Cmpe 493 Introduction to Information Retrieval, Fall 2021
Term Project - Text Classiification for Covid-19 Scientific Literature

---

In this project, you will work on Track 5 of BioCreative VII, namely the LitCovid track on Multi-label topic classification for COVID-19 literature annotation (https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/).

The aim of the LitCovid track is to develop a text classification system that will assign an input scientific article into one or more of the following seven classes: Treatment, Diagnosis, Prevention, Mechanism, Transmission, Epidemic Forecasting, and Case Report. A training set of 24,960 articles and a development set of 6,239 articles manually annotated by human experts are provided at the following link: https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/biocreative/. The training data is included in the BC7-LitCovid-Train.csv file and the development data is included in the BC7-LitCovid-Dev.csv file. Note that a test set is also provided (BC7-LitCovid-Test.csv). However, we will NOT be using it, since the labels for the test set are not available. A readme file including further information about the task and the data sets is available at the BC7-LitCovid-Readme.pdf file.

Please use the training set for developing (training and tuning your system) and the development data set for evaluation/testing. That is, we will use the development set as a test set. So, you should NOT use the development set for developing/training/tuning your systems.

You should use the evaluation tool provided at https://github.com/ncbi/biocreative_litcovid for evaluating the performance of your systems.

In order to design your systems you are suggested to read the prior work on this task. A set of reference papers is available at https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/ and the papers of the shared task participants presented at the BioCreative VII Workshop are available at https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/.

Deliverables:

1. Project progress presentation (**December 29, 2021 (11:00 o'clock)**; 30% of your project score): You should prepare a 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 (describe how you did the preprocessing and provide summary statistics such as number of documents in each class, number of tokens, number of terms etc.) and implemented and tested a baseline approach (such as Naive Bayes, kNN etc.). You should also provide clear plans about how you will improve your system by the end of the semester. We will not have in-class presentations, rather you will only send me your presentation files. Each team should submit their presentation through Moodle (only one team member should make the submission on behalf of the team. The names of all team members SHOULD be written on the cover page (first slide) of the presentation.

2. Project final presentation (On the final exam date/slot (**January 15, 2022, 13:00 o'clock**); 70% of your project score): You should prepare a 10min presentation describing your final system and your results. I also suggest you to include an error analysis and possible directions

for improvement. We WILL have in-class presentations. Each team should submit the slides and all source code and accompanying readme documents through Moodle. Only one team member should make the submission on behalf of the team. The names of all team members should be included on the first slide (cover slide).

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.