Proposition de projet

GIF-7005: Introduction à l'apprentissage machine

Équipe 10: Stéphane Caron, Philippe Blais, Philippe Blouin-Leclerc, Samuel Lévesque et William Bourget

26 octobre 2018

Description du projet

Dans le cadre du projet à réaliser dans le cours GIF-7005, notre équipe avons choisi la problématique proposée par l'entreprise Coveo. Le problème consiste à bâtir un modèle d'apprentissage supervisé qui, à partir d'une nouvelle recherche, retourne un ensemble d'au plus 5 documents pertinents à un utilisateur donné.

Jeu de données

Pour entraîner notre modèle, nous aurons accès à des données portant sur différentes recherches réalisées par des utilisateurs. Nous pourrons donc utiliser certaines informations générales sur l'utilisateur et sur la recherche effectuée, par exemple:

- le langue de la recherche
- le pays de provenance
- la ville
- la requête effectuée
- ...

De plus, nous aurons accès aux clics effectués par l'utilisateur en réponse à sa recherche. Ces clics nous donnent l'information sur les documents choisis par celui-ci. Cela sera d'ailleurs ce que nous tenterons de modéliser avec notre algorithme d'apprentissage.

Méthodologie anticipée

Pour bâtir notre modèle, nous prévoyons utiliser deux types de méthodes. Dans la première phase, nous pensions commencer par utiliser un modèle simple et intuitif dans un contexte de classification. Pour ce faire, nous pensions utiliser un algorithme commme celui des *k-PPV* (Alpaydin 2010). Dans la seconde phase, nous prévoyons tester des réseaux de neurones (Goodfellow, Bengio, and Courville 2016), qui sont d'ailleurs très populaires pour les systèmes de recommendation.

Les méthodes citées plus haut sont des méthodes supervisées. Cela veut dire que le modèle apprend à prédire une réponse connue. Nous aimerions aussi tester si des méthodes non-supervisées pourraient nous aider à entraîner des modèles supersivés par la suite. En d'autres mots, nous allons considérer la possibilité de regrouper les caractéristiques des recherches dans des groupes similaires et utiliser cette information comme variable dans notre modèle supervisé.

Finalement, nous prévoyons aussi utiliser le concept de *word2vec* pour trouver des représentations vectorielles des mots qui feront partis de nos documents et des titres de nos documents.

Références

Alpaydin, Ethem. 2010. *Introduction to Machine Learning.* 2nd ed. The MIT Press.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press.