TP WebScience: Structured Information extraction from real estate offers on the Web

1. Choose a website of real estate offers, e.g.:
    1. CraigList (cotedazur.fr.craigslist.fr) (in French)
    2. PAP (http://www.pap.fr) (in French)
    3. Any international real estate on the web (in English)

2. Extract at least 15 texts containing offers, and save them in .txt files. 5 texts should be used as development set, and the others as test set. You can extract them using one of the following two strategies:
a) Web Crawler (automatic extraction from Web pages, using for instance the following Java library http://jsoup.org/)
b) Copying the html source of the page

3. Analyze the offers of the development set to identify at least 10 features that you judge as relevant to describe the real estate (for instance, its type -house/apartment-, the price, the surface, the number of rooms, etc.)

4. Create a CSV file (how? http://www.computerhope.com/issues/ch001356.htm), or Excell, to store relevant information for each offer of the test set, corresponding to the features chosen before. The extraction phase should be automated, using one of the methods explained during today class:
    1. Hand written patterns (manually written regex)
    2. Classifiers (machine learning algorithms)

NLP tools (Stanford parser, Tree Tagger) can be used to tokenize, lemmatize or PoS detection.

5. Expected result:

Reference offer	Type of real estate	Price	Surface
1234	Apartment	230000 euros	60m²

6. Evaluation:
    1. Goldstandard creation: you need to manually annotate the offers in the test set using xml tags as follows:
    Superbe <TYPEDUBIEN> appartement </TYPEDUBIEN> standing <SURFACE>73m²</SURFACE>
    2. Apply your algorithm on the same dataset (not annotated)
    3. Calculate precision, recall and F-measure on the table obtained from your IE algorithm with respect to the goldstandard you have manually annotated, to see how well you algorithm performs.

7. Write a 2-pages report describing the different steps you have carried out, the strategies you have chosen and the obtained results (with some error analysis).

8. **TP rendu** : report, CSV file, goldstandard, code. Deadline: **Tuesday 24 Mai**. To be sent to **elena.cabrio@unice.fr**.