Data Science Lab: Process and methods
Politecnico di Torino

Project description
September call, A.Y. 2019/2020

Last update: August 19, 2020

1 Competition dates

Start date: August 20, 2020 at 20.00 CEST
Due date: September 20, 2020 at 20.00 CEST

Due date is a **strict deadline**.

2 Problem description

This project consists in the identification of the speaker in short audio recordings. You are required to build a robust classifier capable of distinguishing among a number of different speakers.

2.1 Dataset

The dataset for this project has been extracted from LibriVox, a website providing free public domain audiobooks. These audiobooks are read by volunteers all over the world.

The dataset is comprised of 30,281 recordings extracted from different audiobooks. Each recording lasts approximately 0.5 seconds and is sampled at 24 kHz (for a total of approximately 12,000 samples per recording).

The recordings are collected from a total of 10 speakers, identified by the labels *a,b,c,d,e,f,g,h,i,j*. Each recording is identified by a unique id, a 64-characters hexadecimal string.

2.1.1 Development set

The development set is characterized by 24,449 recordings as WAV files compressed in a single ZIP file. Once extracted, the directory structure is as follows:

```
development/
    a/
        id_a_1.wav
        id_a_2.wav
        ...
    b/
        id_b_1.wav
        id_b_2.wav
        ...
    ...
    j/
```

```
        id_j_1.wav
        id_j_2.wav
        ...
```

The name of each subdirectory defines the label of the speaker whose recordings are contained within the respective folder.

The "`id_*_*`" syntax is used as a shorthand for the 64 characters unique identifiers.

2.1.2 Evaluation set

The evaluation set is comprised of 5,832 WAV recordings compressed in a single ZIP file. Once extracted, the directory structure is as follows:

```
evaluation/
    id_1.wav
    id_2.wav
    ...
    id_XXX.wav
```

The "`id_*`" syntax is used as a shorthand for the 64 characters unique identifiers.

Download You will be redirected to a Microsoft OneDrive directory. **Use your account credentials from Politecnico di Torino to access the material.** The dataset is available at:

`https://bit.ly/DSL1920_dataset_sept`

2.2 Task

The task for this project is to build a classification model capable of identifying the speaker of each recording in the evaluation set.

Please note that the recordings for the evaluation set have been extracted from different chapters/books than those in the development set. As such, overfitting will be particularly penalized in this context (a model that learns to identify the "book" rather than the speaker will have troubles generalizing to never-before-seen recordings).

2.2.1 Evaluation metric

The evalution metric used will be the macro F_1 score (`average=macro` as a parameter of scikit-learn's `f1_score`).

3 Submissions

The solution needs to be uploaded on the evaluation platform as a CSV file. The columns required are two, *Id* and *Predicted*. The first column contains the 64 characters identifier, while the second column contains the predicted label (i.e. a letter from *a* to *j*). The following is an example of the first 10 lines of a submission file (you may find a complete submission file example along with the dataset 2.1.2).

```
Id,Predicted
dfab30d7761cb9128284920eb2088fbf4f84be5aa0d48b60cb81beb5c8a17b99,i
55efe35a9997cee4afbca75001d82a2bdc14b668dd7f65800b67a6e85ca02a48,g
5fa19536788e0ad47b422affe11ef0f7175df551874b0be465681a3d9ee250b7,b
a7183e90defe318a75d42815a1b7b6c7991db9395cbce08b527f135c490ebba2,b
e259a29f18a0a2a09275eee62d1772d2e6a8effa430c7610e22ecdffd6e1d058,a
fa69cd8622c52ecb5d8a0307cb232ebb8af835210c01b15e0ea3fef5dfc5cb98,g
8773d931d0eeb0074eb3442c967c9d731b2aa87c5c89952ba369f70f15d976ef,j
```

`bca6b48fbf2c5156462faa78b9b4d96e82f3e51e132a5a6d36948b7bf4b5830d,d`
`e5168e107289ea2653dca4982ea36a191142d5dc24d4d6ff93885bb92892281d,d`

Submission platform The submission platform is the same you used during the course laboratories. Therefore, you have to use the same key. Please refer to the guide on the course website, to go through the submission procedure.
You can find the competition platform at http://35.158.140.217/

3.1 Upload the report and the software

Warning: The report and the software have to be submitted by the due date reported in Section 1. This is a **strict deadline**.

All the required files (i.e. for the report and the software) must be included in a **single *.zip*** file. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section.
Please use as description: **report_exam_september_2020**.

Formatting rules The formatting rules for both the report and the software are described in a dedicated document. You can find it on the course website.