**Web Search, Renmin University of China**

# Lab 3: non-programmers

Pierre Senellart (pierre@senellart.com)

15 July 2011

Labs can be made individually or by groups of two. Make sure to send by email to pierre@senellart.com, either at the end of the lab session or at the latest at midnight of the same day, an informal report about what you did in the lab, and the results you obtained. You do not have to finish all exercises, if you advance reasonably during the lab session but do not go to the end of the assignment, you will still get a passing grade.

## 1 Hearst Patterns

1. Consider the four following Hearst patterns:
   - $X$ is a $Y$
   - $Y$s, such as $X_1$, $X_2$, ...
   - $X_1$, $X_2$, ... and other $Y$
   - many $Y$s, including $X$,

   For each of these patterns, and for each of the following four types:
   - country
   - island
   - rock singer
   - search engine

   use the "*" advanced operator of Google for looking for instances of these types. Extract the corresponding instances, and compute the precision of the top-10 results for each of these 16 combinations.
2. Come up with two Hearst patterns for Chinese, and repeat the same evaluation on the four types (translated into Chinese) of the previous question.

## 2 Instance Extraction with POS Tagging

Have a look at the POS-annotated corpus for the programmers' assignment. Try to come up with extraction rules (either written as regular expressions, or informally defined) to extract instances and their types, using the POS annotations. Apply manually these rules on the first few articles, and compute their precision. You should strive to have a precision of at least 60%.

## 3 Set Expansion

Play with the set expansion algorithm of Google Sets (http://labs.google.com/sets). Find three examples where it is useful; compute the precision each time.

## 4 Google Squared

Google Squared is a Google prototype which apply information extraction techniques to build dynamic tabular results about entities. Columns and lines of tables can be automatically added. Play with it to understand how it works. Use it to obtain the following information:

- The phone numbers of Chinese universities;
- French cheeses with a soft texture made out of cow milk;
- Doctoral advisors of Turing award winners.

How many results do you get this way? Try using a classical search engine to get the same results, does it work?

## 5 Exploring Yago (by Fabian Suchanek)

YAGO is a large ontology obtained using information extraction technologies, that is being developed by the Max-Planck Institute for Informatics in Germany. Go to the YAGO Web site, `http://mpii.de/yago`, click on the "Demo" tab and start the textual browser. This browser allows navigating through the YAGO ontology.

1. Choose a person of public interest. Type the name in the box and hit ENTER. Why do you not land directly at the entity itself?
2. Click on the link with the entity (on the right). Follow the `type` and `subclassOf` links to the root class. Note down all classes you encounter on one path to the root, hand them in.
3. Find at least one incorrect statement in the ontology.

## 6 OpenCalais

Pick an English-language news article of your choice and submit its content to the OpenCalais information extractor (`http://viewer.opencalais.com/`). For each of the annotations made by the system, suggest a technique that has probably been used to extract the information.

## 7 Wikipedia IE

Pick a Wikipedia article, not too short. Explain which information can be extracted in an easy manner from it, describing the technique(s) that can be used for this purpose.