IST 2018/2019

Processamento e Recuperação de Informação

Lab 04: Information Extraction

The Python extension package named `nltk`[1] provides a set of tools that are useful for addressing information extraction problems such as Named Entity Recognition (NER). More specifically, you can use the following methods:

- `nltk.sent_tokenize(d)`, which splits a document d into a list of sentences;
- `nltk.word_tokenize(s)`, which splits a sentence s into a list of words;
- `nltk.pos_tag(w)`, which leverages a sequence classification model to tag the words in list w according to their part-of-speech (i.e., tag words according to morphosyntactic classes such as noun, verb, adjective, ...);
- `nltk.ne_chunk(p, binary=True)`, which tags the words in list p as named entities or not (where each word in p was previously tagged with a part-of-speech tag).

Note that the output of each of these tools can be used as input to the next tool.

1

Test the tools with a few sentences of your own, or extracted from Web sites. Try text from different contexts (e.g. news, blogs, etc.).

Notes:

- These tools are trained for the English language, and thus they do not perform well on other languages.
- Take a look at the output format of each tool, so that you can use it on the next exercise.

2

Using the above tools, print all named entities found in the documents of the 20 newsgroups collection[2]. This document collection can be conveniently accessed through the scikit-learn library, as shown in the previous lab class.

[1] http://www.nltk.org

[2] http://qwone.com/~jason/20Newsgroups/

3 Pen and Paper Exercise

Consider the Hidden Markov Model represented by the following probabilities. Remember that π corresponds to the initial probabilities of each state, B corresponds to the state emission probabilities, and A corresponds to the transition probabilities.

The symbols corresponding to each line in matrix B are a, b, and c.

$$\pi = \begin{pmatrix} 0.8 & 0.2 \end{pmatrix} B = \begin{pmatrix} 0.1 & 0.6 \\ 0.7 & 0.2 \\ 0.2 & 0.2 \end{pmatrix} A = \begin{pmatrix} 0.1 & 0.5 \\ 0.9 & 0.5 \end{pmatrix}$$

(a) Compute the total probability of occurrence for the sequence **acbc**.

(b) What would be the probability for the sequence **acbc** occurring, if the sequence of states was known as being **1212**.

(c) What is the most likely sequence of states for the sequence of symbols **acbc**?

(d) Starting from the model defined above, compute a new model $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$ using one iteration of the Baum-Welch method, assuming that you had only one observation available: **acb**.