# NLP lab 2

## Textual Information Retrieval

We are going to discover some practical aspects of textual information retrieval (IR) via some on-line tools.

**Exercise 1. (Phonetic search)**

**a.** Search with the French Google (www.google.fr) the incorrect word ***povr***. Can you find any entry citing the correctly spelled word *pauvre*?

**b.** Go to the site of *Windex* tool (http://www.lug.com/) which offers a search engine including the phonetic search. Select *Decouvrir* → *Examples* and choose Recueil des Fables de Jean de La Fontaine. After having checked *Recherche phonétique* search for the keyword ***povr***. Examine the results. When is the phonetic search useful?

**Exercise 2. (Question answering)** We are going experiment with Web browsers using more or less linguistic knowledge.

**a.** Ask the question *Where was John Lennon killed?* to the Google browser (www.google.com).

i. What is the rank of the relevant page (where the answer to the question can be found)?

ii. Among the best rank results you'll find a sentence where "John Lennon" and "killed" appear closely one to another. Whose death is described in this sentence?

**b.** Ask the same question to the http://www.factbites.com, a browser that performs content analysis of the analyzed pages.

i. What is the form of the answers with respect to Google?

ii. What is the rank of the relevant answer?

**c.** Ask the question *When did Marco Polo travel to China?* to the same tools. Where did you find the correct answer?

**Exercise 3. (Vocal IR)** *This experiment can be done if you have headphones at hand.* The idea behind the vocal IR is to enter a keyword on the keyboard and to find audio records containing this keyword.

**a.** Go to the *Voxalead* demo page: http://voxalead.labs.exalead.com/ and tape *tour Eiffel*. Listen to the results and navigate within them.

**b.** Try other keywords in English, e.g. *black market*.

**c.** When is the vocal IP useful?

**Exercise 4. (Intelligent Web browsers)** Intelligence in Web browsing can mean different things. One of them is a "two-dimensional" search. When a keyword is searched two results are shown: (i) pages containing the keyword, (ii) topics to which the keyword is connected.

**a.** Search Google or Yahoo for the keyword *hereditary disease*. The results are one-dimensional: they consist only of pages containing *hereditary disease*.

**b.** Go to the *Gigablast* website (http://sitesearch.gigablast.com/) and search for the same keyword.

i. What are the notions connected to *hereditary disease*?

ii. What is the relation between *hereditary disease* and *monogenic disorder*?

**c.** Go to the *Exalead* search engine (http://www.exalead.fr) and search for the same term. On the right hand side you'll see the associated term. Which three of them are synonyms of *hereditary disease*.

**d.** Why is the 2-dimensional search more useful than a "flat" search?

**Exercise 5. (Term acquisition)** The contents of a document may be characterized by a set of terms that are most relevant to it.

**a.** Copy the *NLPforIR.pdf* file, which is an introduction to natural language processing for information retrieval, from *P:\public_m2\TAL\savary*.

**b.** Go to the *TermExtractor* demo page http://lcl2.di.uniroma1.it/termextractor/ and apply the online demo to the .pdf file.

**c.** Examine the list of extracted terms and judge which of them are good descriptors of the text domain.

**d.** Do the same for another term extractor: http://labs.translated.net/terminology-extraction/. Compare the results to the previous list.