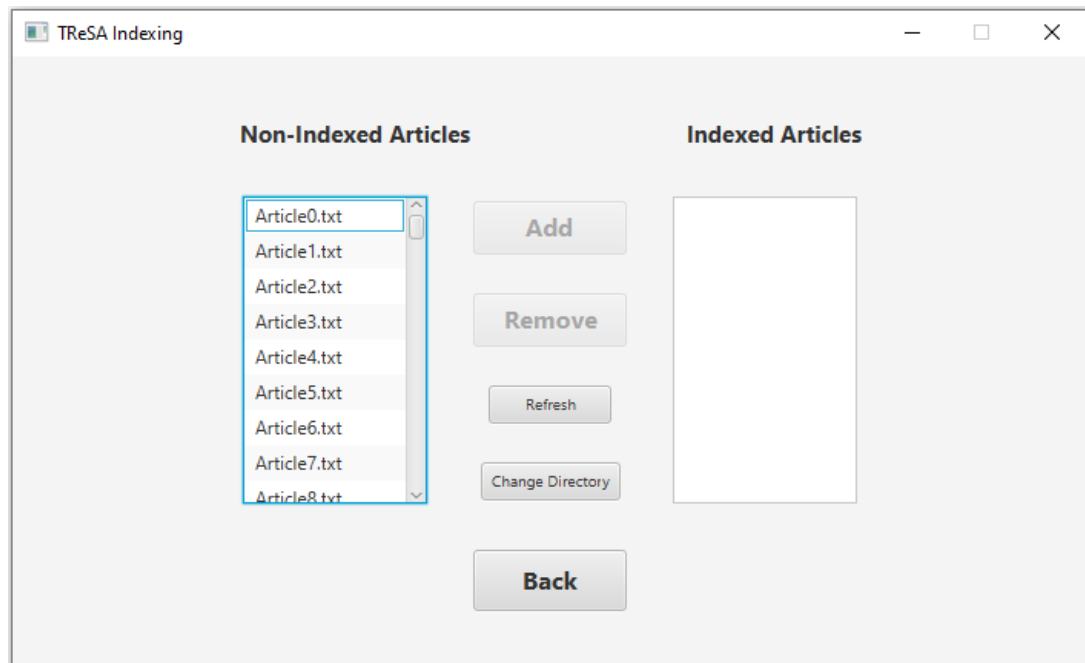
Στην αναζήτηση με βάση κάποιο/α πεδίο/α ο χρήστης μεταφέρεται σε μια νέα οθόνη, στην οποία υπάρχουν 4 text fields στα οποία μπορεί να γράψει ο χρήστης. Το καθένα αντιστοιχεί σε ένα από τα πεδία που έχει το κάθε άρθρο και αυτό αναγράφεται από δίπλα του. Ο χρήστης μπορεί να δώσει όρους προς αναζήτηση σε ένα ή περισσότερα πεδία και να επιλέξει τον αριθμό επιθυμητών αποτελεσμάτων. Αφότου πατήσει αναζήτηση, δημιουργείται query με βάση τα πεδία και τους όρους που έχει δώσει, τέτοιο ώστε να χρειάζεται να χρειάζεται να ταιριάζουν σε όλα τα πεδία που έχει δώσει οι όροι προκειμένου να πάρει αποτέλεσμα. Γίνεται δηλαδή λογικό και μεταξύ των πεδίων. Από εκεί οδηγείται στην κλασσική οθόνη με τα αποτελέσματα, τα σκορ τους και τις κοινές λέξεις που βρέθηκαν στο σώμα του άρθρου.

Τέλος, επιλέγοντας την Boolean αναζήτηση, ο χρήστης μεταφέρεται σε μια νέα οθόνη όπου του παρέχεται ένα σετ οδηγιών με ορισμένα παραδείγματα για την σύνταξη των boolean ερωτήσεων. Δυστυχώς, ο 1ος τρόπος χρήσης από το σετ οδηγιών δεν υλοποιήθηκε, καθώς θεωρήθηκε πως ήταν πολύ παρόμοιος με την γρήγορη αναζήτηση σαν λογική, αλλά ξεχάσαμε να αλλάξουμε τις οδηγίες. Κρατώντας τον 2ο τρόπο, ο χρήστης μπορεί να γράψει ένα Boolean Query χρησιμοποιώντας τους τελεστές AND, OR και NOT μαζί με ένα ή περισσότερα πεδία από αυτά του άρθρου. Από εκεί και ύστερα, υπάρχει μια σχετικά μεγάλη διαδικασία λεκτικής ανάλυσης, αφού πρέπει να σιγουρευτούμε ότι οι λογικοί τελεστές δεν θα αφαιρεθούν από το Query κατά την επεξεργασία του ως stop-words, αλλά θα παραμείνουν μέχρι το τέλος για να μπορέσει να γίνει σωστά η αναζήτηση. Αυτό το επιτυγχάνουμε αλλάζοντας προσωρινά τις λέξεις σε άλλες που έχουν μορφή τέτοια ώστε να μείνουν αναλλοίωτες από την επεξεργασία και έπειτα τις ξανα αλλάζουμε στην αρχική τους μορφή. Με βάση το Query αυτό γίνεται αναζήτηση και τα αποτελέσματα προβάλλονται στον χρήστη μαζί με το σκορ τους και τις κοινές λέξεις που βρέθηκαν.

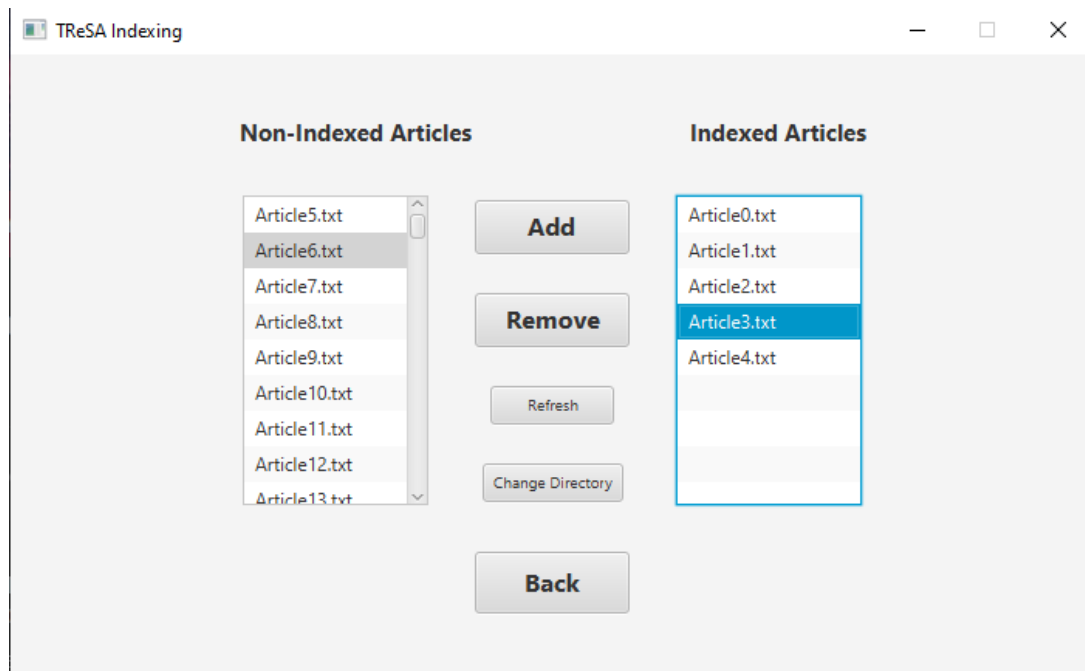
Πρακτική Παρουσίαση των Προηγούμενων/GUI



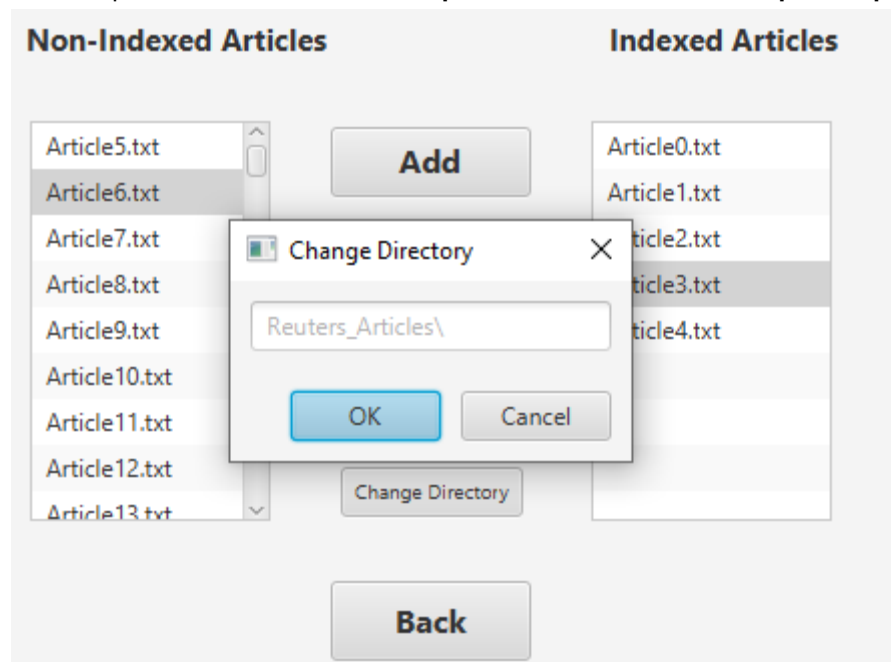
Αρχική οθόνη



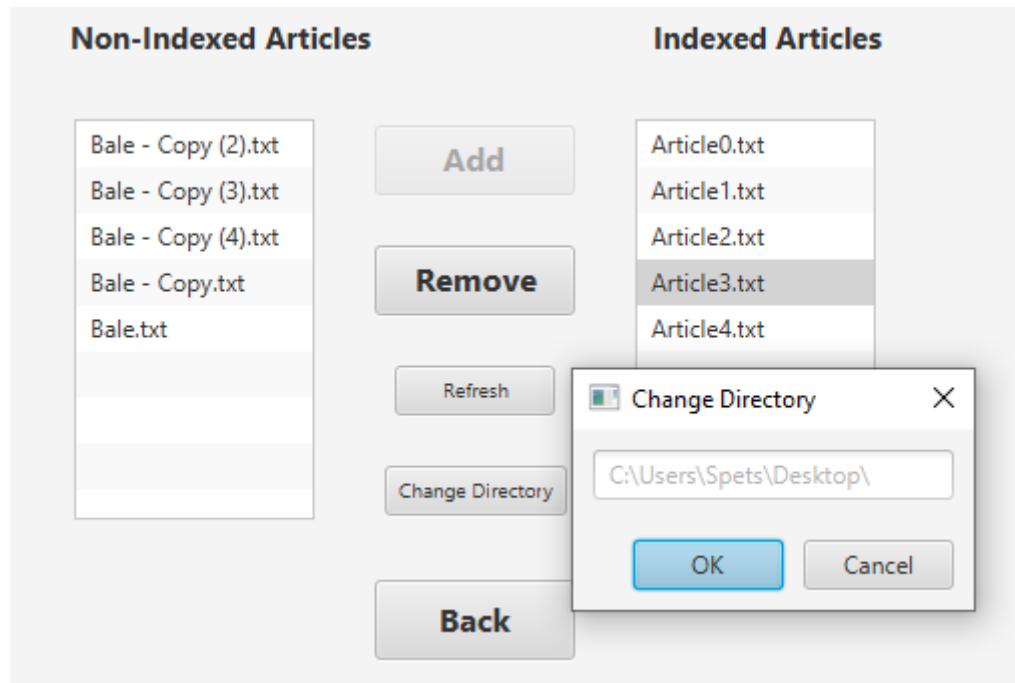
Edit Article List. Βρισκόμαστε στο default directory "Reuters_articles". Τα κουμπιά Add και Remove παραμένουν κλειδωμένα μέχρι να επιλεχθούν αντικείμενα από τις κατάλληλες λίστες.



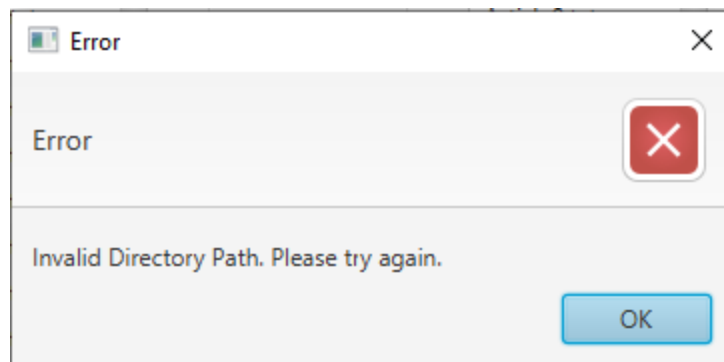
Με τα κουμπιά add και remove ο χρήστης μπορεί να προσθέτει και να αφαιρεί άρθρα από το ευρετήριο. Με το refresh ανανεώνει τον φάκελο στον οποίο βλέπει τα non indexed άρθρα (χρήσιμο αν πειράζει manually τον φάκελο και προσθέσει/αφαιρέσει άρθρα) ενώ με το change directory μπορεί να αλλάξει τον φάκελο στον οποίο βλέπει όπως θα δούμε παρακάτω.



Pop up window αφότου πατήσουμε change directory. Με γκρι γράμματα αναγράφεται το directory που βρίσκεται τώρα ο χρήστης.



Αλλαγή directory στην επιφάνεια εργασίας όπου ανιχνεύτηκαν ορισμένα άλλα άρθρα. Φαίνεται το directory στο prompt ενώ το κουμπί add είναι απενεργοποιημένο αφού δεν έχει επιλεγθεί κανένα άρθρο.



Μήνυμα λάθους σε περίπτωση που δοθεί directory που δεν μπορεί να προσδιοριστεί/δεν υπάρχει.

The screenshot shows a window titled "TReSA Search". In the top-left corner is a "Back" button. The main area features the "TReSA" logo in the center. Below the logo is a search input field containing the text "Iran-Iraq war", followed by a "Search" button. Underneath the search field is a label "Number of results:" followed by a text box containing the number "10". At the bottom center is a button labeled "► Advanced Search".

Οθόνη γρήγορης αναζήτησης που περιγράψαμε στην προηγούμενη ενότητα.

The screenshot shows an "Error" dialog box. It has a title bar with the word "Error" and a close button. The main area contains the word "Error" and a red square icon with a white "X". Below this, a message reads: "Invalid Input. Please insert a positive integer at the 'Number of results' field." At the bottom right is an "OK" button.

Πορ up παράθυρο λάθους σε περίπτωση που αναζητήσουμε με μη έγκυρο αριθμό αποτελεσμάτων (γράμματα, 0, αρνητικοί αριθμοί κλπ)

TReSA Search

Back

IRAN CLAIMS NEW VICTORIES NEAR BASRA
Relevant words: ...war...
Matching Score: 60.316402

IRAN ANNOUNCES END OF MAJOR OFFENSIVE IN GULF WAR
Relevant words: ...war...
Matching Score: 60.159836

IRAQ SAYS IT REPELS IRANIAN ATTACK
Relevant words:
Matching Score: 23.066011

IRAQ DEFERS PAYMENTS ON 500 MLN DLR EUROLOAN
Relevant words: ...war...
Matching Score: 22.017673

SHULTZ SAYS NO RESIGNATION OVER IRAN REPORT

Αποτελέσματα της προηγούμενης αναζήτησης. Οι τίτλοι υπογραμμίζονται όταν το ποντίκι περνάει από πάνω για να ξέρει ο χρήστης πως μπορεί να κλικάρει.

TReSA Search

Back

IRAN CLAIMS NEW VICTORIES NEAR BASRA
Places involved: ukiraniraq

Iran said its forces had captured one of Iraq's strongest fortifications east of Basra on the Gulf War southern front in a major battle overnight.

The Iranian National News Agency, received here, said Iranian forces smashed four Iraqi brigades, killed or wounded 1,500 Iraqi soldiers and destroyed 45 enemy tanks and personnel carriers.

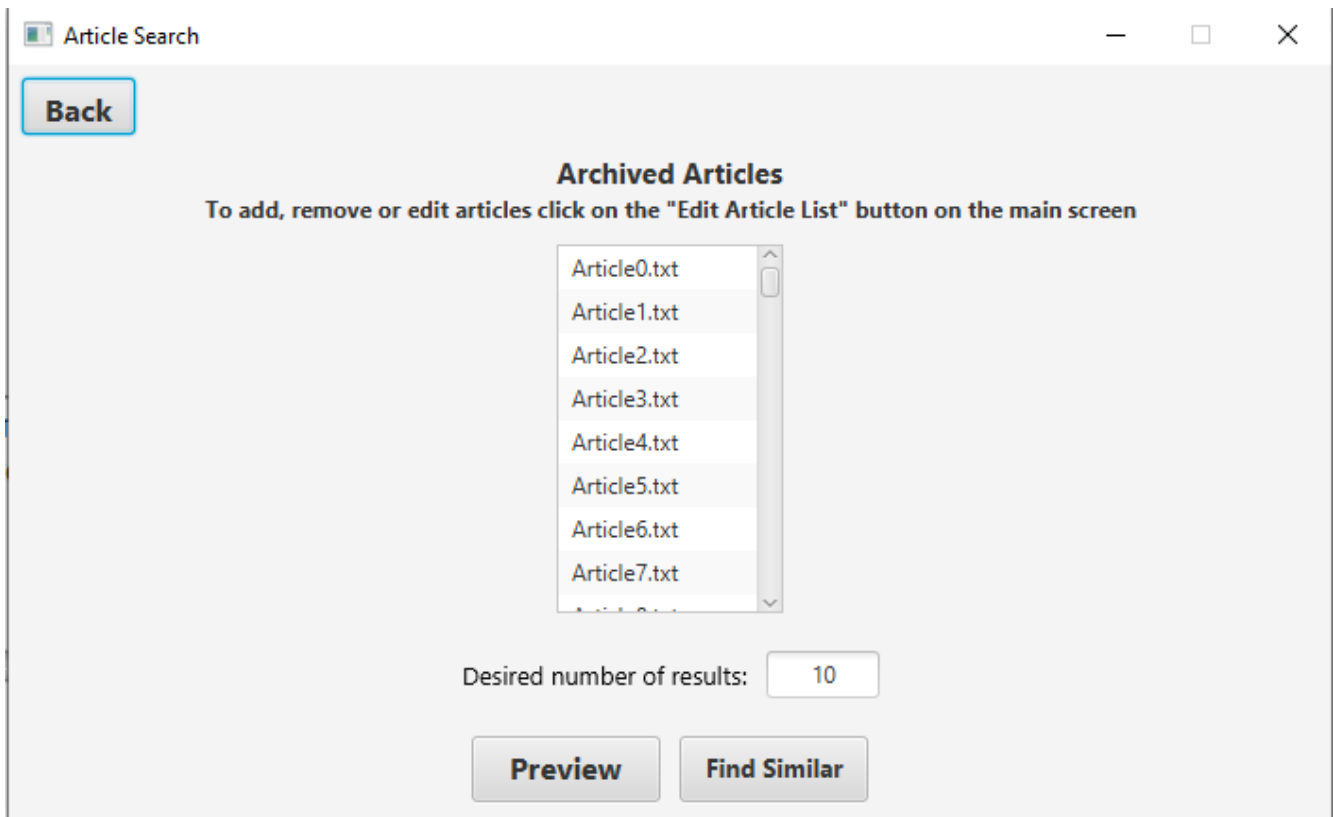
IRNA said the Iranian troops seized one of the strongest Iraqi fortifications and breached Iraqi defence lines southwest of Fish Lake, 10 kilometres (six miles) east of Iraq's second largest city of Basra.

REUTER

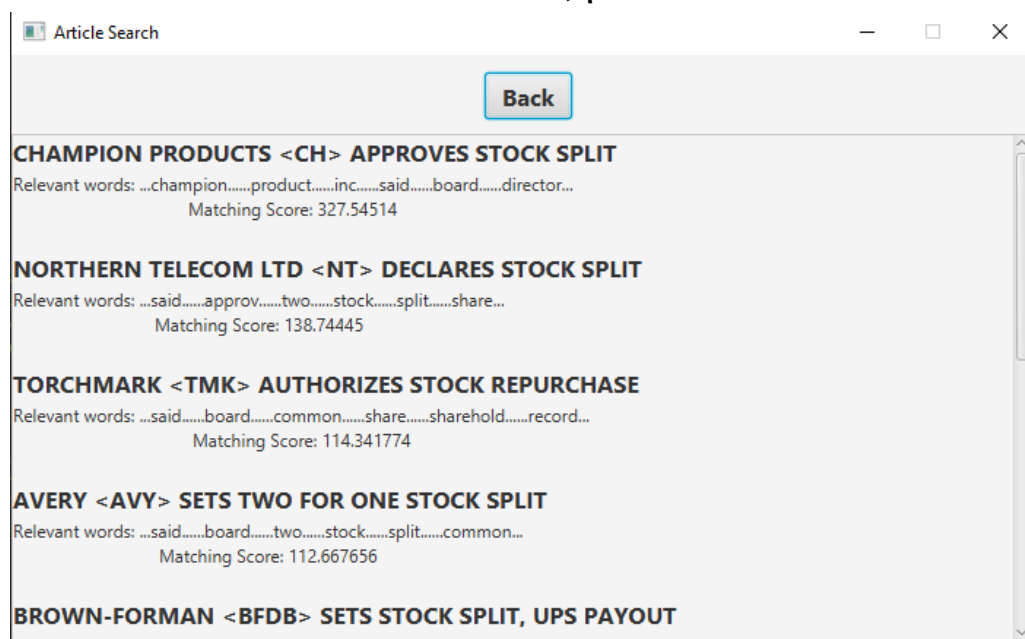
□

Παρουσίαση άρθρου. Φαίνονται ο τίτλος, οι τοποθεσίες και το σώμα του άρθρου. Οι τοποθεσίες καθώς και τα άτομα γίνονται εμφανή κάτω από τον

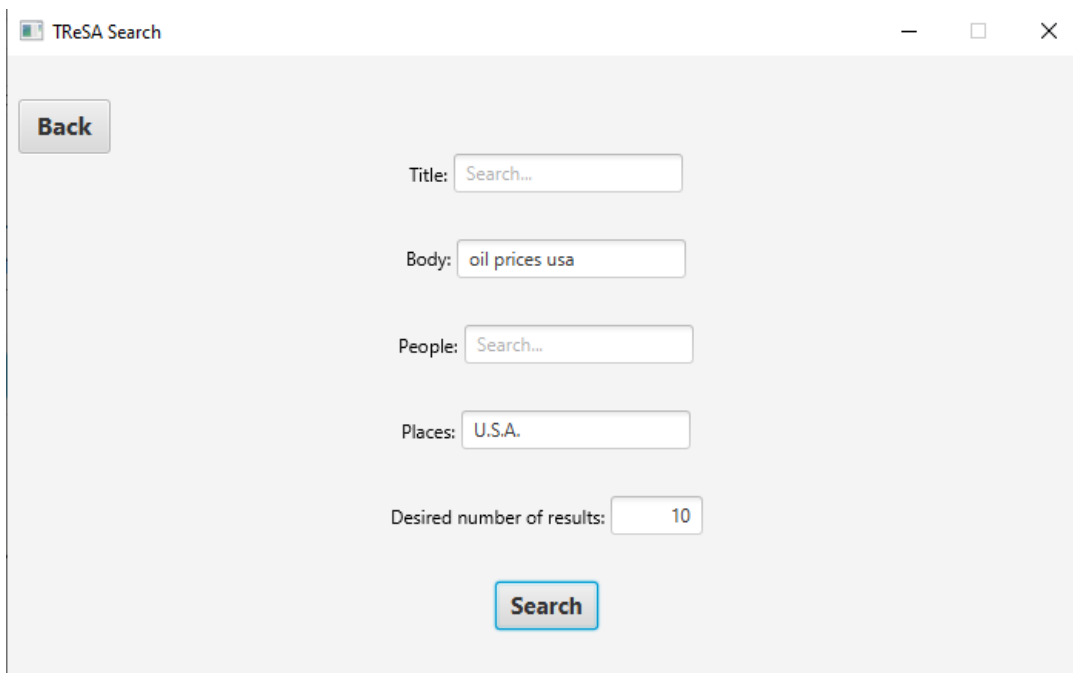
τίτλο μόνο εφόσον υπάρχουν στο άρθρο. Στην συγκεκριμένη περίπτωση δεν υπήρχαν άτομα.



Αναζήτηση παρομοίων άρθρων. Το Preview επιτρέπει στον χρήστη να διαβάσει το άρθρο όπως φαίνεται στην προηγούμενη εικόνα, ενώ με το find similar αναζητά.



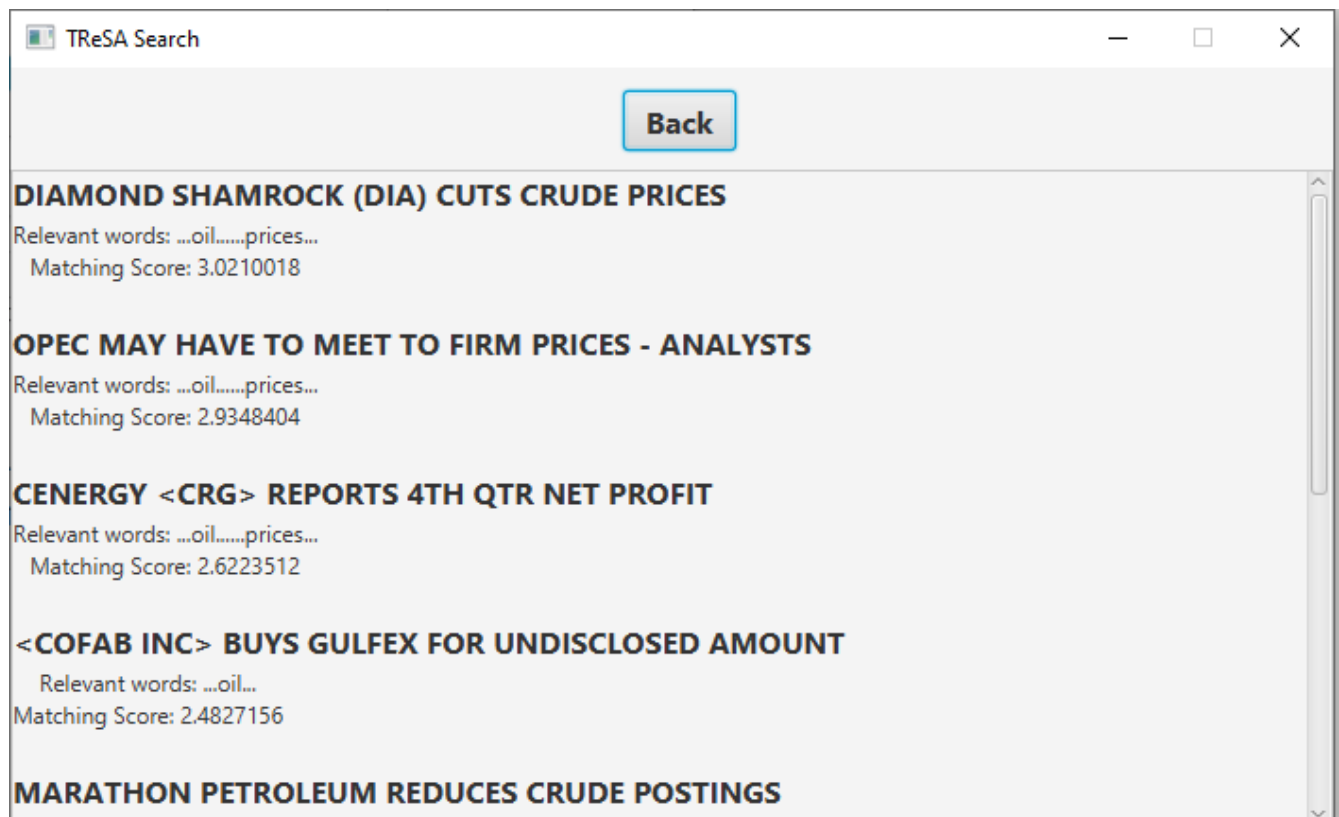
Αποτελέσματα αναζήτησης για το άρθρο 8.



The screenshot shows the TReSA Search window with the following fields and values:

- Back** button
- Title:** Search...
- Body:** oil prices usa
- People:** Search...
- Places:** U.S.A.
- Desired number of results:** 10
- Search** button

Παράδειγμα αναζήτησης ανά πεδίο



The screenshot shows the TReSA Search results window with a **Back** button at the top. The results are listed below:

- DIAMOND SHAMROCK (DIA) CUTS CRUDE PRICES**
Relevant words: ...oil.....prices...
Matching Score: 3.0210018
- OPEC MAY HAVE TO MEET TO FIRM PRICES - ANALYSTS**
Relevant words: ...oil.....prices...
Matching Score: 2.9348404
- CENERGY <CRG> REPORTS 4TH QTR NET PROFIT**
Relevant words: ...oil.....prices...
Matching Score: 2.6223512
- <COFAB INC> BUYS GULFEX FOR UNDISCLOSED AMOUNT**
Relevant words: ...oil...
Matching Score: 2.4827156
- MARATHON PETROLEUM REDUCES CRUDE POSTINGS**

Αποτελέσματα. Όπως περιμένουμε, σε όλα τα άρθρα αν τα ανοίξουμε, στο Places έχει usa.

Boolean Search

Back

TReSA Boolean Search

Instructions:

There are 2 ways to use this search feature.

1) You can search for terms or sentences separated by the Boolean operators (OR, AND, NOT) and can also use parenthesis.

Example: (USA OR BRITAIN) AND OIL PRICES

2) Same as the above, but you can also specify specific fields of the article in your query.

Example: (Title: USA AND Oil OR Places: USA AND Body: oil) NOT (People: George)

Valid fields are "Title", "Body", "Places", "People" and MUST be written in that order. If any are omitted, the order remains the same without them.

Title: oil AND Body: (oil OR prices) NOT Places: usa **Search**

Desired number of results:

Boolean αναζήτηση. Όπως αναφέρθηκε και πριν, το 1ο σετ οδηγιών δεν ισχύει, αλλά το 2ο μπορεί να ακολουθηθεί.

Boolean Search

Back

GULF ARAB OIL MEETING ENDS
Relevant words: ...oil.....AND.....(oil.....OR.....prices).....NOT...
Matching Score: 5.069264

OIL PRICES RISE ON SAUDI EFFORT
Relevant words: ...oil.....AND.....(oil.....OR.....prices).....NOT...
Matching Score: 4.8835273

MALAYSIA RAISES DUTY ON PROCESSED PALM OIL
Relevant words: ...oil.....AND.....(oil.....OR.....prices).....NOT...
Matching Score: 4.837585

GULF ARAB DEPUTY OIL MINISTERS TO MEET IN BAHRAIN
Relevant words: ...oil.....AND.....(oil.....OR.....prices).....NOT...
Matching Score: 4.6199827

OILS/FATS STOCKS SEEN FALLING SHARPLY IN 1986/87

Αποτέλεσμα αναζήτησης. Δυστυχώς αφού πλέον ορισμένα stopwords είναι μέρος του query, όταν θέλουμε να βρούμε matching words του query αργότερα, μας δείχνει και αυτά.

Για να τρέξει

Προκειμένου να τρέξει το πρότζεκτ χρειάζεται να γίνουν τα ακόλουθα: Αρχικά πρέπει να εγκαταστήσουμε και να προσαρμόσουμε το JavaFX για Eclipse όπως φαίνεται στο παρακάτω βίντεο:

<https://www.youtube.com/watch?v=bk28ytggz7E>

Έπειτα, επειδή το lucene δεν συνεργάζεται πολύ καλά με το package management της Java, αναγκάστηκα να συνενώσω δύο jar (common-analyzers και core) σε ένα άλλο jar, μια διαδικασία που οι δημιουργοί του lucene ονόμασαν “δημιουργία uberjar” οπότε κατά συνέπεια του έδωσα κ εγώ αυτό το όνομα. Το αρχείο αυτό παρέχεται μέσα στο zip της εργασίας.

Από το eclipse θα χρειαστεί να πατήσουμε δεξί κλικ στο πρότζεκτ αφότου το φορτώσουμε, να πάμε στο build path → configure build path και εκεί στην καρτέλα libraries να προσθέσουμε στο modulepath ως external jars τα “lucene-queryparser-8.9.0.jar”, “lucene-queries-8.9.0.jar” και το “uberjar.jar. Έπειτα πατάμε apply και κλείνουμε το παράθυρο. Στην συνέχεια κάνουμε δεξί κλικ πάνω στο Project → Run As → Run Configurations και πηγαίνουμε στα dependencies και προσθέτουμε τα παραπάνω αρχεία στο module path και στο classpath. Επίσης προσθέτουμε ό,τι αρχεία JavaFX χρειάζονται όπως φαίνεται στο βίντεο. Τέλος, πάλι όπως φαίνεται στο βίντεο, στο μενού που βρισκόμαστε ήδη πατάμε στην καρτέλα “Arguments” και στα VM arguments προσθέτουμε “--module-path "C:\Path\to\lib" --add-modules javafx.controls,javafx.fxml”

Μετά από αυτό, μπορούμε να τρέξουμε κανονικά το πρόγραμμα.

Πηγές:

<https://www.youtube.com/watch?v=bk28ytggz7E>

<https://www.baeldung.com/lucene-analyzers>

https://lucene.apache.org/core/9_0_0/index.html

<https://stackoverflow.com/questions/2005084/how-to-specify-two-fields-in-lucene-queryparser>

https://lucene.apache.org/core/2_9_4/queryparsersyntax.html

<https://stackoverflow.com/questions/968297/how-to-do-search-of-part-of-a-word-using-lucene>