GARDESO - Groupe sur l'Analyse, la Recherche et le Développement En Source Ouverte

Introduction à la collecte de données sur le Web

Date : Jeudi 18 juin 2015
Horaire : 9h à 17h

Local : C-2059

Adresse :
Université de Montréal
Pavillon Lionel Groulx
3150, rue Jean-Brillant
Montréal QC H3T 1N8

Description :

Au cours des vingt dernières années, la croissance exponentielle du volume de données produites et disponibles sur le Web a considérablement transformé nos capacités de connaissance. Rendu possible par la convergence des avancées technologiques – augmentation de la puissance de calcul et de stockage des serveurs informatiques, diminution des coûts économiques associés à la virtualisation des données –, des évolutions propres au Web – montée en puissance des médias sociaux et de l'infonuagique – et des transformations sociétales plus générales – économie de l'information, transparence et e-gouvernement, quantification des données personnelles du Web 2.0, etc. – cette « numérisation du tout », aussi parfois qualifiée de révolution du *big data,* offre en retour de nouvelles opportunités d'analyse et de recherche. Grâce aux données produites ou stockées en ligne, de nouveaux chantiers scientifiques s'ouvrent et de nouvelles stratégies de recherche émergent. Diverses dans les informations et les éclairages qu'elles peuvent fournir sur les problématiques contemporaines, les données disponibles sur le Web constituent aujourd'hui de nouvelles ressources pour les chercheurs, incluant les criminologues, mais également pour les milieux plus pratiques.

La collecte, le stockage et l'utilisation des données disponibles sur le Web ne sont toutefois pas sans défis, notamment en raison des évolutions techniques constantes et rapides qui animent l'Internet. Diverses dans leurs formats – HTML, CSV, JSON, etc. – comme dans leur degré de structuration – données brutes, semi-structurées, voire non structurées –, les données disponibles sur le Web nécessitent l'apprentissage d'une série de stratégies et de méthodes avant de pouvoir être mobilisées.

Face à ces données numériques, notre journée de formation se propose d'introduire les participants aux fondements de la collecte de données sur le Web. La formation ne requiert pas de connaissances fondamentales, même si une connaissance minimum en informatique constitue un atout certain. La formation s'adresse à un public souhaitant comprendre les fondements généraux de la collecte de données en ligne et apprendre les étapes initiales d'une stratégie de collecte de données numériques.

La première section de la journée (A) aura pour objectif de présenter la variété de formats de données auxquels nous sommes confrontés dans toute volonté de collecte de données sur le Web. Après avoir introduit les formats de données du Web, la formation présentera les grandes lignes du langage de programmation *Python* et son utilisation dans le cadre de stratégies de collecte de données en ligne. Les sections d'avant-midi et de début d'après-midi (B & C) aborderont plus en détail la collecte de données à partir de *Python* et de modules comme *BeautifulSoup*. La dernière section de l'après-midi (D) sera consacrée à la collecte de données sur *Twitter* et aux multiples stratégies de collecte de données applicables sur ce service.

Prérequis :

- Avoir accès à un ordinateur portable;
- Installer une série de logiciels et de modules de programmation nécessaires à la formation (la liste et les instructions vous seront communiquées prochainement);
- Être en mesure de préparer les bases de la formation à partir d'une série de lectures et vidéos (la liste et les instructions vous seront communiquées prochainement).

Inscription obligatoire : https://eventbrite.ca/event/16976995651

Programme préliminaire

9h-9h45	*Présentation*
	i. Présentation des objectifs et du déroulement de la journée; ii. Vérification d'installation des logiciels et modules nécessaires; iii. Introduction : pourquoi collecter des données en ligne? iv. Panorama général des stratégies et des outils disponibles; v. Problématiques légales et éthiques de la collecte de données en ligne.
9h45-10 h30	*A. Les bases : formats de données en ligne et Python*
	i. Comprendre les formats de données en ligne (HTML, XML, JSON, YAML…) ii. *Python* : introduction générale iii. Pourquoi utiliser *Python* pour collecter des données? (Avantages et inconvénients) iv. Prise en main de *Python* et comprendre le langage v. Exemple pratique n° 1
10h30 - 10h45	*Pause*
10h45-12h	*B. Collecter les données en ligne I*
	i. *Python* et les modules pour la collecte de données sur le Web; ii. Présentation du module *BeautifulSoup* : introduction et bases; iii. L'utilisation du module *BeautifulSoup* pour collecter des données en ligne (HTML, XML); iv. Exemple pratique n° 2;
12h - 13h	*Repas*
13h00-14 h30	*C. Collecter les données en ligne II*
	i. Retour sur l'exemple pratique n° 2 ii. Approfondissement des fonctionnalités du module *BeautifulSoup* iii. Exemple pratique n° 3
14h30-14 h45	*Pause*
14h30 - 17h	*D. Collecter des données sur Twitter*
	i. Stratégies et avenues de collecte de données sur *Twitter* ii. Comprendre l'API *Twitter* et son fonctionnement iii. Compte développeur et dispositifs d'authentification iv. *Tweepy* : Introduction et bases v. Les stratégies de collecte de données sur *Twitter* à partir de *Tweepy* vi. Exemple pratique n° 4