Ανάκτηση Πληροφορίας

GitHub Repository Link

Βασίλειος Σκαραφίγκας ΑΜ: 4491

Κωνσταντίνος Χριστόπουλος ΑΜ: 4527

Ενότητα 1 : Εισαγωγή

Lucene πρόγραμμα, ειδικά σχεδιασμένο για στίχους τραγουδιών. Το σύστημα διαβάζει δεδομένα στίχων από ένα αρχείο CSV, ευρετηριάζει τα δεδομένα αυτά με τη χρήση του Lucene και δίνει τη δυνατότητα στους χρήστες να αναζητήσουν τα ευρετηριασμένα αυτά δεδομένα χρησιμοποιώντας λέξεις-κλειδιά. Οι χρήστες μπορούν να φιλτράρουν την αναζήτησή τους με βάση πεδία όπως το όνομα του καλλιτέχνη, το όνομα του τραγουδιού και τους στίχους. Το σύστημα διατηρεί επίσης ένα ιστορικό των ερωτημάτων των χρηστών και παρέχει προτάσεις για μελλοντικά ερωτήματα.

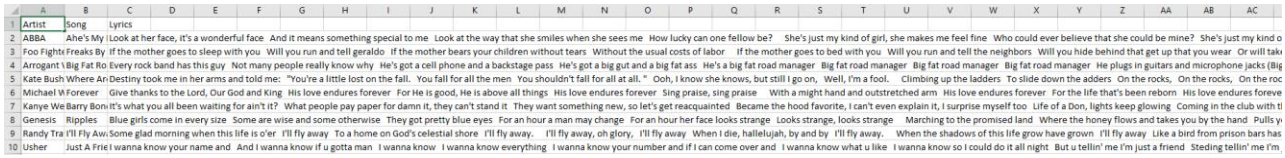
Ενότητα 2 : Συλλογή

Για την δημιουργία της συλλογής (*corpus*) κατεβάσαμε μια έτοιμη συλλογή από το kaggle:

- spotify million song dataset

Η συλλογή αυτή περιέχει:

- 643 διαφορετικούς artists
- 44824 διαφορετικά ονόματα τραγουδιών
- 57650 διαφορετικά links στα τραγούδια
- 57494 διαφορετικοί στίχοι τραγουδιών



	A	B	C
1	Artist	Song	Lyrics
2	ABBA	Ahe's My	Look at her face, it's a wonderful face And it means something special to me Look at the way that she smiles when she sees me How lucky can one fellow be? She's just my kind of girl, she makes me feel fine Who could ever believe that she could be mine? She's just my kind o
3	Foo Fighte	Freaks By	If the mother goes to sleep with you Will you run and tell geraldo If the mother bears your children without tears Without the usual costs of labor If the mother goes to bed with you Will you run and tell the neighbors Will you hide behind that get up that you wear Or will tak
4	Arrogant \	Big Fat Ro	Every rock band has this guy Not many people really know why He's got a cell phone and a backstage pass He's got a big gut and a big fat ass He's a big fat road manager Big fat road manager Big fat road manager Big fat road manager He plugs in guitars and microphone jacks (Big
5	Kate Bush	Where An	Destiny took me in her arms and told me: "You're a little lost on the fall. You fall for all the men You shouldn't fall for all at all. " Ooh, I know she knows, but still I go on, Well, I'm a fool. Climbing up the ladders To slide down the adders On the rocks, On the rocks, On the roc
6	Michael W	Forever	Give thanks to the Lord, Our God and King His love endures forever For He is good, He is above all things His love endures forever Sing praise, sing praise With a might hand and outstretched arm His love endures forever For the life that's been reborn His love endures foreve
7	Kanye We	Barry Bon	It's what you all been waiting for ain't it? What people pay paper for damn it, they can't stand it They want something new, so let's get reacquainted Became the hood favorite, I can't even explain it, I surprise myself too Life of a Don, lights keep glowing Coming in the club with t
8	Genesis	Ripples	Blue girls come in every size Some are wise and some otherwise They got pretty blue eyes For an hour a man may change For an hour her face looks strange Looks strange, looks strange Marching to the promised land Where the honey flows and takes you by the hand Pulls y
9	Randy Tra	I'll Fly Aw	Some glad morning when this life is o'er I'll fly away To a home on God's celestial shore I'll fly away. I'll fly away, oh glory, I'll fly away When I die, hallelujah, by and by I'll fly away. When the shadows of this life grow have grown I'll fly away Like a bird from prison bars has
10	Usher	Just A Frie	I wanna know your name and And I wanna know if u gotta man I wanna know I wanna know everything I wanna know your number and if I can come over and I wanna know what u like I wanna know so I could do it all night But u tellin' me I'm just a friend Steding tellin' me I'm

Εικόνα 1

Από όλα τα δεδομένα της παραπάνω συλλογής, επιλέχθηκαν τυχαία 1000 γραμμές απο τα πεδία “artist”, “song”, “text” το οποίο μετονομάστηκε σε “lyrics”. Η στήλη "link" αφαιρέθηκε. Από τα περιεχόμενα κάθε γραμμής του πεδίου "lyrics", αφαιρέσαμε το χαρακτήρα αλλαγή γραμμής "\n" για να είναι πιο εύκολη η εκπόνηση των fields (artist, song, lyrics)(βλέπε εικόνα 1). Οι συγκεκριμένες ενέργειες έγιναν με τη βοήθεια python scripts.

Ενότητα 3 : Ανάλυση κειμένου και κατασκευή ευρετηρίου:

Το αρχείο CSV, το οποίο αποτελεί την πηγή δεδομένων, αναμένεται να έχει τρεις στήλες: όνομα καλλιτέχνη, όνομα τραγουδιού και στίχους τραγουδιού. Το σύστημα διαβάζει το αρχείο CSV γραμμή προς γραμμή και αναλύει αυτά τα τρία πεδία. Κάθε τραγούδι αντιμετωπίζεται ως έγγραφο και το αντίστοιχο όνομα καλλιτέχνη, το όνομα τραγουδιού και οι στίχοι ευρετηριάζονται στο Lucene χρησιμοποιώντας έναν StandardAnalyzer.

Για την αποθήκευση του ευρετηρίου στο δίσκο θα χρησιμοποιηθεί η συνάρτηση FSDirectory.open() του Lucene API org.apache.lucene.store. Με την κλάση IndexWriter() του package org.apache.lucene.index θα δημιουργηθεί ένα καινούργιο **index** στο οποίο θα αποθηκεύθούν τα Documents αντικείμενα.

Πιο συγκεκριμένα, το πεδίο "song" προστίθεται στο έγγραφο ως πεδίο TextField. Αυτό το πεδίο αποθηκεύει τις πληροφορίες για το τραγούδι και επιτρέπει την αναζήτηση πλήρους κειμένου. Επιπλέον, το πεδίο "song" προστίθεται επίσης ως πεδίο SortedDocValuesField. Αυτό το πεδίο επιτρέπει λειτουργίες ταξινόμησης, παρέχοντας αποτελεσματικές δυνατότητες ταξινόμησης και ομαδοποίησης. Η ίδια διαδικασία ακολουθείται για το πεδίο "artist" και το πεδίο "lyrics".

Για την ανάλυση των πεδίων θα χρησιμοποιηθεί ο StandardAnalyzer() ο οποίος διαχωρίζει το κείμενο σε μεμονωμένους όρους, χρησιμοποιώντας ένα tokenizer που διαχωρίζει τα whitespace και τα σημεία στίξης, μετατρέπει όλους τους όρους σε lowercase, αφαιρεί τα stop words και παρέχει τη δυνατότητα του stemming. Αποτελεί μια λογική επιλογή όταν δεν γνωρίζουμε τα χαρακτηριστικά του κειμένου μας. Το API που χρησιμοποιήθηκε για αυτή την περίπτωση είναι το “org.apache.lucene.analysis”.

Το ευρετήριο κατασκευάζεται χρησιμοποιώντας τρία πεδία (καλλιτέχνης, τραγούδι, στίχοι), καθένα από τα οποία είναι ένα TextField. Τα TextFields τόσο αναλύονται όσο και ευρετηριάζονται, πράγμα που σημαίνει ότι πριν από την ευρετηρίαση, γίνεται το tokenized και επίσης αποθηκεύονται στο ευρετήριο για ανάκτηση. Για την δημιουργία του **Document** θα χρησιμοποιήθεί το package της Lucene: org.apache.lucene.document και ύστερα θα προσθεθεί στο ευρετήριο.

Ενότητα 4 : Αναζήτηση

Για τα ερωτήματα αναζήτησης, το σύστημα υποστηρίζει αναζήτηση με λέξεις-κλειδιά και αναζήτηση με βάση τα πεδία. Ο χρήστης μπορεί να κάνει αναζήτηση με λέξεις-κλειδιά σε όλα τα πεδία ή να περιορίσει την αναζήτηση σε ένα συγκεκριμένο πεδίο (π.χ. καλλιτέχνης, τραγούδι, στίχοι).

Ένας MultiFieldQueryParser χρησιμοποιείται για την ανάλυση του ερωτήματος σε πολλαπλά πεδία. Αυτό επιτρέπει στο σύστημα να αναζητήσει το κείμενο του ερωτήματος σε όλα τα καθορισμένα πεδία του εγγράφου και χρησιμοποιεί το πακέτο “org.apache.lucene.queryparser.classic” της Lucene. Τα αποτελέσματα επιστρέφονται ανά δεκάδες με βάση τη συνάφειά τους με το ερώτημα χρησιμοποιώντας τον searcher της “org.apache.lucene.search”. Η συνάφεια (ή η βαθμολογία) υπολογίζεται με βάση διάφορους παράγοντες, όπως η συχνότητα όρων, η αντίστροφη συχνότητα εγγράφων, το μήκος πεδίου κ.λπ. σύμφωνα με τον προεπιλεγμένο αλγόριθμο ομοιότητας που χρησιμοποιεί το Lucene.

Ακόμα, το σύστημα αναζήτησης διατηρεί ιστορικό των ερωτημάτων του χρήστη. Κάθε νέο ερώτημα προστίθεται σε ένα ευρετήριο προηγούμενων ερωτημάτων. Και αυτό το ευρετήριο αποθηκεύεται στο δίσκο μέσω των API’s που συζητήθηκαν στην «Ενότητα 3». Εάν ένα ερώτημα υπάρχει ήδη στο ευρετήριο, δεν προστίθεται ξανά, διασφαλίζοντας ότι το ιστορικό περιέχει μοναδικά ερωτήματα αναζήτησης. Αυτό το ευρετήριο ιστορικού χρησιμοποιείται για να παρέχει στους χρήστες προτάσεις αναζήτησης. Όταν ένας χρήστης εισάγει ένα μερικό ερώτημα (δηλ. όταν έχει πληκτρολογήσει μερικά

γράμματα του query που θέλει να αναζητήσει), το σύστημα χρησιμοποιεί την PrefixQuery για να βρει προηγούμενα ερωτήματα που ξεκίνησαν με το ίδιο κείμενο επιστρέφοντας το πολύ 5 αποτελέσματα σε κάθε αναζήτηση στο ευρετήριο. Αυτά τα προηγούμενα ερωτήματα επιστρέφονται ως προτάσεις στον χρήστη. Αυτή δυνατότητα παρέχεται μέσω του "org.apache.lucene.search.PrefixQuery" της Lucene. Έτσι, ο χρήστης μπορεί να αναζητήσει παλαιότερα queries.

Ενότητα 5 : Παρουσίαση αποτελεσμάτων

Για την παρουσίαση των αποτελεσμάτων χρησιμοποιήσαμε JavaFX. Συγκεκριμένα τα αποτελέσματα παρουσιάζονται σε ένα περιβάλλον που μοιάζει με ιστοσελίδα (WebView) με τη βοήθεια ετικετών HTML. Κάθε αποτέλεσμα αναζήτησης εμφανίζεται ως συνδυασμός του ονόματος του καλλιτέχνη, του τίτλου του τραγουδιού και των στίχων. Κάθε συναφές έγγραφο ενός αποτελέσματος αναζήτησης παρουσιάζεται σε ξεχωριστές ετικέτες παραγράφου, με το όνομα του καλλιτέχνη, τον τίτλο του τραγουδιού και τους στίχους να επισημαίνονται ανάλογα. Τα αποτελέσματα αναζήτησης είναι σελιδοποιημένα μέσω μιας μεταβλητής resultsPerPage, η οποία αρχικοποιείται στο 10. Αυτό σημαίνει ότι μόνο 10 αποτελέσματα εμφανίζονται σε μια σελίδα κάθε φορά.

Ο χρήστης μπορεί να επιλέξει να κάνει αναζήτηση σε συγκεκριμένα πεδία - Artist, Song ή Lyrics- χρησιμοποιώντας check boxes. Εάν επιλεγούν πολλαπλά πλαίσια ελέγχου, η αναζήτηση θα εκτελεστεί σε όλα αυτά τα πεδία με την ίδια λογική που αναφέρθηκε στην «Ενότητα 4» μέσω της MultiFieldQueryParser. Όταν εκτελείται μια αναζήτηση, το ερώτημα αναζήτησης προστίθεται στο ιστορικό αναζήτησης. Εάν τα ευρετήρια δεν υπάρχουν κατά την εκκίνηση της εφαρμογής, δημιουργούνται.

Ακόμα, χρησιμοποιείται ξανά η MultiFieldQueryParser (όπως στην «Ενότητα 4»), κατά την διαδικασία της εναλλαγής των σελίδων μεταξύ των αποτελεσμάτων αναζήτησης καθώς σε κάθε σελίδα επιστρέφουμε μόνο 10 αποτελέσματα. Στην συγκεκριμένη περίπτωση, χρησιμοποιείται, ακόμα, η searchAfter της "org.apache.lucene.search" η οποία επιστρέφει τα επόμενα 10 πιο συναφή έγγραφα βάση του κειμένου της αναζήτησης. Η σελιδοποίηση παρέχεται με τα κουμπιά "Previous" και "Next". Η εφαρμογή παρακολουθεί την τρέχουσα σελίδα και τα κουμπιά αυτά ενεργοποιούνται ή απενεργοποιούνται ανάλογα με τη σελίδα στην οποία βρίσκεται ο χρήστης.

Ο χρήστης έχει επίσης τη δυνατότητα να ταξινομήσει τα αποτελέσματα της αναζήτησης με αλφαβητική σειρά, εδώ λοιπόν φαίνεται και η δυνατότητα ομαδοποίησης. Αυτό επιτυγχάνεται ενεργοποιώντας ένα πλαίσιο ελέγχου με τίτλο " Alphabetical Grouping". Όταν επιλεγεί, τα αποτελέσματα αναζήτησης θα παρουσιάζονται με αλφαβητική σειρά με βάση τα πεδία "καλλιτέχνης" και "τραγούδι", δίνοντας στους χρήστες μια τακτοποιημένη, ευανάγνωστη λίστα. Αυτό είναι ιδιαίτερα επωφελές για μεγαλύτερα σύνολα αποτελεσμάτων, καθώς απλοποιεί τον εντοπισμό συγκεκριμένων στοιχείων ενδιαφέροντος. Αυτή η λειτουργία ενεργοποιείται με τη χρήση της κλάσης "Sort" της "org.apache.lucene.search", η οποία επιτρέπει την ταξινόμηση των αποτελεσμάτων αναζήτησης σύμφωνα με διάφορα πεδία. Ένα αντικείμενο "Sort" δημιουργείται με αντικείμενα "SortField" που καθορίζουν τα πεδία 'artist' και 'song', και τον τύπο τους ως STRING. Όταν αυτό το αντικείμενο "sort" περάσει στη μέθοδο αναζήτησης του αντικειμένου IndexSearcher, τα επιστρεφόμενα αποτελέσματα θα ταξινομηθούν σύμφωνα με τα πεδία που έχουν καθοριστεί (αν αυτά έχουν επιλεχθεί από τον χρήστη).

Για την προβολή των προτάσεων εναλλακτικών ερωτημάτων, οι προτάσεις αντλούνται από το ευρετήριο ιστορικού αναζήτησης 1500 χιλιοστά του δευτερολέπτου (1,5 δευτερόλεπτο) αφού ο χρήστης σταματήσει να πληκτρολογεί στο πλαίσιο αναζήτησης. Τα αποτελέσματα, εμφανίζονται κάτω από την μπάρα αναζήτησης.

Εμφάνιση GUI.

Η μέθοδος updateResultArea(DocumentsSearcher.SearchResult result) έχει σχεδιαστεί για να εμφανίζει τα αποτελέσματα της αναζήτησης σε μια διεπαφή χρήστη βασισμένη στο διαδίκτυο. Λειτουργεί με την κατασκευή μιας συμβολοσειράς HTML που αναπαριστά τα αποτελέσματα αναζήτησης και τα εμφανίζει σε ευανάγνωστη και διαδραστική μορφή.

Αρχικά, πραγματοποιεί επανάληψη της λίστας των αντικειμένων Document στα αποτελέσματα αναζήτησης. Για κάθε έγγραφο, εξάγει τα πεδία καλλιτέχνης, τραγούδι και στίχοι. Εάν το μήκος των στίχων είναι μεγαλύτερο από 200 χαρακτήρες, περικόπτει τους στίχους σε μια προεπισκόπηση 200 χαρακτήρων για συμπαγή εμφάνιση. Στη συνέχεια, η μέθοδος υπογραμμίζει όλες τις εμφανίσεις των λέξεων-κλειδιών αναζήτησης στην προεπισκόπηση των στίχων, περικλείοντάς τες με ετικέτες HTML span που ορίζουν το χρώμα φόντου σε κίτρινο.

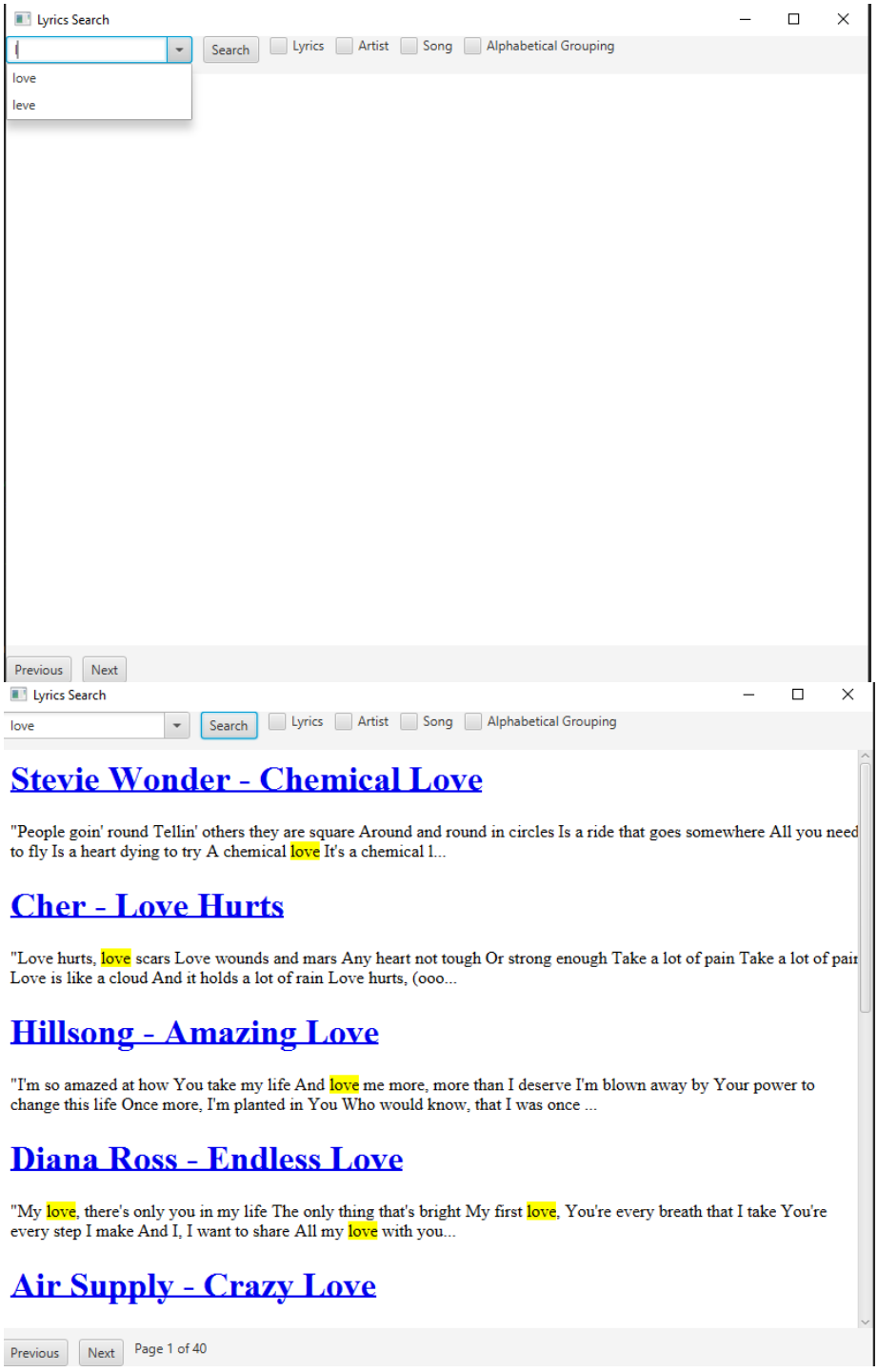
Στη συνέχεια, δημιουργεί μια κεφαλίδα HTML (H1) για κάθε αποτέλεσμα που περιέχει τον καλλιτέχνη και τον τίτλο του τραγουδιού. Κάνοντας κλικ σε αυτήν την επικεφαλίδα ενεργοποιείται μια συνάρτηση JavaScript που εμφανίζει τους πλήρεις στίχους του τραγουδιού. Η προεπισκόπηση των στίχων εμφανίζεται κάτω από την επικεφαλίδα. Η συνάρτηση JavaScript showFullLyrics υλοποιείται στο τέλος της συμβολοσειράς HTML.

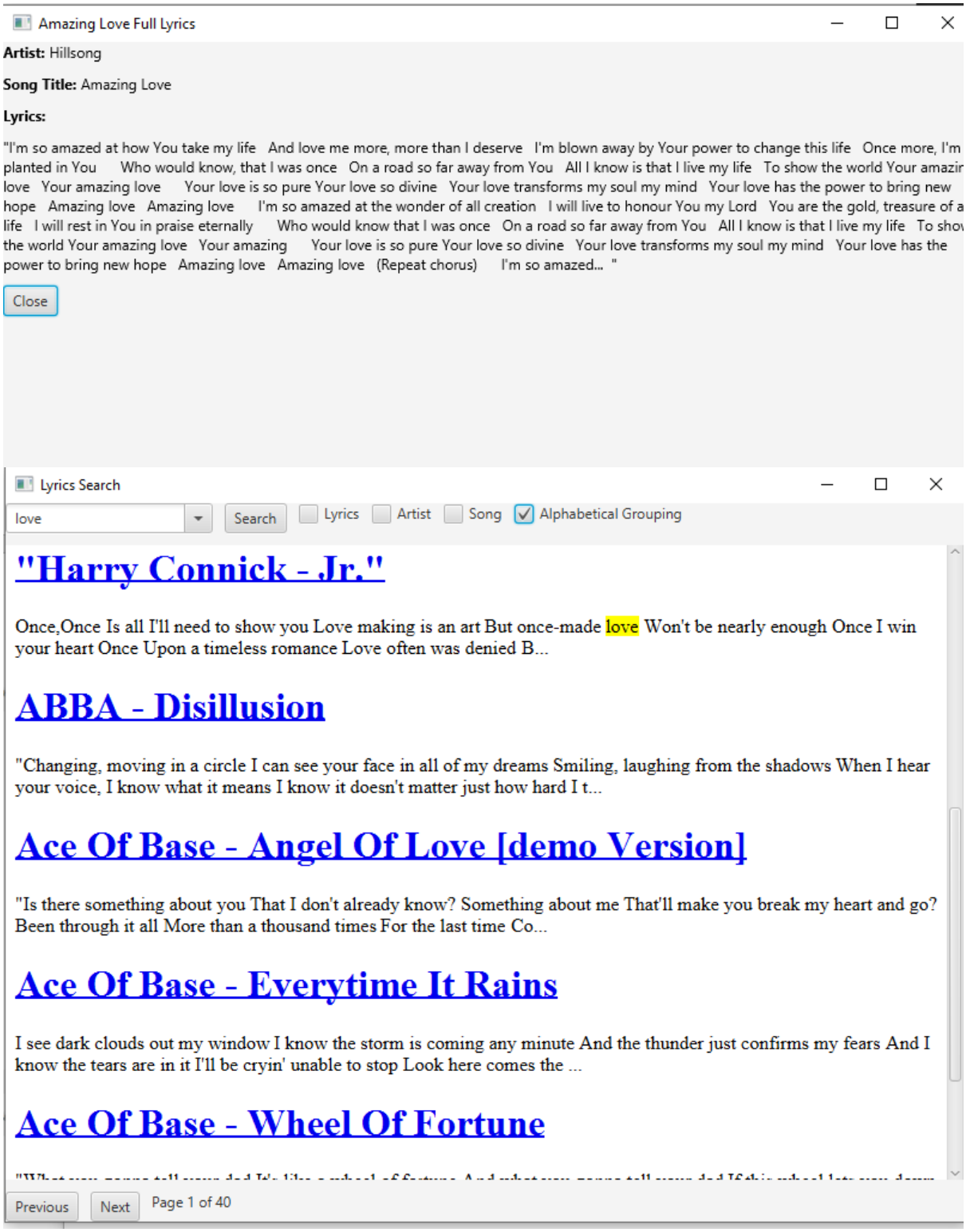
Η συμβολοσειρά HTML φορτώνεται σε ένα αντικείμενο WebEngine, το οποίο αποτελεί μέρος της υλοποίησης του προγράμματος περιήγησης JavaFX WebKit. Η JavaScript είναι ενεργοποιημένη στη μηχανή ιστού, ώστε η συνάρτηση showFullLyrics να μπορεί να αλληλεπιδράσει με τον κώδικα Java.

Ένα αντικείμενο JavaBridge δημιουργείται και προστίθεται στο JavaScript της μηχανής ιστού μετά την επιτυχή φόρτωση του περιεχομένου HTML. Αυτή η γέφυρα επιτρέπει στον κώδικα JavaScript να καλεί μεθόδους Java. Η συνάρτηση showFullLyrics JavaScript χρησιμοποιεί αυτή τη γέφυρα για να καλέσει τη μέθοδο JavaBridge.showFullLyrics της Java.

Η μέθοδος JavaBridge.showFullLyrics έχει σχεδιαστεί για να εμφανίζει τους πλήρεις στίχους ενός τραγουδιού σε ένα νέο παράθυρο. Ψάχνει στη λίστα των εγγράφων στα αποτελέσματα αναζήτησης για ένα έγγραφο που ταιριάζει με τον τίτλο του τραγουδιού. Μόλις βρεθεί, ανακτά το όνομα του καλλιτέχνη και τους πλήρεις στίχους από το έγγραφο. Στη συνέχεια δημιουργεί ένα νέο (παράθυρο) JavaFX που περιέχει ετικέτες και πεδία κειμένου για το όνομα του καλλιτέχνη, τον τίτλο του τραγουδιού και τους πλήρεις στίχους. Παρέχεται επίσης ένα κουμπί κλεισίματος για να μπορεί ο χρήστης να κλείσει το παράθυρο. Η διάταξη του παραθύρου οργανώνεται χρησιμοποιώντας ένα παράθυρο διάταξης VBox που διατάσσει τους κόμβους-παιδιά του σε μία μόνο κατακόρυφη στήλη.

Ακολουθεί παράδειγμα αναζήτησης με εικόνες:


Lyrics Search
Search
Lyrics
Artist
Song
Alphabetical Grouping
love
leve
Previous
Next
Lyrics Search
love
Search
Lyrics
Artist
Song
Alphabetical Grouping
Stevie Wonder - Chemical Love
"People goin' round Tellin' others they are square Around and round in circles Is a ride that goes somewhere All you need to fly Is a heart dying to try A chemical love It's a chemical l...
Cher - Love Hurts
"Love hurts, love scars Love wounds and mars Any heart not tough Or strong enough Take a lot of pain Take a lot of pain Love is like a cloud And it holds a lot of rain Love hurts, (ooo...
Hillsong - Amazing Love
"I'm so amazed at how You take my life And love me more, more than I deserve I'm blown away by Your power to change this life Once more, I'm planted in You Who would know, that I was once ...
Diana Ross - Endless Love
"My love, there's only you in my life The only thing that's bright My first love, You're every breath that I take You're every step I make And I, I want to share All my love with you...
Air Supply - Crazy Love
Previous
Next
Page 1 of 40


Amazing Love Full Lyrics
Artist: Hillsong
Song Title: Amazing Love
Lyrics:
"I'm so amazed at how You take my life And love me more, more than I deserve I'm blown away by Your power to change this life Once more, I'm planted in You Who would know, that I was once On a road so far away from You All I know is that I live my life To show the world Your amazing love Your amazing love Your love is so pure Your love so divine Your love transforms my soul my mind Your love has the power to bring new hope Amazing love Amazing love I'm so amazed at the wonder of all creation I will live to honour You my Lord You are the gold, treasure of a life I will rest in You in praise eternally Who would know that I was once On a road so far away from You All I know is that I live my life To show the world Your amazing love Your amazing Your love is so pure Your love so divine Your love transforms my soul my mind Your love has the power to bring new hope Amazing love Amazing love (Repeat chorus) I'm so amazed... "
Close
Lyrics Search
love
Search
Lyrics
Artist
Song
Alphabetical Grouping
"Harry Connick - Jr."
Once,Once Is all I'll need to show you Love making is an art But once-made love Won't be nearly enough Once I win your heart Once Upon a timeless romance Love often was denied B...
ABBA - Disillusion
"Changing, moving in a circle I can see your face in all of my dreams Smiling, laughing from the shadows When I hear your voice, I know what it means I know it doesn't matter just how hard I t...
Ace Of Base - Angel Of Love [demo Version]
"Is there something about you That I don't already know? Something about me That'll make you break my heart and go? Been through it all More than a thousand times For the last time Co...
Ace Of Base - Everytime It Rains
I see dark clouds out my window I know the storm is coming any minute And the thunder just confirms my fears And I know the tears are in it I'll be cryin' unable to stop Look here comes the ...
Ace Of Base - Wheel Of Fortune
Previous
Next
Page 1 of 40