Cmpe 493 Introduction to Information Retrieval, Spring 2023
Term Project - Authorship Attribution

In this project, you will work on the PAN 2019 Shared Task on Cross-Domain Authorship Attribution. The details about the task are available at `https://pan.webis.de/clef19/pan19-web/authorship-attribution.html`. In this task you are given a set of authors and sample text documents written by them. The goal of the task is to determine the author of a given new text document (i.e., a test document).

You should request the data set as soon as possible from `https://doi.org/10.5281/zenodo.3530313` by specifying that you will be using it for your term project in the "CMPE 493 Introduction to Information Retrieval course taught by Arzucan Ozgur at Bogazici University".

The data set is split into problems (each in a separate folder) ranging from problem00001 to problem00020. We will use the first five problems (problem00001, problem00002, problem00003, problem00004, problem00005), which is the English subset of the data set. Each problem will be handled separately. *problem00002* and *problem00004* will be used as a development set while developing your systems and the other three problems will be used as test set to report your results. Each problem0000x folder contains text documents written by *nine* different authors. The documents of each author are included in the *candidate0000x* folders. Note that the authors <u>are different</u> for different problems, that is candidate00001 in problem00001 is <u>not</u> the same author as candidate00001 in problem00002. The *unknown* folder contains the test text documents whose authors have to be predicted. The ground truth (the author of each unknown text) is available in the *groundtruth.json* file. Note that for some of the unknown texts, the ground truth author is specified as $< UNK >$. This means that the author of this text is not one of the nine authors in the provided author set, and your systems are expected to determine that.

The macro-averaged F1-score will be used as the primary evaluation metric. You should use the evaluation tool provided at `https://github.com/pan-webis-de/pan-code/tree/master/clef19/authorship-attribution` for evaluating the performance of your systems for *problem00001, problem00003*, and *problem00005*.

Details about the task, the data set and the participating systems in 2019 are described in the overview paper about the shared task (available at `https://webis.de/downloads/publications/papers/kestemont_2019.pdf`), which I strongly recommend you to read.

Deliverables:

1. Half page project progress report (**April 28, 2023 (23:59 o'clock)**): You should briefly explain 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 implemented and tested a baseline approach. You should also provide clear plans about how you will improve your system by the end of the semester. Each team should submit their report 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 report.

2. Project final presentation (**May 8, 2023, lecture hour**): You should prepare a 10min presen-

tation describing your final system and your results. I also suggest you to include an error analysis and possible directions for improvement. 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.