

Bases de données, ENS Cachan & Ulm

TP n° 8 – Recherche en texte intégral

Pierre Senellart (pierre@senellart.com)

16 mai 2008

Le but de ce TP est d'exploiter les fonctionnalités de recherche en texte intégral (*full-text search*) de MySQL.

1 Recherche en texte intégral

1. Le fichier `/home/senellar/imdb/plots.csv` contient des données issues d'IMDB, un site Internet d'informations sur les films. La première colonne de ce fichier donne le titre d'un film, la deuxième un résumé, la troisième l'auteur de ce résumé. Plusieurs résumés peuvent être indiqués pour un film donné. Concevoir un schéma de table permettant d'accueillir ces données et les importer dans MySQL. L'outil d'importation de PHPMyAdmin ne permettant pas d'importer de trop gros fichiers, on utilisera l'outil en ligne de commande `mysql`, et l'ordre :

```
LOAD DATA LOCAL INFILE 'fichier' INTO TABLE nom_table
  FIELDS TERMINATED BY 'char'
  ENCLOSED BY 'char2'
  ESCAPED BY 'char3';
```

On adaptera bien sûr `fichier`, `nom_table`, `char1`, `char2` et `char3` au format du fichier.

2. Écrire une requête SQL pour calculer le nombre de films distincts dans la table. On pourra utiliser une requête imbriquée :
`SELECT ... FROM (SELECT ...) nom_table_temporaire;`
3. Écrire une requête SQL pour calculer, pour tout n , le nombre de films ayant exactement n résumés dans la table.
4. Utiliser une requête avec l'opérateur `LIKE` pour répondre à la question suivante : « Dans quel film apparaissent à la fois un archéologue (*archeologist*) et des Nazis ? ».
5. Le support MySQL des requêtes en texte intégral est présenté dans la documentation à l'URL suivante :
<http://dev.mysql.com/doc/refman/5.0/fr/fulltext-search.html>
Étudier cette page, et utiliser l'opérateur `MATCH(...) AGAINST(...)` pour répondre aux questions suivantes (on n'aura besoin ni de `IN BOOLEAN MODE`, ni de `WITH QUERY EXPANSION`) :
 - (a) Quels films mettent en scène un archéologue ?
 - (b) Quel film est centré sur un Nazi avec une liste de Juifs à sauver (*a nazi with a list of jews to save*) ?
 - (c) Dans quels films peut-on voir des grands singes (*apes*) et un vaisseau spatial (*spaceship*) ?
6. Le fichier `/home/senellar/imdb/ratings.csv` donne, pour chaque film, une note d'appréciation moyenne ainsi que le nombre d'évaluateurs. Ces deux données peuvent servir à définir une notion d'*importance* pour un film. Importer la table dans MySQL, proposer une définition d'importance, et l'utiliser pour améliorer l'ordre des films renvoyés par une requête en texte intégral sur les résumés. Appliquer aux trois requêtes précédentes, et comparer les résultats.

2 Interface

Concevoir une interface Web de recherche en texte intégral dans les résumés de films d'IMDB, avec les fonctionnalités suivantes :

- classement ou non selon l'importance du film ;
- possibilité de rechercher dans le titre, les résumés, ou les deux ;
- accès par un lien à la fiche du film dans IMDB (accessible grâce à l'URL `http://www.imdb.com/find?q=toto` où `toto` est le nom du film) ;
- affichage de la liste des résultats avec résumés complets.

Réfléchir également à la mise en valeur des termes recherchés dans les résultats affichés.

3 Compléments

Les données des fichiers `plots.csv` et `ratings.csv` ont été extraites de la base de données d'IMDB, téléchargeable depuis `ftp://ftp.fu-berlin.de/pub/misc/movies/database/` et, en particulier, des fichiers `plot.list.gz` et `ratings.list.gz`. Effectuer le traitement permettant d'obtenir depuis ces derniers des fichiers importables directement dans MySQL, avec l'outil ou langage de programmation de votre choix.