le cnam

Intéraction avec une base Elasticsearch

Travaux Pratiques

CNAM Paris
traversn / fournier (at) cnam.fr

Chapitre

Table des matières

Chapitre 1

Mise en place

Ces travaux pratiques ont pour but de comprendre l'interrogation d'une base *elasticsearch*. Pour installer et comprendre le lancement de la base, veuillez vous référer au guide pratique sur `http://chewbii.com/elasticsearch/`. Veuillez installer et lancer *elasticsearch*.

1.1 Installation du jeu de données

Pour ce TP, nous allons utiliser un jeu de données contenant des films :

1.1.1 Le jeu de données peut être téléchargé ici :
`http://chewbii.com/wp-content/uploads/2016/03/movies_elastic.json_.zip`

1.1.2 Décompresser le fichier `movies.json.zip`

1.1.3 Pour importer le fichier [1], dans une console :

```
curl -XPUT localhost:9200/_bulk --data-binary @movies.json
```

1.2 Tester la base

Pour tester l'importation de la base, veuillez ouvrir un navigateur et ouvrir l'URL suivante : `http://localhost:9200/_plugin/head`

Vous devrez y voir le nom de la base (`index` sous *elasticsearch*) : 'movies'. Et la collection (`type`) : 'movie'.

1.3 Document 'movies'

Les documents de l'index 'movies' de type 'movie' ont la structure suivante :

```
{
    "directors" : ["Joseph Gordon-Levitt"],
    "release_date" : "2013-01-18T00:00:00Z",
    "rating" : 7.4,
    "genres" : ["Comedy","Drama"],
    "image_url" : "http://ia.media-imdb.com/images/M/MV5BMTQxNTc3NDM2MF5BMl5BanBnXkFtZTcwNzQ5NTQ3OQ@@._
    "plot" : "A New Jersey guy dedicated to his family, friends, and church, develops
      unrealistic expectations from watching porn and works to find happiness and
      intimacy with his potential true love.",
    "title" : "Don Jon",
    "rank" : 1,
    "running_time_secs" : 5400,
    "actors" : ["Joseph Gordon-Levitt","Scarlett Johansson","Julianne Moore"],
    "year" : 2013
}
```

1. Si vous regardez le fichier, vous constaterez que chaque ligne est précédée par les informations d'index' et de 'type' pour l'importation

Chapitre 2

Interrogation

L'interface Web 'head' de *elasticsearch* permet d'exécuter des requêtes textuelles. Pour ce faire, allez sur l'onglet 'Autres requêtes', sélectionnez l'index 'movies' avec le type 'movie'.

Pour chaque requête, écrire (lorsque c'est possible) sous les deux formes suivantes :

— Méthode HTTP GET (URL dans le navigateur ou via *curl*), avec le paramètre : **q=**
`curl -XGET 'localhost:9200/movies/movie/_search?q=XX:YY&pretty=1'`

— Méthode HTTP POST (dans le formulaire de requête), à l'aide d'une requête sous forme de document JSon (DSL)

— Avec curl avec un fichier 'query.txt' contenant le fichier JSon :
`curl -XGET 'localhost:9200/movies/movie/_search?pretty=1' -d @query.txt`

2.1 Requêtes simples

2.1.1 Donner la liste des films dont le titre contient 'Star Wars' (requête de type 'match')

2.1.2 Films 'Star Wars' dont le réalisateur (directors) est 'Georges Lucas' (requête booléenne)

2.1.3 Films dans lesquels 'Harrison Ford' a joué

2.1.4 Films dans lesquels 'Harrison Ford' a joué dont le résumé (plot) contient 'Jones'

2.1.5 Films dans lesquels 'Harrison Ford' a joué dont le résumé (plot) contient 'Jones' mais sans le mot 'Nazis'

2.1.6 Films de 'James Cameron' dont le rang devrait être inférieur à 1000 (boolean + range query)

2.1.7 Films de 'James Cameron' dont le rang doit être inférieur à 100

2.1.8 Films de 'James Cameron' dont le rang doit être supérieur à 5, sans être un film d'action ni un drame

2.1.9 Films de 'J.J. Abrams' sorties (released) entre 2010 et 2015 (filtered query)

2.2 Agregation

Nous souhaiterions maintenant réaliser quelques agrégats sur la base de films. Pour cela, utilisez la commande suivante :

`curl -XGET 'localhost:9200/movies/movie/_search?search_type=count&pretty=1' -d @query.txt`

Le fichier "query.txt" contiendra la requête d'agrégat souhaité.

2.2.1 Donner par année le nombre de film sortie,

2.2.2 Par catégorie de film, donner leurs occurences,

2.2.3 Dans le titre de film, donner les occurences des termes utilisés,

2.2.4 Dans la note (rating) moyenne des films,

2.2.5 Dans la note (rating) moyenne des films de Georges Lucas,

2.2.6 Dans la note (rating) moyenne des films par genre,

2.2.7 Dans la note (rating) minimum, maximum et moyenne des films par genre,

2.2.8 Dans le rang (rank) moyen des films par metteur en scène (directors),

2.2.9 Compter le nombre de films par tranche de note (0-1.9, 2-3.9, 4-5.9...),

2.2.10 Termes les plus utilisés (agrégat : significant_terms) dans les descriptions des films de Georges Lucas,

2.2.11 Nombre de metteur en scène distincts pour les films d'aventure.