Cmpe 493 Introduction to Information Retrieval, Spring 2020
Term Project - Named Entity Normalization

In this project, you will work on the Bacteria Biotopes Entity Normalization (BB-norm) task (https://sites.google.com/view/bb-2019/home) addressed at the BioNLP Open Shared Tasks 2019 (BioNLP-OST:http://2019.bionlp-ost.org).

In the BB-norm task, a data set containing biomedical text from scientific publications, where the Microorganism, Habitat, Geographical, and Phenotype named entities have already been annotated, is provided. Your goal in this project will be to normalize the Habitat named entities to their corresponding concepts in the given OntoBiotope ontology.

The training, development, and test sets as well as the OntoBiotope Ontology are available at https://sites.google.com/view/bb-2019/dataset. A link to an evaluation software is also provided. Please ONLY use the training and development data sets for developing your systems. The gold standard annotations for the test set are not available. You should use the online evaluation tool to obtain your final scores on the test set in the end of the semester (http://bibliome.jouy.inra.fr/demo/BioNLP-OST-2019-Evaluation/index.html). You should NOT use the test set throughout the semester to develop/tune your systems.

Please make sure that you use the data sets for the *BB-norm* task and refer to the shared task web site (https://sites.google.com/view/bb-2019/home) for the details of the task, the data sets, the annotation and evaluation guidelines.

Deliverables:

1. Project progress presentation (May 11, 2020 (in the lecture hour); 30% of your project score): You should prepare a 10min presentation describing what you have done so far and what your plan is for the remaining time period. You should have completed at least the preprocessing of the data set and implemented and tested a baseline approach (such as simple exact matching). You should also have clear plans about how you will improve your system by the end of the semester.

2. Project final presentation (On the final exam date/slot; 70% of your project score): You should prepare a 15min presentation describing your final system and your results. I also suggest you to include an error analysis.

3. Prior to each presentation (latest 1 hour before the presentations start) you should send me by email your slides and all source code and accompanying readme documents.

Honor Code: You should work in teams of two or three people. Each team member should contribute equally to the development of the project and to the presentations. All team members will get the same score. You are allowed to use external libraries/resources for the project. However, you SHOULD properly acknowledge and cite these in your presentations and source code.

Late Submission: Late submissions are NOT allowed.