

Lucene Search Engine



Ιωάννης Τσιρώνης 4908

Κωνσταντίνος Ανδρέου 4316

[GitHub](#)

[YouTube](#)

Εισαγωγή

Στόχος της εργασίας είναι να δημιουργήσουμε μια εφαρμογή που επιτρέπει την αναζήτηση τραγουδιών από ένα ευρετήριο, καθώς και σημασιολογική αναζήτηση χρησιμοποιώντας ένα Google embedding.

Η εφαρμογή αποτελείται από μία σειρά από services, γραμμένα σε Java και Python που δουλεύουν μεταξύ τους για να προσφέρουν στον χρήστη ένα πλήθος λειτουργιών σε ένα εύχρηστο περιβάλλον.

Lucene Search Service

Η κύρια υπηρεσία αναζήτησης υλοποιείται χρησιμοποιώντας Apache Lucene και είναι υπεύθυνη για την αναζήτηση αποτελεσμάτων στο ευρετήριο.

Create Index Service

Η υπηρεσία δημιουργίας ευρετηρίου δημιουργεί ένα ευρετήριο όπου εισάγει τα δεδομένα τραγουδιών για τη γρήγορη αναζήτηση τους.

Semantic Search Service

Η υπηρεσία σημασιολογικής αναζήτησης υλοποιείται σε Python χρησιμοποιώντας το Google Word2Vec embedding, εκπαιδευμένο σε πάνω 100 δισεκατομμύρια λέξεις από το Google News. Δίνει τη δυνατότητα εμφάνισης αποτελεσμάτων τα οποία παρότι δεν περιέχουν τις ίδιες λέξεις με αυτές του ερωτήματος, έχουν παρόμοια σημασία.

Backend

Η βάση που οργανώνονται όλα αυτά τα services είναι το Back-End API γραμμένο σε Django(Python web framework).

Το API λαμβάνει τα αιτήματα του χρήστη από το Front-End και καλεί τα κατάλληλα services για να τα ικανοποιήσει.

Frontend

Front-End υλοποιείται με Django-Templates και με ένα εύχρηστο τρόπο προσφέρει στο χρήστη όλες τις υπηρεσίες της Εφαρμογής.

Υπηρεσίες της Lucene Search Engine

- Υπηρεσία Αναζήτησης(Lucene Search Service)
- Υπηρεσία Σημασιολογικής Αναζήτησης(Semantic Search Service)
- Ιστορικό Αναζητήσεων
- Πρόταση Αναζήτησης με βάση το Ιστορικό Αναζήτησης(History based recommendation),
- Οπτικοποίηση των αποτελεσμάτων
- Έμφαση στους ορούς της αναζήτησης που βρίσκονται στα αποτελέσματα
- Αλλαγή σελίδας(μπρος και πίσω)
- Υπηρεσία Ομαδοποίησης

Συλλογή και Προ επεξεργασία αρχείων

Η συλλογή τραγουδιών αποτελείται από 57,650 τραγούδια από το `spotify_millsong_data.csv` dataset που βρίσκεται στο Kaggle. Κάθε τραγούδι περιγράφεται από τον καλλιτέχνη, τον υπερ-σύνδεσμο του, τον τίτλο και τους στίχους του.

Στο στάδιο της προ επεξεργασίας των δεδομένων:

- αφαιρέθηκε το πεδίο του υπερ-συνδέσμου του
- αφαιρέθηκαν οι χαρακτήρες `\n`, `\r` και `\t` από όλα τα πεδία
- αφαιρέθηκαν τα τραγούδια που έχουν ακριβώς ίδιους στίχους με άλλα τραγούδια και αυτά που έχουν ίδιο καλλιτέχνη με ίδιο τίτλο
- δημιουργήθηκε ένα αρχείο `clean_songs.tsv` που θα χρησιμοποιηθεί για τη δημιουργία του ευρετηρίου.

Create Index Service

Για την υπηρεσία δημιουργίας ευρετηρίου χρησιμοποιήθηκε η κλάση `CreateIndex`, η οποία αρχικά φτιάχνει το `directory` στο οποίο θα αποθηκευτεί το ευρετήριο, στον δίσκο και στην συνέχεια προσθέτει την συλλογή τραγουδιών στο ευρετήριο(Ευρετηριοποίηση).

Lucene Search Service

Για την κύρια υπηρεσία αναζήτησης χρησιμοποιείται την κλάση LuceneModule, η οποία δέχεται ως command-line arguments, το ερώτημα προς αναζήτηση, το πεδίο αναζήτησης και τον αριθμό της σελίδας εμφάνισης των αποτελεσμάτων.

Στη συνέχεια, διαβάζεται το ευρετήριο και αναζητούνται τα πιο συναφή τραγούδια ως προς το ερώτημα αναζήτησης. Οι λέξεις της ερώτησης που εμφανίζονται στα τραγούδια παρουσιάζονται με bold.

Τέλος, τα αποτελέσματα αντιστοιχούν στον αριθμό της σελίδας μεταφέρονται στο standard output.

Το Lucene Search Service εξάγεται σε ένα jar αρχείο(search_app.jar), για να μπορεί να κληθεί από το Backend.

Backend

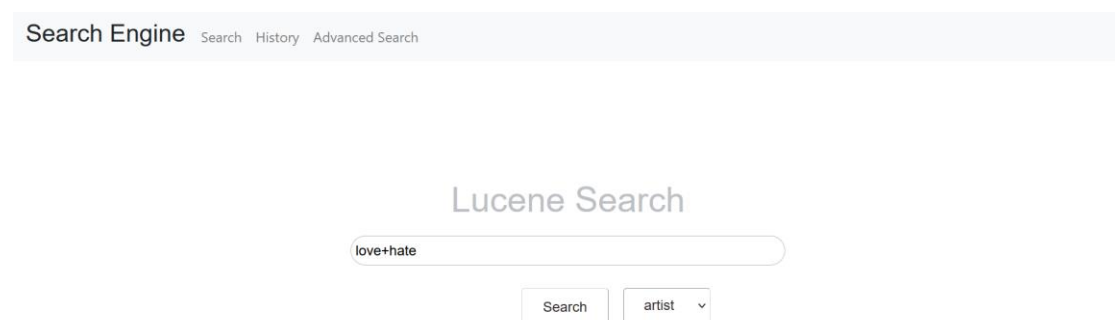
Search controller

Όταν ο χρήστης κάνει αίτημα για αναζήτηση(Εικόνα 1), ο search controller καλεί το search service μεταφέροντας του το αίτημα του χρήστη.

Το search service καλεί το Lucene Search Service για να κάνει την αναζήτηση, μεταφέροντας τα αποτελέσματα της, στο process_output service για να τα επεξεργαστεί.

Το process_output service επιστρέφει τα αποτελέσματα σε μορφή pandas DataFrame στον search controller, ο οποίος με τη σειρά του τα μετατρέπει σε JSON και τα εμφανίζει στον χρήστη(Εικόνα 2).

Αφότου τα αποτελέσματα εμφανιστούν, ο χρήστης έχει στη διάθεση του και άλλες υπηρεσίες.



Εικόνα 1. Σελίδα Lucene Αναζήτησης.

Search Engine Search History Advanced Search			
Results			
<div>Group By Lyrics Length</div> <div>Current Page: 1</div> <div> <div></div> <div>1</div> </div>			
Rank	Artist	Title	Lyrics
1	Lloyd Cole	I Hate To See You Baby Doing That Stuff	I love to see you in your sweater girl I love you in your high tall boots I love to see you in your sweater girl Love to see you walking that cute, you know it I love to see you in your leather skin Love your lipstick on my sheets And I love to see you in your alligator shoes Reciting to me my poetry, I really do but I hate to see you baby doing that stuff I hate to see you doing that stuff I hate to see you baby doing that stuff Really hate to see you doing that stuff I love you baby when you bite my ear I don't worry about your uptown geek And I love you baby don't you bite your lip 'Cause it's me you talk to in your sleep and I hate to see you baby doing that stuff I hate to see you doing that stuff I hate to see you baby doing that stuff Really hate to see you doing that stuff I love to see you in your sweater girl Love your alligator chic Et je sais te vois dans ton christian la croix But I hate to see you walking my street, but I hate to see you baby doing that stuff I hate to see you doing that stuff I hate to see you baby doing that stuff Really hate to see you doing that stuff, ha ha
2	Britney Spears	Love To Hate You	Love to hate you I'm doing fine I'm doing fine For seven months and fifteen days I've been telling you lies in different ways I know things will change But I don't know when You keep talking about us Like we still were friends I just hate the way you talk to me I just love to say what's bothering me I love to hate you I really love to hate you And nothing you do Will ever change my mind I'm doing fine Without you it wasn't meant to be like this All you gave her was one little innocent kiss I didn't believe She would turn to be The love of your life Just like me I just hate the way you talk to me I just love to say what you mean to me I love to hate you I really love to hate you And nothing you do Will ever change my mind I'm doing fine Without you I'm doing fine I love to hate you I really love to hate you And nothing you do Will ever change my mind I'm doing fine I love to hate you I really love to hate you And nothing you do Will ever change my mind I'm doing fine Without you
3	Christina Aguilera	i hate boys	No-no-no-no, I'm not bitter, I'm not mad. Well, maybe just a little, just a tad. I know every apple here ain't bad, But I found a worm in every single

Εικόνα 2. Σελίδα εμφάνισης των αποτελεσμάτων

Group by length controller

Όταν ο χρήστης κάνει αίτημα για ομαδοποίηση των αποτελεσμάτων με βάση το μέγεθος των στίχων τους, ο group by length controller καλεί το group_by_len service το οποίο επιστρέφει τα ομαδοποιημένα αποτελέσματα στον controller για να εμφανιστούν στον χρήστη.

Next/Previous Ten controller

Όταν ο χρήστης θέλει να μεταφερθεί στην επόμενη σελίδα αναζήτησης, ο next_ten controller καλεί το search service, μεταφέροντας του, το αίτημα του χρήστη μαζί με τον αριθμό της επομένης σελίδας. Τα τραγούδια της επομένης σελίδας εμφανίζονται στο χρήστη.

Ομοίως, ο previous_ten επιστρέφει τα τραγούδια της προηγούμενης σελίδας. (Εικόνα 3).

View History controller

Κάθε φορά που ο χρήστης κάνει αίτημα αναζήτησης, αυτό αποθηκεύεται σε μια SQLite3 Βάση Δεδομένων. Όταν ο χρήστης θελήσει να δει το ιστορικό του, ο view_history controller εμφανίζει όλες τις αναζητήσεις που έχει κάνει(Εικόνα 5).

Ο χρήστης έχει τη δυνατότητα να μεταφερθεί στα αποτελέσματα της κάθε αναζήτησης επιλέγοντας αυτή.

Επίσης, έχει τη δυνατότητα να επιλέξει την Πρόταση Αναζήτησης, η οποία είναι φτιαγμένη από τις πιο σύνηθες λέξεις και πεδία αναζήτησης που έχει στο ιστορικό του.

Search Engine <small>Search History Advanced Search</small>			
Results			
<small>Group By Lyrics Length</small> <small>Current Page: 2</small>			
Rank	Artist	Title	Lyrics
11	Cyndi Lauper	A Part Hate	Somber sister This is a strange and bitter fruit Because you taught me to sing And the rhythm in my heart And the rhythm in my feet is - Why are the rainbows Stolen from the sky And locked up in boxes Yellow, black, red and white Like birds in their cages Beating their wings on the bars And there's a song that they're singing It's a word in the world It's a word in their hearts A part hate I heard a man say Tear apart hate And I saw hope in his face A part hate Where the color of love Slips away Why are the children Carrying guns, not books Drug dealing, not learning The golden rule And the idea of freedom Not just the same Castle in the sky Haunted by white-sheeted ghouls Filled with hate me And hate you And proud of it too A part hate Heard a woman saying Tear apart hate And I saw hope in her face A part hate Where the color of love Slips away Why are the people Running down the block Rock throwing, not knowing What else to do But I'm just a spectator And I can never know the pain But when I hear That whip cracking I cry out tears of anger I cry out tears of shame A part hate I heard myself say Tear apart hate And I saw hope in my face A part hate Where the color of love Slips away... Tear apart hate tear apart hate Tear apart hate tear apart hate tear apart hate ...
12	Alice In Chains	Love, Hate, Love	I tried to love you I thought I could I tried to own you I thought I would I want to peel the skin from your face Before the real you lays to waste You told me I'm the only one Sweet little angel you should have run Lying, crying, dying to leave Innocence creates my hell Cheating myself still you know more It would be so easy with a whore Try to understand me little girl My twisted passion to be your world Lost inside my sick head I live for you but I'm not alive Take my hand before I kill I still love you, but, I still burn Yeah, love, hate, love Yeah, love, hate, love Yeah, love, hate, love Oh, Love, hate, love Yeah, Love, hate, love
13	Ne-Yo	Hate That I Love You	That's how much I love you That's how much I need you And I can't stand you Must everything you do make me wanna smile Can I not like you for awhile? (No) But you won't let me You upset me, girl And then you kiss my lips All of a sudden I forget (that I was upset) Can't remember what you did (But I hate it) You know exactly what to do So that I can't stay mad at you For too long, that's wrong (But I hate it) You know exactly how to touch So that I don't want to fuss and fight no

Εικόνα 3. Εμφάνιση αποτελεσμάτων επομένης σελίδας.

Semantic Search Service (Bonus)

Η σημασιολογική αναζήτηση, δίνει τη δυνατότητα εύρεσης αποτελεσμάτων που παρότι δεν περιέχουν τις ίδιες λέξεις με αυτές του ερωτήματος, έχουν παρόμοια σημασία. Αυτό θα βοηθήσει τον χρήστη να βρει κάποιο τραγούδι χωρίς να θυμάται τις ακριβές λέξεις (Εικόνα 4).

Για την επίτευξη αυτού του προβλήματος χρησιμοποιήθηκε το Word2Vec embedding της Google εκπαιδευμένο σε 100 δισεκατομμύρια λέξεις.

- Οι σίχοι προ-επεξεργάζονται μετατρέποντας τις λέξεις σε πεζά γράμματα και αφαιρώντας τις πολύ μεγάλες και τις πολύ μικρές λέξεις, μέσω της συνάρτησης `preprocess_lyrics`.
- Στη συνέχεια, για να βρεθούν οι διανυσματικές αναπαραστάσεις(δ.α.) των στίχων κάθε τραγουδιού, πρέπει:
 - Για κάθε τραγούδι να βρεθεί η δ.α. κάθε λέξης των στίχων του, με τη χρήση του `embedding`.
 - Από αυτές τις δ.α. υπολογίζεται η μέση δ.α. που αναπαριστά το μέσο διάνυσμα των στίχων του τραγουδιού.
 - Το σύνολο αυτών των δ.α. για κάθε τραγούδι αποθηκεύονται στο αρχείο `encoded_data.npy`.
- Το ερώτημα του χρήστη επεξεργάζεται και κωδικοποιείται χρησιμοποιώντας το `embedding`.
 Στη συνέχεια, υπολογίζεται το cosine distance του διανύσματος της ερώτησης, με κάθε άλλη διανυσματική αναπαράσταση στίχων που προέκυψε από το παραπάνω βήμα.
 Από εκεί επιλέγονται τα διανύσματα των στίχων που έχουν τη μικρότερη απόσταση από το διάνυσμα της ερώτησης και αυτά αντιστοιχίζονται στον πίνακα τραγουδιών.

Τέλος, οι λέξεις της ερώτησης που εμφανίζονται στο αποτέλεσμα παρουσιάζονται με bold (συνάρτηση highlight_words).

4. Το τελικό αποτέλεσμα επιστρέφεται στον advanced_search controller, ο οποίος με τη σειρά του εμφανίζει τα αποτελέσματα στον χρήστη(Εικόνα 5).

Search Engine

SearchHistoryAdvanced Search

Advanced Lyric Search

apple orange

Search

Εικόνα 4. Σελίδα Σημασιολογικής Αναζήτησης

Search Engine

SearchHistoryAdvanced Search

Results

Current Page: 1

<

1

Rank	Artist	Title	Lyrics
	Deep Purple	The Orange Juice Song	Orange juice, Just thinkin bout that orange orange juice Orange juice, Just thinkin bout that orange orange juice Some peeps Say orange juice is yellow But I say Orange juice is mellow With a lil more orange In that orange orange juice Orange juice, Just thinkin bout that orange orange juice, Orange juice, Just thinkin bout that orange orange juice Some peeps, Say orange juice is yucky But I say Orange juice is lucky With a lil more spunk In that orange orange juice Orange juice, Just thinkin bout that orange orange juice, Orange juice, Just thinkin bout that orange orange ju-you-you-you-uice.
	Children	Apples And Oranges	I like apples and oranges . I like apples and oranges . Apples and oranges are so sweet. Apples and oranges are good to eat. I like apples and oranges . Orange juice is so sweet, Apple sauce is fun to eat, Apple pie with ice cream -- what a tasty treat. I like apples and oranges . Apples and oranges are so sweet. Apples and oranges are good to eat. I like apples and oranges . Orange juice is so sweet, Apple sauce is fun to eat, And apple pie with ice cream -- what a tasty treat. I like apples and oranges . I like apples and oranges . Apples and oranges are so sweet. Apples and oranges are good to eat. I like apples and oranges . Orange juice is so sweet, Apple sauce is fun to eat, Apple pie with ice cream -- oooh, what a tasty treat. I like apples and oranges . I like apples and oranges . Apples and oranges (repeat to fade)
	Tori Amos	Datura	Get out of my garden Passion vine Texas sage Indigo spires salvia Confederate jasmine Royal cape plumbago Arica palm Pygmy date palm Snow-on-the-mountain Pink powderpuff Datura Crinum lily St. Christopher's lily Silver dollar eucalyptus White african iris Katie's charm ruellia Variegated shell ginger Florida coontie Datura Ming fern Sword fern Dianella Walking iris Chocolate cherries allamanda Awabuki viburnum Natal plum Black magic ti Mexican bush sage Gumbo limbo Golden shrimp Belize shrimp Senna Weeping sabicu Golden shower tree Golden trumpet tree Bird of paradise Come in Valegated shell ginger Datura Ionicera Red velvet costus Xanadu philodendron Snow queen hibiscus Frangipani Bleeding heart Persian shield Cat's whiskers Royal palm Sweet slyssum Petting bamboo Orange jasmine Clitoria blue pea Downy jasmine Datura Frangipani Dividing Cannaan Piece by piece O let me see

Εικόνα 5. Σελίδα εμφάνισης αποτελεσμάτων Σημασιολογικής Αναζήτησης