Projet Solr

Un moteur de recherche sur les personnages Marvel et DC

Alexane Jouglar

Sorbonne Université

Dans le cadre de notre enseignement d'indexation sémantique et de recherche d'information, nous avons travaillé au développement d'un moteur de recherche. Ce travail se divisait en trois grandes parties. Il a fallu tout d'abord choisir un corpus, l'indexer puis mettre en place le moteur de recherche grâce à Velocity. C'est dans cet ordre que nous parlerons de ce travail.

Le choix du corpus

Le corpus utilisé renseigne les caractéristiques de 16 376 personnages Marvel et 6 896 personnages DC comics. Les deux ensembles ont été fusionnés au sein d'un même fichier renommé marvel_dc_data.csv et enregistré au sein du dossier ExempleDocs de Solr. Pour chaque personnage, nous avons douze caractéristiques.

Les données viennent des wiki respectivement dédiés aux personnages Marvel et DC comics, et sont librement disponibles depuis un repository GitHub (https://github.com/fivethirtyeight/data/tree/master/comic-characters). La licence est la suivante : CC-BY-4.0 License.

A l'aide d'expressions régulières, certaines colonnes ont été modifiées. L'url de chaque document a été complétée avec l'ajout d'un http://... qui n'existait pas dans le fichier original. La colonne *alive* est devenue booléenne : si le personnage est vivant, il est marqué *true*, sinon *false*.

Table 1: Colonnes du fichier csv : caractéristiques pour chaque document d'après le README du corpus.[a]

page$_id$	L'identifiant de la page du personnage au sein du wiki dédié au monde en question
name	Le nom du personnage
urlslug	L'url qui mène vers la page wiki du personnage
ID	si l'identité du personnage est secrète, publique ou les deux
align	si le personnage fait partie des méchants ou des gentils
eye	la couleur des yeux
hair	la couleur des cheveux
sex	le sexe du personnage
gsm	si le personnage fait partie d'une minorité sexuelle
alive	si le personnage est vivant ou décédé
appearances	le nombre d'apparition du personnage dans les comics
year	l'année d'apparition
id	L'identifiant unique donné par défaut pour ce travail

[a]de ces douze colonnes, seule une n'est pas d'origine :l'id qui a été généré ultérieurement pour obtenir un identifiant unique pour chaque personnage. Cet identifiant peut être incrémenté pour ajouter des personnages au fichier.

Indexation des données

La première étape consistait en la création d'une collection MarvelDC en solr. Pour ce faire, en Linux, il suffit de se rendre dans le dossier bin de la plateforme Solr, d'y ouvrir un terminal et d'exécuter la commande suivante :

```
./solr create -c MarvelDC
```

Pour indéxer les documents à la collection créée, nous nous rendons dans le dossier ExampleDocs (au sein d'Example), où l'on a précédemment ajouté le fichier marvel_dc_data.csv, y ouvrons un terminal et d'exécutons la commande suivante :

```
java -Dtype=text/csv -Dcommit=yes -Dc=MarvelDC -jar post.jar mar-
vel_dc_data.csv
```

Il s'agit alors de modifier le schéma d'indexation présent dans le fichier xml

Table 2: Type pour chaque champ d'un document.

align	text_general
alive	booleans
appearances	plong
eye	text_general
genre	string
gsm	text_general
hair	text_general
id	string
name	text_general
page_id	pint
secrecy	text_general
sex	text_general
urlslug	text_general

managed-schema de la collection MarvelDC et d'indiquer le type de chaque champ. Nous récapitulons dans le tableau ci-dessus ces informations.

Lancement du moteur de recherche avec Velocity

La troisième étape repose sur l'utilisation de Velocity. Pour ce faire, nous utilisons le fichier Velocity d'ores-et-déjà formé et présent dans la collection technoProducts précédemment réalisé en cours. Nous copions ce fichier tel quel dans notre serveur au sein du dossier conf. Plusieurs modifications sont nécessaires pour que le moteur de recherche soit fonctionnel.

Nous proposons des facettes et nous les inscrivons dans le fichier solrconfig.xml. Le tableau ci-dessous reprend les facettes utilisées et leur type.

Nous inscrivons dans le fichier solrconfig.xml de notre collection les deux balises *lib*, *QueryResponseWriter*, et *resquestHandler* tel qu'indiqué dans le td5.

Nous choisissons ensuite les champs qui seront visibles pour un document donné dans les résultats d'une recherche. Les champs choisis sont : *name*, *id*, *urlslug*,

Table 3: Champs utilisés pour chaque type de facette.

4*Facettes de champ	genre
	align
	secrecy
	sex
2*Facettes d'intervalles	appearances
	year
Facettes pivot	alive

secrecy, *align*, *year*. On les renseigne dans le fichier product_doc.vm présent dans Velocity.

Par ailleurs, il est nécessaire de supprimer dans les fichiers adéquats les éléments de l'interface que nous ne souhaitons pas voir.

La dernière étape repose sur la modification du main.css. Cette étape n'est pas indispensable, elle sert simplement à personnaliser l'environnement. Nous changeons quelques couleurs et arrondissons le bord des cadres.

Conclusion

Dans ce rapport nous avons présenté le travail réalisé pour le module d'indexation sémantique. Ce travail consistait en la construction d'un moteur de recherche à partir d'une liste de documents. Nous avons choisi un fichier csv portant sur les personnages des univers Marvel et DC. Ce travail nous a permis d'en apprendre davantage sur le fonctionnement des moteurs de recherche et sur les possibilités qu'ils offrent.