Στόχος αυτής της φάσης είναι να περιγράψουμε εν συντομία την δημιουργία της συλλογής των δεδομένων που θα χρησιμοποιηθούν για την εκπόνηση της εργασίας μας, την κατανόηση των βασικών βημάτων αυτής και την ανάλυση ενός αρχικού σχεδιασμού.

Η συλλογή από έγγραφα θα έχει τα εξής χαρακτηριστικά:

Τα αρχεία είναι με την μορφή csv και περιέχουν πληροφορίες για ταινίες. Πιο συγκεκριμένα είναι μια συλλογή από 8800 ταινίες και σειρές, οι οποίες προβάλλονται στο Netflix. Την συλλογή αυτήν την κατεβάσαμε από το site (https://www.kaggle.com/datasets/shivamb/netflix-shows), που είναι μια έτοιμη συλλογή με όλα όσα αναγράφονται προηγουμένως.

Περιγραφή σχεδιασμού του συστήματος:

Τα αρχεία, όπως αναφέραμε και πριν, είναι στην μορφή csv. Αυτό το αρχείο έχει πεδία, τα οποία αντιπροσωπεύουν την κάθε ταινία ή σειρά που αναγράφεται. Τα πεδία αυτά είναι show_id(το αναγνωριστικό id της εκάστοτε ταινίας ή σειράς), type(ο τύπος, ταινία ή σειρά), title(ο τίτλος), director(ο σκηνοθέτης), cast(οι ηθοποιοί που έχουν λάβει μέρος), country(η χώρα που έγιναν τα γυρίσματα), date_added(η ημερομηνία που προστέθηκε στην συλλογή του Kaggle),release_year(η χρονιά που προβλήθηκε),rating(οι ηλικίες που προορίζεται),duration(η διάρκεια, σε λεπτά όταν αναφέρεται σε ταινίες και σε κύκλους όταν αναφέρεται σε σειρές), listed_in(σε τί είδος κατατάσσεται, πχ δράμα) και το τελευταίο πεδίο αναφέρεται στο discerption(η περιγραφή).

Από αυτά τα πεδία που αναφέρθηκαν προηγουμένως, για την εργασία μας, κρατήσαμε τα εξής: title, release_year,duration,listed_in και description

Πως έγινε η επεξεργασία των άρθρων:

Η κάθε ταινία-σειρά θα είναι ένα document στην Lucene. Κάθε πεδίο χωρίστηκε μα βάση το κόμμα του, αλλά επειδή πολλά πεδία είχαν κόμματα, κάναμε μια προεπεξεργασία και όλα τα κόμματα έγιναν κάτω παύλα(_). Αυτό αποσκοπούσε στο να μπορούμε να διαβάσουμε το αρχείο με βάση τα πεδία του. Κατά τη διαδικασία του φορτώματος είχε μερικά λάθη, συγκεκριμένα στις γραμμές 8204 και 8422, στις οποίες το κείμενο που αναγραφόταν δεν ήταν σε μια γραμμή και την μεταφέραμε από πάνω ώστε να αποτελεί μια συνέχεια και συνάφεια με το προηγούμενο. Η μονάδα εγγράφου έχει ταινία-σειρά και αποτελείται από τα πεδία που αναφέρθηκαν πριν.

Αναζήτηση ταινιών και τα είδη ερωτημάτων:

Τα ευρετήρια θα είναι 4, για τίτλο, για χρονολογία, για είδος και για περιγραφή, με τα αντίστοιχα πεδία title, release_year, description, listed_in, δηλαδή τα πεδία εκείνα που κρατήσαμε κατά την διάρκεια του σχεδιασμού του συστήματος.

Όσο αναφορά την διαδικασία της αναζήτησης, θα υπάρχει μια μπάρα αναζήτησης και δίπλα από αυτήν θα υπάρχουν επιλογές για κάθε πεδίο αναζήτησης.

1. Για τον τίτλο θα υπάρχει ένα checkbox δηλαδή αν ο χρήστης θέλει να γίνει αναζήτηση με βάση το τίτλο.
2. Για την χρονολογία ένα dropdown menu ή ένα πεδίο, στο οποίο θα γράφει ο χρήστης την χρονολογία
3. Το ίδιο θα συμβεί με το είδος της ταινίας-σειράς.

Αν ο χρήστης δεν επιλέξει κάποιες από τις διαθέσιμες επιλογές, τότε η αναζήτηση θα γίνει με κατάλληλους συντέλεσες, οι οποίοι θα οριστούν και θα αποδοθούν στα εκάστοτε πεδία που έχουμε δημιουργήσει.