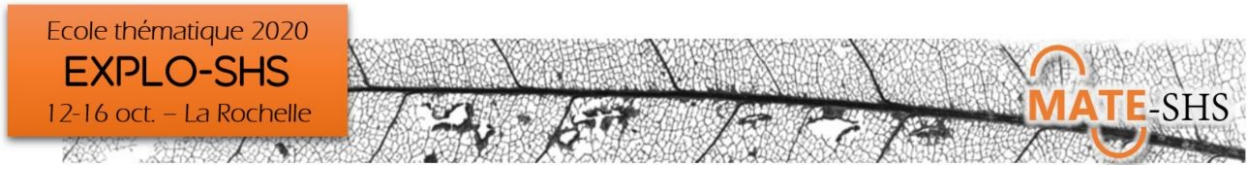

Le scraping de données structurées web à l'aide d'Extractify. Focus sur les données conversationnelles

Frédéric Vergnaud, Pascal Cristofoli

Atelier, jeudi 15 octobre 9h30-12h45

Présentation générale

Si en théorie la manière de structurer en HTML et CSS des données sur le web est plutôt bien définie par tout un ensemble de normes et de standards énoncés par différentes instances promouvant la compatibilité des technologies web, en pratique on se rend compte assez vite de la grande hétérogénéité qui prévaut dans ce domaine, rendant la plupart des méthodes et logiciels inopérants s'ils reposent sur l'identification des structures classiques pour en extraire l'information voulue.

L'atelier présente le logiciel libre Extractify, un plugin pour le navigateur Chrome, qui se propose de fournir à son utilisateur une interface simplifiée lui permettant de récolter n'importe quel type de données structurées en ligne. Après avoir décrit le logiciel, nous en étudierons les fonctions automatiques d'identification des structures HTML englobant les données recherchées. Dans un second temps, nous verrons qu'il est possible d'aller plus loin en utilisant les sélecteurs CSS. Enfin, dans le cadre d'un focus sur des données issues de forums de discussions, nous utiliserons le logiciel libre L@ME pour visualiser les données extraites et les exporter en vue de traitements statistiques ultérieurs.

Environnement informatique

Extractify : Liaison Internet & Navigateur Chrome L@ME : Java 7

Type de données traitées - droit d'accès

Tout type de données structurées et librement accessibles sur le web. Focus sur des forums de discussions publics.

Niveau requis

Tous niveaux. Un tour d'horizon rapide des sélecteurs CSS sera nécessaire pour les fonctions avancées.

Objectifs

Initiation au « scraping » de données en sciences sociales à l'aide d'une suite de logiciels libres développés pour des non-initiés.

Formule pédagogique

Après une présentation générale et un premier exemple rapide suivi par tous, chaque stagiaire pourra s'exercer à récolter son propre corpus.

Références bibliographiques

Frédéric Vergnaud. L@ME : un logiciel libre d'analyse et de traitement de messages électroniques. Tuto@MATE, 2017. ⟨hal-02393861⟩

Liens

https://github.com/fredericvergnaud/extractify
https://github.com/fredericvergnaud/lame