Etude de cas avec le package R rvest

Pascal BIZZARI [a] et Amélie NEVEUX [b]

[a] Directeur Performance Analytique
Groupe AVISIA
8 av. Kléber 75016 PARIS
pbizzari@capmarket.fr

[b] Chef de projets Data
Groupe AVISIA
8 av. Kléber 75016 PARIS
aneveux@capmarket.fr

Mots clefs : Crawling, modélisation, machine learning.

Le Groupe AVISIA est un cabinet de conseil né en 2007 de la réflexion d'entrepreneurs spécialisés en décisionnel ayant l'ambition de faire profiter leurs clients d'une réelle expertise, tout en y ajoutant le pragmatisme et la réactivité d'un Cabinet à taille humaine. Au fil des années, nous n'avons cessé d'enrichir notre Offre jusqu'à devenir un acteur de référence en conseil, intégration et réalisation de projets DATA. Le Groupe AVISIA offre une réelle continuité d'accompagnement et de services reconnue par bon nombre de clients Grands Comptes où nos 120 consultants sont chaque jour mobilisés sur des projets d'envergure. Dans le cadre de nos travaux de veille sur les sujets de Big Data et de R&D nous avons mis en place un DataLab Interne pour mettre en place des projets ambitieux et innovants.

Un de nos projets consiste à mettre en œuvre une démarche analytique permettant de répondre à la question : Quelles sont les compétences les plus recherchées sur le marché de la DATA ? Et quels sont les profils les plus appétents ? Pour répondre à cette question, nous avons entrepris une démarche de récupération et d'analyse des offres d'emploi publiées sur le web avec pour objectif de bâtir un modèle de recommandation de profil pour chaque offre. Cette recommandation est ensuite mise à disposition auprès des utilisateurs de l'application qui qualifient la pertinence de la réponse (OK / KO) et peuvent indiquer la caractéristique expliquant le refus (ceci dans le but d'améliorer par la suite la pertinence du modèle). Ainsi, nous avons dû mettre en œuvre une procédure de récupération de l'information (offre & profil) sur le web à l'aide du package « rvest » de R, de structuration & qualification du contenu de cette information, puis de modélisation (supervisée & non supervisée) pour la recommandation. La complexité de cette application réside dans la profusion des offres et des profils sur la thématique qui nécessite d'avoir un processus automatisé (notamment pour la qualification du contenu de l'information) mais également auto-apprenant pour s'enrichir des retours des utilisateurs et proposer en quasi temps réel des recommandations adaptées.

Références

[1] http://cran.r-project.org/web/packages/rvest/rvest.pdf