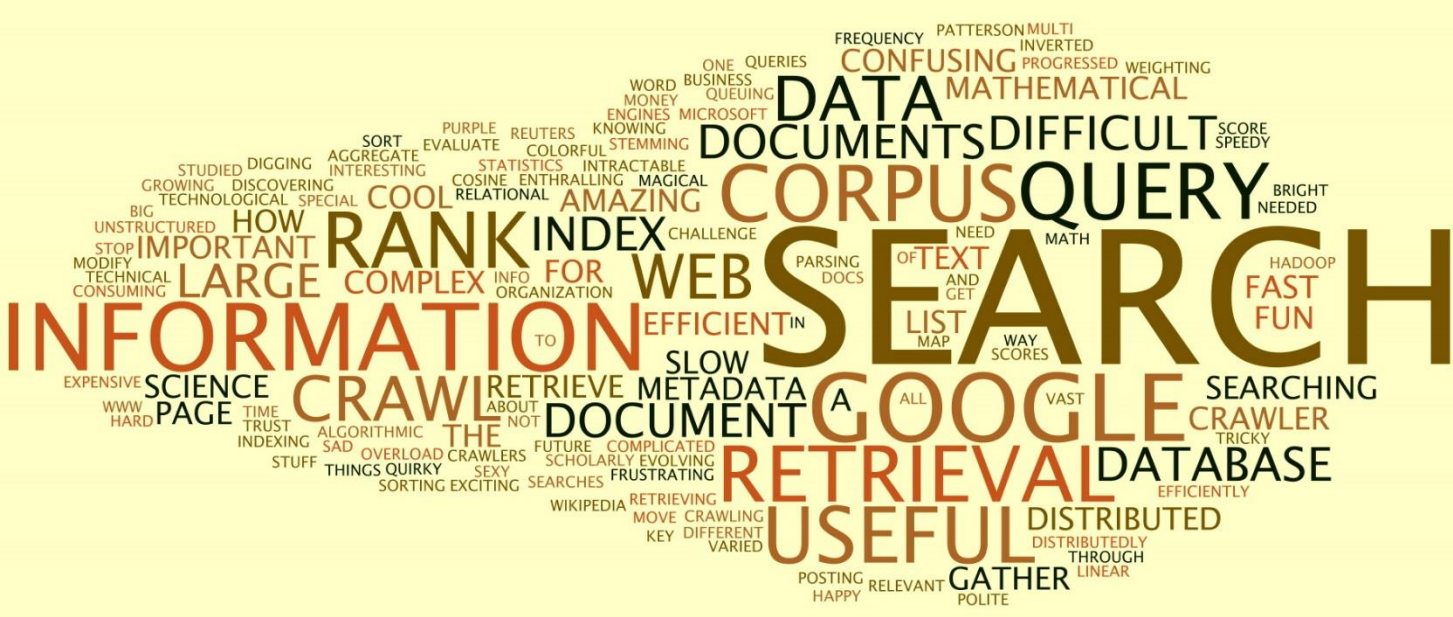

Information Retrieval

PHASE 1

Σταματίνα Λυμπέρη | AM:4410 | Project 2021-2022

ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΣΥΛΛΟΓΗΣ

Η συλλογή(corpus) που επιλέχθηκε να χρησιμοποιηθεί για την συγκεκριμένη λειτουργία της μηχανής είναι κομμάτι από την έτοιμη συλλογή της ιστοσελίδας Kaggle, η οποία αποτελείται από data διαφόρων shows της πλατφόρμας Netflix.

Η αρχική συλλογή (τύπου .csv) αποτελούνταν από 8.807 ταινίες και σειρές, η οποία ενημερώνεται κάθε μήνα και αποτελούνταν από 12 στήλες. Στην συγκεκριμένη χρονική στιγμή κρατήσαμε το αρχείο που υπήρχε μέχρι και τον μήνα Μάρτιο. Αρχικά κρατήθηκε ένας αριθμός ταινιών και σειρών του ύψους 800, ο οποίος μετατράπηκε σε .txt αρχεία για κάθε ένα show με τον παρακάτω κώδικα σε γλώσσα Python.

```
1. import csv
2.
3. fileName = "netflix_titles"
4. numberOfRows = 800
5. counterForRows = 0
6. with open('netflix_titles.csv', encoding="utf8") as csv_file:
7.     csv_reader = csv.reader(csv_file)
8.     for row in csv_reader:
9.         if counterForRows == 0:
10.            print(f'Columns in file are {", ".join(row)}')
11.            counterForRows += 1
12.        else:
13.            text_file = open(''.join([fileName, str(counterForRows),
   '.txt']), 'w', encoding="utf8")
14.            text_file.write("Title: "+row[2]+'\n'+"Type: "+row[1]+'\n' +
   "Release year: "+row[7]+'\n'+"Listed in :
   "+row[10]+'\n\n'+"Description:\n"+row[11])
15.            text_file.flush()
16.            text_file.close()
17.            counterForRows += 1
18.
19.        if counterForRows > numberOfRows:
20.            print('txt files created! We all done!')
21.            break
22.    print(f'Processed {numberOfRows} rows.')
```

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΣΧΕΔΙΑΣΜΟΥ ΤΗΣ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ

Σκοπός είναι η δημιουργία μιας μηχανής αναζήτησης , η οποία θα μας επιστρέφει αποτελέσματα σχετικά με τις ταινίες και τις σειρές της πλατφόρμας Netflix μέσω μιας συλλογής που έχει συγκεντρωθεί για την εν λόγο διαδικασία.

Η συγκεκριμένη μηχανή θα υποστηρίζει και λειτουργίες όπως:

- ➔ Αναζήτηση με βάση λέξεις κλειδιά
- ➔ Επιστροφή των αποτελεσμάτων με βάση την συνάφεια που θα υπάρχει με το συγκεκριμένο ερώτημα.

ΚΑΤΑΣΚΕΥΗ ΤΟΥ ΕΥΡΕΤΗΡΙΟΥ ΚΑΙ Η ΑΝΑΛΥΣΗ ΤΟΥ ΚΕΙΜΕΝΟΥ

Γνωρίζουμε ότι οι βασικές έννοιες που χρησιμοποιεί η βιβλιοθήκη Lucene είναι:

- Indexes
- Fields
- Documents
- Terms

Για κάθε αρχείο που έχουμε στην διάθεση μας θα μπορέσουμε να εξάγουμε τα πεδία Title, Description και στην συνέχεια και τα πεδία Type,Release Year, Listed in.

Για να μπορέσουμε να αξιοποιήσουμε καλυτέρα τα δεδομένα μας θα χρησιμοποιήσουμε στο Title και στο Description έναν αναλυτή, και συγκεκριμένα τον Standard Analyzer, Ο οποίος θα μετατρέπει όλες τις λέξεις σε lowercase ,αφαιρεί κοινές λέξεις και τα σημεία στίξης.

Το πεδίο Title και Description θα κρατηθούν ως μια μονάδα token για την αναζήτηση.

Τα αποτελέσματα της παραπάνω διαδικασίας θα αποθηκευτούν για χρήση στην παρακάτω εκτέλεση.

ΑΝΑΖΗΤΗΣΗ

Για να πραγματοποιηθεί η αναζήτηση, ο χρήστης θα θέτει το ερώτημα του στο UI. Στην συνέχεια αυτό θα αναλύεται και θα υλοποιείται στην συνέχεια η αναζήτηση στα Documents που θα έχουμε. Ακόμα το σύστημα μας θα πρέπει να υποστηρίζει και ερωτήσεις της μορφής:

1. Αναζήτηση με λέξεις κλειδιά σε όλα τα πεδία, τα οποία είναι Title,Type,Release Year, Listed in, Description.
2. Αναζήτηση σε ένα συγκεκριμένο πεδίο με μια λέξη κλειδί.

Ακόμα, στόχος αυτής της μηχανής είναι και η χρήση και διατήρηση ενός ιστορικού αναζητήσεων που έχει πραγματοποιήσει ο χρήστης, αλλά και η εμφάνιση μηνυμάτων σε περίπτωση που χρειαστεί να δοθούν αποτελέσματα παρόμοια με αυτά που αναζήτησέ , τύπου " maybe you are looking for .."

ΕΜΦΑΝΙΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Για να επιτευχθεί η σωστή εμφάνιση των αποτελεσμάτων στον χρήστη θα χρησιμοποιηθεί ένα GUI το οποίο:

1. Θα χρησιμοποιεί ένα κουτί αναζήτησης(search box) όπου ο χρήστης θα εισάγει το ερώτημα του .
2. Θα παρουσιάζει τα αποτελέσματα ιο κάθε φορά(με βάση την συνάφεια που θα έχουν)
3. Στα αποτελέσματα θα εμφανίζονται οι λέξεις -κλειδιά τονισμένες
4. Θα δίνεται η δυνατότητα αναδιάταξης των αποτελεσμάτων που θα εμφανίζονται με βάση τον τύπο(type) και το Release Year.

ΣΗΜΕΙΩΣΗ

Η παραπάνω περιγραφή αποτελεί μια προσέγγιση της υλοποίησης και υπάρχει μεγάλη πιθανότητα να διαφοροποιηθεί στην εκτέλεση για να είναι εφικτή η σωστή λειτουργία της μηχανής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. https://www.kaggle.com/datasets/shivamb/netflix-shows

ΓΛΩΣΣΑΡΙ

- UI: User Interface