Projet Interpromo 2021

Groupe 8 : Innovation

Mise à jour des mots-clés

Cheffe : Flora Estermann
Cheffe adjointe : Katia Belaid
Chef qualité : Tanguy Dirat

Samba Seye
Célya Marcélo
Théo Saccareau
Damien Sonneville
Mélina Audiger
Jordi Mora Fernandez

SOMMAIRE

1. Introduction

Actuellement, le scraping est réalisé en utilisant une liste de keywords issus d'un produit cartésien entre des lexiques de gestion et d'innovation.

Le problème concerne le fait que les résultats obtenus ne sont pas forcément tous pertinents et génèrent parfois beaucoup de données à enregistrer inutilement.
Comment évaluer la pertinence de cette liste de keywords et la faire évoluer pour optimiser le scrapping réalisé lors des veilles ?

Dans un premier temps, nous présenterons une méthode permettant d'ajouter de nouveaux mots-clés à la liste en prenant en considération le contenu des requêtes utilisateurs vis-à-vis d'un thème donné.
Dans un second temps, nous verrons comment évaluer la pertinence de cette liste, pour à terme, pouvoir supprimer les mots-clés inutiles ou obsolètes, en utilisant des caractéristiques adaptées et un seuillage pertinent en fonction du temps.

2. État de l'art

Ajout de nouveaux mots-clés

Pour ajouter de nouveaux mots-clés pertinents, nous avons décidé de nous baser sur le contenu des requêtes utilisateurs. En effet, le fait de récupérer les données des requêtes précédentes permet d'identifier des mots représentatifs qui vont, à terme, servir dans le cadre du scraping. Plus il y aura de requêtes utilisateurs, plus les résultats de cette méthode seront pertinents.
En prenant en considération un thème donné, nous évaluons la similarité entre ce thème et chacune de nos requêtes afin de les classer par degré de similarité. Le degré de similarité thème/requête est calculé en utilisant une représentation vectorielle du thème et de la requête et en évaluant la Word Mover Distance entre les deux. La Word Mover Distance calcule la distance minimale que les mots d'un texte doivent parcourir pour atteindre les mots d'un autre texte.
Les requêtes ainsi classées par similarité, nous créons un nuage de mots à partir des requêtes les plus similaires au thème. Les mots clés que nous identifions sont les mots les plus importants, en termes de fréquence, retournés par le nuage de mots.
Nous intégrerons par la suite ces nouveaux mots-clés au lexique initial.

Suppression des mots-clés inutiles ou obsolètes

La suppression des mots-clés inutiles pourrait s'effectuer par rapport à leur importance dans les recherches web. L'analyse va donc suivre deux axes différents : la récurrence avec

laquelle les gens cherchent les mots qui composent les tuples, et leur présence dans les résultats des moteurs de recherche.
Pour la première méthode, on se base sur une mesure temporelle de la présence des mots dans le web. En effet, on calcule la fréquence avec laquelle le mot a été recherché dans les moteurs de recherches, pour ainsi pouvoir décider s'il est important ou bien obsolète à l'analyse.
Pour ceci, on exploite la librairie pytrend, une API de Google Trend qui nous permet de récupérer les occurrences d'un mot en fonction du temps. On peut ainsi étudier sa pertinence et l'évolution des caractéristiques choisies au fil du temps. Nous avons limité la période d'étude aux 12 derniers mois.
La deuxième méthode consiste à analyser si le mot a une grande présence dans le web. Cela consiste à récupérer le nombre de résultats qu'on obtient dans les moteurs de recherche pour un mot donné. La pertinence de celui-ci va donc être déterminée par le nombre de résultats obtenus.
Pour la conception, on exploite la librairie BeautifulSoup, qui permet de parcourir les urls correspondants à chaque ensemble de mots-clés, et ainsi d'extraire des balises HTML l'information correspondante au nombre de résultats obtenus.

3. Résultats

La méthode d'identification des mots clés nous permet déjà d'identifier des mots significatifs pour un thème donné. Néanmoins, pour différents thèmes, nous notons l'apparition de mots clés identiques ainsi que de mots sans lien significatif avec le sujet. Cela est dû au fait que les requêtes utilisateurs relèvent à ce stade de la simulation uniquement. De plus, le volume des données de requêtes ne nous permet pas, pour le moment, d'avoir des résultats représentatifs. Ci-dessous un exemple de sorties pour deux thèmes différents : rh et informatique.

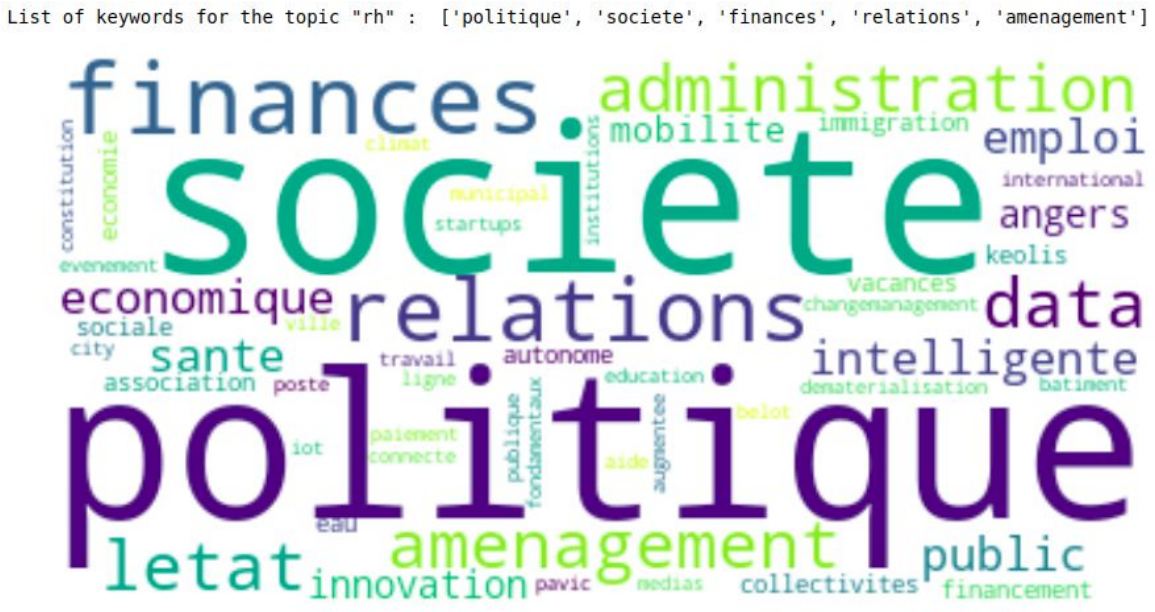

Figure 1. Sortie des mots clés et du nuage de mots associés au thème "RH"

Figure 2. Sortie des mots clés et du nuage de mots associés au thème “Informatique”

4. Discussion

On pourrait développer d’autres méthodes pour l’analyse et le traitement des mots-clés.

Pour l’analyse des mots-clés, on pourrait envisager de mettre en place des méthodes d’analyse statistique (représentation des features sous forme de séries temporelles, régression linéaire pour étudier les tendances saisonnières et faire de la prédiction sur les tendances à venir, etc).
De plus, il faudrait faire varier la période d’études des données pour vérifier si les tendances dépendent davantage de facteurs récents ou bien plus anciens.
Enfin, on pourrait développer un outil qui évalue automatiquement si un mot-clé a une grande importance dans le web avant de le définir en tant que mot-clé. Cette étape permettrait d’ajouter de la pertinence aux mots-clés retournés par notre algorithme.