Laboratory Project of INLP for the course 2015-2016.

The Laboratory Project of INLP for this course consists of building a system for parsing geographically oriented queries.

In 2007, within the framework of GEOCLEF-2007 challenge, a task was organized for evaluating the parsing of geographically oriented queries. The task was organized and evaluated by Microsoft Research Asia. We attach a paper[1] describing the task and the results obtained by the participants in it (a team from TALP-UPC participated in the task obtaining medium results (3rd from 6 participants).

We propose a laboratory case based on this task.

The description of the task and the evaluation metrics are the same described in the attached paper. The only difference is that we have reduced the dataset to be evaluated.

The training corpus consists of 100 queries and is attached (GQ_Tr_100.xml). The test corpus consists also of 100 queries and will be provided one week before the deadline of the laboratory project (GQ_Test_100.xml). The only difference in the format of both files is that in the later the answers fields are obviously not included (after the deadline a file, GQ_Test_with_answers_100.xml, with answers will be provided).

Students should organize themselves into groups of two or three members. Tasks are organized into two sets that have to be delivered in two dates (the first one month after the start of the course and the second at the end of the course). The first set consists of building the infrastructure for the task and will be faced in a collaborative way from all the groups (perhaps more than one group will produce some of the programs) in a way that the programs produced will be (after being evaluated) made available to all the groups. The idea is that common modules, probably needed for everybody, would be built first, with limited effort (only one module per group) and made available to all the students.

Tasks to be done for the first set:

- Write in python[2] a corpus accessor for accessing the .xml file and building a more suitable representation of all the queries (the accessor could be used for accessing both training (with answers) and test (without) files).
- Write in python a program for saving the representation structure of a set of queries into a .xml file
- Write in python a scorer that allows to score a query (and a set of queries) from its internal representation according to the evaluation criteria contained in the paper. The scorer will compare two datasets, one corresponding to a dataset produced by a group and a golden dataset.
- Write in python a program for generating some baseline datasets (in order to get lowerbound scores for the sake of comparison.

[1] LI_OverviewCLEF2007.pdf. Zhisheng Li, Chong Wang, Xing Xie, Wei-Ying Ma, “Query Parsing Task for GeoCLEF2007 Report”

[2] https://www.python.org/

- Write in python an interface for accessing to Freeling[3].
- Write in python an interface for accessing DBpedia[4] (in order to check whether a string could correspond to a location (or a city, a country, a state, etc).

Each group will be assigned one of these tasks and are expected to provide:

- The corresponding python program
- A short (1 page) description of the approach

Task to be done by all the groups (second deliverable):

- Write a grammar (or a set of grammars, one for each query component) for facing the task of parsing a geographically oriented query, according to the description in the paper. You are free to use the formalism you consider more appropriate (FSA, CFG, DCG, … probabilistic or not). You can write the grammar manually or learn it from the training corpus. If you need to tokenize the texts you are free to use whatever tokenizer, morphological analyzer, POS tagger, NE recognizer, etc. We encourage the use of Freeling (through the interface built in our first set of tasks). If you need to access some geographical gazetteer we encourage to use the use of DBpedia (through the interface built in our first set of tasks).
- Include your grammar into a python program for analyzing a dataset.
- Apply the scorer to your output using the golden dataset

Each group is expected to provide:

- The grammar (or grammars)
- The python program
- The results
- A short (max 5 pages) description of the approach. You should include an analysis of your errors and ideas for avoiding them (just ideas, no implementation is needed!!!)

Barcelona, September 2015

[3] http://nlp.lsi.upc.edu/freeling/

[4] http://wiki.dbpedia.org/