

TP – WEB CRAWLERS (cours WEB SCIENCE)

Un crawler Web pour extraire informations à propos des clients pour les restaurants à utiliser pour la publicité

1. Choisir un site web où les clients des restaurants proposent des évaluations :
 1. TripAdvisor (www.tripadvisor.com)
 2. Yelp (<http://www.yelp.com/>)

Choisissez une ville.

2. Développer un crawler Web pour adresser les étapes suivantes :
 1. Sélectionner les top 500 restaurants de la ville sélectionnée ;
 2. Pour chaque restaurant, sélectionner tous les utilisateurs à avoir écrit des reviews pour ce restaurant ;
 3. Pour chacun de ces utilisateurs, sélectionner leur nationalité ;
 4. Retourner le classement des nationalités les plus actives dans la ville sélectionnée ;
3. Écrire un rapport de au moins deux pages, qui décrit avec précision toutes les étapes que vous avez parcouru, les stratégies que vous avez choisi dans chaque étape, et les résultats obtenus (résultat du classement et performances). Le TP peut être fait par deux personnes, mais en retournant un classement supplémentaire au classement des nationalités.

4. TP rendu : rapport et code source.

5. Date limite : vendredi 13 juin.

Envoyer par mail aux adresses : elena.cabrio@inria.fr ; serena.villata@inria.fr