Cours de Recherche Intelligente dans les Textes

CASALIS
Garance

Rapport du projet avec Solr

1) Présentation du corpus

Pour ce projet de Recherche Intelligente dans les Textes, nous avons décidé de traiter un corpus qui liste les plaintes effectuées à New York (au New York City Police Department) entre septembre 2016 et septembre 2017. Il comporte 351 509 entrées, et 24 colonnes. Nous l'avons téléchargé sur le site NYC Open Data – il est rédigé en anglais et est libre de droits. Il est disponible à l'adresse suivante : https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-YTD/5uac-w243. Il est fourni par le New York Police department (NYPD).
Nous l'avons téléchargé au format CSV.

2) Nettoyage du Corpus

Dans un premier temps, il s'agit de nettoyer le corpus. Bien qu'étant très clair et ne comportant aucune cellule descriptive avec un long texte, nous avons décidé de supprimer certaines colonnes et de les renommer pour qu'elles soient plus explicites – avec le corpus était fourni un document explicitant les données fournies dans les colonnes.

Nous avons utilisé le langage de programmation Python, plus précisément, la librairie d'analyse de données Pandas.
Auparavant, dans les colonnes, il y avait la latitude, la longitude de l'endroit d'où provenait la plainte, ou encore le code de référence du type de crime commis.

Nous avons gardé et renommé 9 colonnes :

- Complaint_ID : l'identifiant sous lequel est enregistrée la plainte
- Start_Date : La date à laquelle a été émise la plainte
- End_Date : La date à laquelle a été résolue la plainte, si elle l'a été
- Crime : Le crime signalé
- Type_of_Crime : la catégorie dans laquelle a été classifiée le crime
- State : (completed / attempted) indique si le problème a été résolu ou non
- Jurisdiction : le département de police où la plainte a été enregistrée
- Borough : l'arrondissement de New York d'où provient la plainte
- Exact_Place : le type de bâtiment ou d'endroit où a été commis le crime, s'il a été décrit par le plaignant (dans la rue, dans une épicerie, dans un bar, etc.)

Après ce premier nettoyage avec Python, nous avons utilisé des expressions régulières dans Notepad++ pour effectuer quelques mises en formes supplémentaires, de façon à ce que le corpus soit compatible avec Solr.

La colonne concernant la date et celle concernant l'heure ont été regroupées en une seule. Les dates n'étant pas au bon format, nous avons modifié celui-ci :

Recherche	Remplacement
([0-9]{2})/([0-9]{2})/([0-9]{4}),([[0-9]{2}:[0-9]{2}:[0-9]{2})	(\3)-(\1)-(\2)T(\4)Z
09/30/2017,23:46:00	2017-09-30T23:46:00Z

Parfois, dans la partie End_Date, la date était annoncée, et pas l'heure. Dans d'autres cas, l'heure était annoncée et non la date. Quand l'heure seulement était annoncée, cela signifiait que la date était la même que dans la colonne précédente, Start_Date. De fait, nous avons utilisé l'expression régulière suivante :

Recherche	Remplacement
([0-9]{4}-[0-9]{2}-[0-9]{2})(T[0-9]{2}:[0-9]{2}:[0-9]{2}Z),,([[0-9]{2}:[0-9]{2}:[0-9]{2})	(\1)(\2),(\1)T(\3)Z

Quand l'heure de la End_Date n'était pas marquée mais la date si, nous l'avons normalisée à 00:00:00

Recherche	Remplacement
([0-9]{2})/([0-9]{2})/([0-9]{4}),,	(\3)-(\1)-(\2)T00:00:00Z,

3) Indexation dans Solr

Nous avons ensuite créé la collection **Complaint_Data**, et y avons indexé le corpus grâce à la ligne de commande suivante (nous avons pris le parti de garder les virgules comme moyen de séparation entre les différentes colonnes du corpus plutôt que de les remplacer par des tabulations) :

java -jar -Dtype=text/csv -Dcommit=yes
-Durl="http://localhost:8983/solr/Complaint_Data/update/csv?commit=true&separator=**%2C**"
post.jar NYPD_Cleaned_Corpus.csv

4) Création des facettes

Pour les facettes de champs, nous avons décidé de proposer une recherche :

- par type de crime (Crime),
- par état d'avancée de l'affaire (State),
- par arrondissement de New York (Borough)

Pour la facette pivot, nous avons décidé d'associer la colonne « Crime » avec la colonne « State ». En effet, cela permet de vérifier si les vols à main armée sont statistiquement plus résolus que les cas de harcèlement ou de conduite en état d'ivresse.

Pour les facettes de Range, n'ayant pas beaucoup de chiffres comme des prix, nous avons pris le parti de faire une facette de range par Start_Date, et une autre par End_Date ; nous avons fixé l'intervalle à un mois pour chaque facette de range. Cela permet d'accéder aux plaintes en fonction du mois auquel elles ont été effectuées.

Nous avons ensuite modifié le document « richtext.vm » pour afficher le contenu des documents