GIF-4101 / GIF-7005 - Proposition de projet

Ce projet a pour objectif de tester vos connaissances des différents algorithmes d'apprentissage, en résolvant un problème industriel fourni par la compagnie Coveo. Coveo produit un moteur de recherche propulsé par l'intelligence artificielle. En étudiant le comportement des utilisateurs tout au long de leur séance de recherche, les informations recueillies sont utilisées pour améliorer la pertinence des résultats proposés par l'engin de recherche. Dans ce projet, nous vous proposons de vous attaquer à un ensemble de données provenant du site Web des ressources techniques publiques de Coveo, où nous avons extrait des données contenant des informations sur les *recherches* qui ont été effectuées, et les *clics sur des documents* qui en ont résulté.

Le problème à résoudre

Vous vous retrouvez dans la situation où vous avez accès à un jeu historique de données, contenant un ensemble de recherches qui ont été effectuées. Pour chaque recherche, vous avez accès à un certain nombre d'attributs dont la requête en langage naturel et de l'information sur l'utilisateur et sa visite en cours. Pour chaque recherche, vous trouverez un ensemble de documents qui ont été choisis par l'utilisateur. Nous prenons pour acquis que ces documents étaient *pertinents* pour l'utilisateur en question, au moment où il a fait cette recherche. Pour chaque document, vous avez également accès à certains attributs comme le titre de celui-ci.

Vous devez implémenter un algorithme d'apprentissage qui, à partir d'une nouvelle recherche, retourne un ensemble d'au plus 5 documents pertinents. Ce problème peut être vu comme un problème de *classification multi-étiquettes* (où chaque document unique est une classe), ou un problème de régression (où chaque paire requête/document se voit attribuer un score).

Les données

Les données sont fournies en format *csv* (*comma separated value*), où la première ligne est une entête spécifiant le nom de chaque colonne. Nous vous fournissons deux types fichiers : *searches* contenant les données de recherche, et *clicks* contenant les informations sur les clics. Chaque *search* contient les informations suivantes :

- `search_id`: Un identifiant unique pour cette recherche.
- `search_datetime`: La date et l'heure à laquelle été faite cette recherche
- `query_expression`: La requête effectuée par l'utilisateur.
- `visit_id`: Identifiant unique représentant la visite d'un utilisateur. Plusieurs recherches (et clics) peuvent avoir été réalisés dans une même visite.
- `visitor_id`: Identifiant unique représentant un visiteur. Lorsqu'un visiteur est anonyme mais qu'il revient avec le même navigateur (et n'a pas supprimé ses *cookies*), le même `visitor_id` lui sera attribué.
- `user_id`: Lorsque l'utilisateur est connecté, ce champs contient le nom d'utilisateur (anonymisé).
- `user_language`: La langue de l'utilisateur.
- `user_country`: Pays de l'utilisateur.
- `user_city`: Ville de l'utilisateur.

Chaque *click* contient les informations suivantes :

- `search_id`: L'identifiant unique du *search* après lequel a été effectué ce clic.
- `click_datetime`: La date et l'heure à laquelle a été effectué ce clic.
- `document_title`: Le titre du document cliqué.
- `document_author`: L'auteur (anonymisé) du document.
- `document_id`: L'identifiant unique correspondant au document.

D'autres champs sont disponibles dans les fichiers fournis, et pourront vous être expliqués sur demande si vous choisissez ce projet.

Les *searches* et les *clicks* sont divisés en trois ensembles : un ensemble d'entraînement, un ensemble de validation (que vous utiliserez pour rapporter vos performances dans votre rapport), et un ensemble de test pour lequel n'avez que les *searches*. Vous devrez fournir vos prédictions de jusqu'à 5 documents pertinents pour les *searches* de l'ensemble de test, qui sera utilisé par la suite par Coveo pour effectuer une évaluation finale de l'approche.

	Ensemble d'entraînement	Ensemble de validation	Ensemble de test
searches	52133	14895	7448
clicks	24491	6920	3567

La fonction d'évaluation

Afin d'évaluer la qualité de vos prédictions, nous utiliserons la fonction suivante. Soit $T := \{x_i, \mathbf{y}_i\}_{i=1}^n$ un ensemble de test, où chaque x_i correspond à un *search* et chaque $\mathbf{y}_i$ est un ensemble d'identifiants uniques de documents qui ont été cliqués suite à cette recherche. Chaque document dans l'ensemble $\mathbf{y}_i$ est considéré comme un document pertinent lié à la requête x_i.

Pour chaque recherche x_i, nous vous demandons un ensemble d'au plus 5 documents pertinents $\widehat{\mathbf{y}_i}$. Pour évaluer un ensemble de prédictions pour une recherche donnée, nous le considérons comme un succès si *l'un des documents prédits correspond à l'un des documents vraiment pertinents*. En d'autres termes, la fonction de **perte** est la suivante :

$$\ell(\mathbf{y}_i, \widehat{\mathbf{y}_i}) := \begin{cases} 1 & \text{si } \mathbf{y}_i \cap \widehat{\mathbf{y}_i} = \emptyset; \\ 0 & \text{autrement}. \end{cases}$$

La performance de votre approche sera évaluée sur un ensemble de test dont vous n'aurez pas accès pendant la réalisation de ce projet. Vous devrez donc nous fournir un fichier de prédictions en format CSV pour cet ensemble, où :

- Le premier élément de chaque ligne est le `search_id` auquel se rapporte la prédiction.
- Les 5 éléments suivants sont les `document_ids` prédits, en ordre décroissant de pertinence. Le document le plus pertinent selon votre algorithme se retrouve donc au début de la liste, puis le deuxième plus pertinent, et ainsi de suite.

Contact

Si la réalisation de ce projet vous intéresse, veuillez contacter Jean-Francis Roy (jfroy@coveo.com) et Sébastien Paquet (spaquet@coveo.com), avec votre professeur en copie conforme.