LETTRES
SORBONNE
UNIVERSITÉ

Projet final
Moteur de recherche
un cas basé sur les films biographiques

20 nov. 2023

Yuyan QIAN

Indexation sémantique et recherche d'information

Plan

Corpus cinématographique

Aperçu du corpus

Origine du corpus

- *Ensemble de données de films IMDb : tous les films par genre* sur Kaggle[1]
- Licence CC BY-NC-SA 4.0
- 16 fichiers CSV, soit 16 genres du film (130,45 Mb)[2]

Le fichier *biographie.csv* :

- 2,73 Mb
- 8289 instances et 14 attributs
- Nombre entier, nombre décimal, chaîne de caractère, texte long, etc.

1. https ://www.kaggle.com/datasets/rajugc/imdb-movies-dataset-based-on-genre/data
2. *Action, Aventure, Animation, Biographique, Crime, Famille, Fantastique, Film noir, Historique, Horreur, Mystère, Romantique, Science fiction, Sport, Thriller, Guerre.*

Nb	Attribut	Description	Exemple	Type
1	movie_id	ID du film IMDB	*tt3704428*	string
2	movie_name	Titre du film	*Elvis*	TG
3	year	Année de sortie	*2022*	pint
4	certificate	Classification du film	*PG-13*	string
5	run_time	Durée d'exécution	*159 min*	pint
6	genre	Genre du film	*Biography, Drama*	TG
7	rating	Note moyenne du film	*7.3*	pfloat
8	gross (in $)	Revenus bruts en dollar	*151040048.0*	pdouble
9	votes	NB de votes sur IMDB	*189732.0*	pdouble
10	director	Réalisateur/réalisatrice	*Baz Luhrmann*	TG
11	director_id	ID de réalisateur(trice)	*/name/nm0525303/*	–
12	star	Casting principal	*Tom Hanks, etc.*	TG
13	star_id	ID des acteurs(trices)	*/name/nm0000158/, etc.*	–
14	description	Description du film	*The life of American music icon Elvis Presley...*	TG

Table 1 – Illustration des données du fichier *biographie*. Pour raccourcir, TG signifie *Text general*.

Faiblesses du corpus

- Informations redondantes :
  - ID de réalisateur(trice), ID d'acteur(trice)
- Valeurs anormales :
  - I, II, XVI, etc. (qui ne correspond pas à la vraie année de sortie)[3]
- Erreur de Type de Données pour les *Votes* sur IMDb :
  - Nombres décimaux au lieu d'entiers (*189732.0*)
- Absence de la date de récupération des données
- Lacune de valeurs

3. Probablement à cause de l'extraction automatique du site web. En outre, les instances ayant des valeurs anormales sont en manque d'autres valeurs d'attribut

Pré-traitement du corpus

Pré-traitement du corpus

- Enlèvement de deux catégories d'informations redondantes
- Filtrage des valeurs anormales
- Uniformisation des données de la colonne *Durée* (159 ~~min~~)
- Ajout des valeurs booléennes pour l'existence de description (True ou false)

Indexation du fichier *biographie.csv*

Création du schéma

Attribut	Type	Requis	MultiValeurs
id	string	true	false
movie_name	text_general	true	-
certificate	string	-	-
description	text_general	-	-
director	text_general	-	-
genre	text_general	-	-
star	text_general	-	-
runtime_min	pint	-	-
year	pint	-	-
rating	pfloat	-	-
gross_in_dollar	pdouble	-	-
votes	pdouble	-	-
with_description	booleans	-	-

Table 2 – Paramères de différents attributs dans le schéma

Indexation du corpus

Éléments Clés

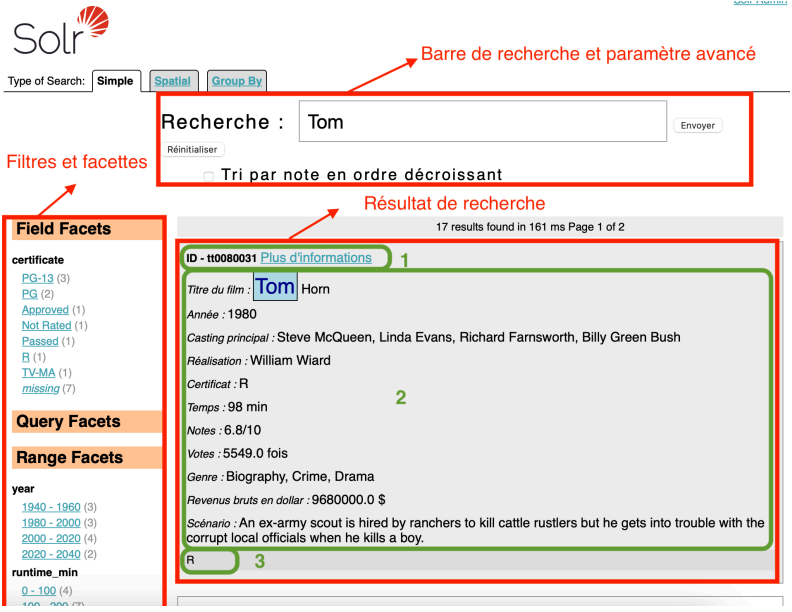
Sur le corpus :

- Noms des colonne ($, (,), espace, etc.)

Sur les lignes de commandes :

- Délimiteur de point-virgule (%3B)
- Espaces inutils dans les champs (f.champs.trim=true)

Modification de 3 parties d'UI


Solr
Type of Search: Simple Spatial Group By
Barre de recherche et paramètre avancé
Recherche : Tom
Envoyer
Réinitialiser
Tri par note en ordre décroissant
Filtres et facettes
Résultat de recherche
Field Facets
certificate
PG-13 (3)
PG (2)
Approved (1)
Not Rated (1)
Passed (1)
R (1)
TV-MA (1)
missing (7)
Query Facets
Range Facets
year
1940 - 1960 (3)
1980 - 2000 (3)
2000 - 2020 (4)
2020 - 2040 (2)
runtime_min
0 - 100 (4)
17 results found in 161 ms Page 1 of 2
ID - tt0080031 Plus d'informations
1
Titre du film : Tom Horn
Année : 1980
Casting principal : Steve McQueen, Linda Evans, Richard Farnsworth, Billy Green Bush
Réalisation : William Wiard
Certificat : R
Temps : 98 min
2
Notes : 6.8/10
Votes : 5549.0 fois
Genre : Biography, Crime, Drama
Revenus bruts en dollar : 9680000.0 $
Scénario : An ex-army scout is hired by ranchers to kill cattle rustlers but he gets into trouble with the corrupt local officials when he kills a boy.
R
3

Création des facettes

- Facette de champs
  - Classification cinématographique
  - Genre [4]
- Facette d'intervalle :
  - Année de sortie : 1903-2024
  - Notes : 1,0-9,8
  - Durée d'exécution : 0-442 min
- Facette de pivot :
  - Existence de la description du film
- Facette de requête : vide

4. Dans le genre cinématographique biographique, il existe également des différences en termes de genres, car un film peut être étiqueté simultanément comme biographique, dramatique, musical, etc.

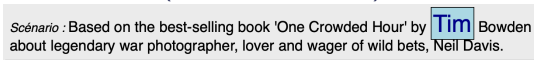
Adaptation de l'interface d'utilisateur

1. Tir en ordre DESC (*query.vm*, *query_form.vm*)

2. Plus de résultats similaires (*richtext_doc.vm*)

3. Badge de classification cinématographique (*richtext_doc.vm*)

4. Highlighting (*richtext_doc.vm*)



Conclusion et perspective

Conclusion

- Bien connaître les données du corpus
- Adapter les données à l'outil d'indexation (comme Solr)
- Améliorer la visualisation des résultats de recherche

Perspectives

- Extension au corpus avec image ou avec coordonnée géographique
- Corpus multilingue
- Enrichissement des facettes de filtrage
- ...

Jouer sur l'interface d'utilisateur

Browse

Jouer sur l'interface administrative

Core-overview