NFE204

Introduction à la Recherche d'Information
S3: la pratique: requêtes booléennes

Auteurs : Raphaël Fournier-S'niehotta, Philippe Rigaux
(fournier@cnam.fr, philippe.rigaux@cnam.fr)

EPN Informatique
Conservatoire National des Arts & Métiers, Paris, France

Plan du cours

Interrogation

- ElasticSearch s'appuie sur le système d'indexation **Lucene**, dont le rôle est essentiellement de créer les index inversés, et d'implanter les algorithmes de parcours brièvement introduits dans la session précédente.
- Lucene propose un langage de recherche basé sur des combinaisons de mot-clés, langage étendu et raffiné par ElasticSearch (cf plus tard)
- La première méthode pour transmettre des recherches est de passer une expression en paramètre à l'URL :

```
$ curl http://localhost:9200//nfe204/movies/_search?q=alien
# utilisable dans l'interface Kopf, onglet Rest
```

la réponse d'ES


```
{
"took": 3,
"timed_out": false,
"_shards": {
        "total": 4,
        "successful": 4,
        "failed": 0
},
"hits": {
        "total": 20,
        "max_score": 1.2078758,
        "hits": [
                {
                        "_index": "movies",
                        "_type": "movie",
                        "_id": "764",
                        "_score": 1.2078758,
                        "_source": {
                                "fields": {
                                        "directors": [
                                                "Duncan Jones"
                                        ],
                                        "genres": [
                                                "Action",
                                                "Adventure",
                                                "Fantasy"
                                        ],
                                        "plot": "An epic fantasy/adventure based on the popular video game series.",
                                        "title": "Warcraft",
                                        "rank": 764,
                                        "actors": [
                                                "Paula Patton",
                                                "Paul Dano",
                                                "Anson Mount"
                                        ],
                                        "year": 2015
                                },
                                "id": "tt0803096",
                                "type": "add"
                        }
                },
```

Termes

- Notion de base : le **terme**
- c'est un mot au sens usuel
- ou une séquence de mots entre apostrophes

On peut interroger un index avec :

```
space vessel
```

Puis :

```
"space vessel"
```

- Première recherche : documents avec "space", "vessel" ou les deux
- Deuxième : seulement "space vessel" (côte à côte)

Termes (suite)

- la recherche d'un terme s'effectue toujours sur un champ.
- La syntaxe complète pour associer le champ et le terme est:

```
champ:terme
```

- si non précisé, c'est le champ par défaut qui est utilisé
- pratique courante : concaténer toutes les chaînes de caractères en un champ "text" général, défini par défaut
- Nos requêtes deviennent :

```
text:space text:vessel
```

- et

```
text:"space vessel"
```

Termes (suite)

- Les valeurs des termes (dans la requête) et le texte indexé sont tous deux soumis à des transformations spécifiées dans le schéma.
- Une transformation simple est de tout transcrire en minuscules.

```
text:"Space Vessel"
```

- Les transformations appliquées à la requête ET au texte indexé doivent être cohérentes : si les termes sont transformés en majuscules, et le texte indexé en minuscules, on n'aura jamais de résultat!

Termes (suite)

On peut spécifier des termes (pas des séquences) incomplets

- le '?' indique un caractère inconnu
  - "opti?al" désigne "optimal", "optical", etc.
- le '*' indique n'importe quelle séquence de caractères
  - "opti*" pour toute chaîne commençant par "opti"

Approximations avec "∼" :

- Rechercher "optimal" et "optimal∼"
- 0 et 1 résultat ("optical")
- Proximité des termes par une distance d'édition :
  (nb opérations pour passer de "optimal" à "optical")

Intervalles :

- [] bornes comprises
- { } bornes exclues

```
%price:[100 TO 200]
```

Requêtes Booléennes

- Les critères peuvent être combinés avec des opérateurs Booléens : **AND**, **OR** et **NOT**
- Attention : majuscules

```
%price:[100 TO 300] OR popularity:5
%price:[100 TO 300] AND NOT popularity:5
%popularity:6 AND features:matrix
```

- Par défaut, c'est un **OR** qui est appliqué
- Recherche sur plusieurs critères ramène l'union des résultats sur chaque critère pris individuellement

Exercice Exprimez les recherches suivantes sur votre base de données :

- les films dans lesquels on parle de "hunter";
- même critère, mais en ajoutant le mot-clé "bounty";
- films avec Kate Winslett et Leonardo di Caprio;
- films qui sont soit des drames, soit du fantastique;
- films avec le mot-clé « France »; obtient-on les films produits en France? Sinon pourquoi? Que faudrait-il faire?
- on recherche le film « Sleepy Hollow »; effectuez une recherche sur le titre (« Sleepy », « Hollow », « Sleepy Hollow ») puis sur le résumé.
- films satisfaisant une combinaison de critères: parus entre 1990 et 2000 et aux USA, ou contenant les mots-clés « Michael » et « Sonny »;

Vous êtes invités à effectuer les recherches avec ou sans majuscules, à chercher des phrases comme « bounty and hunter », à indiquer ou non des noms de champs, et à interpréter les résultats (ou l'absence de résultat) obtenus.