ASSIGNEMENTS FOR NLP COURSE

Students shall implement one among the following possible choices:

1) General assignment;
2) Research-oriented assignment.

General assignments need to be implemented individually. Research-oriented ones can be implmented in groups of maximum three students.

GENERAL ASSIGNEMENT

Implement a python technology employing NLTK (or alternatively, a JAVA technology, employing OPENNLP) that, given a corpus (suggestion for test, the reuters repo nltk.corpus.reuters), executes one and only one of the following tasks:

a) Clusters the corpus in a specified number of classes based on the cosine similarity measure by a classical clustering method, such as K-nearest Neighbours;

b) Asks for a file of keywords to the user, and implements the probabilistic match of keywords on the documents while measuring the tf-idf index of each of the elements in the corpus itself. The feature returns the same list of the keywords with associated their tf-idf index, and marks them in three classes (top, medium, bottom) based on their percentiles in the index itself (use the classical 10-80-10 distribution);

c) Asks for a document to the user, and a match percentile, and returns the documents of the corpus that match the documenta bove the percentile as similar by means of the cosine similarity of the tf-idf (the match has to be on complete docs, no application of the stopword elimination phase).

RESEARCH-ORIENTED ASSIGNMENTS

1. Implement a graphic pipeline for document processing on corpora and single documents;

2. Implement a graphic tool for deploying the following actions:
   2.1 Eliminate stopwords;
   2.2 Lemmatize terms;
   2.3 Compute frequencies;
   2.4 Measure distances from strategic points (start and end);
   2.5 Compute compound relevance indices (50% frequency and 50% earliness).

3. Implement a technology that relates documents based on reference (a document refers things told in another document).