**TP EXTRACTION D'INFORMATION DU WEB (cours WEB SCIENCE)**
**Extraction d 'information structurée a partir des annonces immobiliers sur le Web**

1. Choisir un site web d'annonces immobiliers
   1. CraigList (cotedazur.fr.craigslist.fr)
   2. PAP (http://www.pap.fr)

2. Extraire au moins 15 textes des annonces, et les sauvegarder dans des fichiers textuels. 5 textes serons utilisés comme set de développement, et les autres serons utilisés comme test set.
   Vous pouvez les extraire en choisissant une des stratégies suivantes :
   1. Web Crawler (extraire les textes automatiquement des pages web, par exemple en utilisant la librairie java http://jsoup.org/)
   2. En copiant les textes des sources HTML de la page

3. Analyser les annonces du set de développement pour identifier au moins 10 caractéristiques que vous jugez relevants pour décrire les biens immobiliers (par exemple : le type de bien, le prix, la surface, le nombre de pièces, etc.)

4. Générer un fichier CSV (how ? http://www.computerhope.com/issues/ch001356.htm), ou Excell, pour stocker les informations relevants de chaque annonce du test set, par rapport aux caractéristiques choisies. L'extraction doit être automatique, en utilisant un des méthodes expliqués dans le cours :
   1. Hand written patterns (patrons, regex écrites a la main)
   2. Classifiers (en utilisant des algorithmes d'apprentissage automatique)
   Des outils de TAL (Stanford parser, Tree Tagger) peuvent être utilisés pour tokeniser, lemmatiser, ou détecter les Part-of-Speech du texte.

5. Résultat attendu :

| *Reference annonce* | *Type de bien* | *Prix* | *Surface* |
|---|---|---|---|
| 1234 | Appartement | 230000 euros | $60\text{m}^2$ |

6. Évaluation :
   1. Création du goldstandard : annoter les annonces de test manuellement, en utilisant les balises xml, comme il suit :
      *Superbe <TYPEDUBIEN> appartement </TYPEDUBIEN> standing <SURFACE> 73 m²</SURFACE>*
   2. Calculer la précision, le rappel et la F-measure du tableau obtenu grâce au votre extracteur d 'information, par rapport aux annotations correctes que vous avez annoté manuellement (le goldstandard).

7. 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.

8. **TP rendu : rapport, fichier CSV, fichier goldstandard. Date limite : lundi 2 juin. Envoyer par mail aux adresses : elena.cabrio@inria.fr ; serena.villata@inria.fr**