Data Science Lab: Process and methods
Politecnico di Torino

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

Last update: June 22, 2020

1 Competition dates

Start date: June 17, 2020 at 20.00 CEST
Due date: July 17, 2020 at 20.00 CEST

Due date is a **strict deadline**.

2 Problem description

In this competition, you are required to construct a sentiment analysis pipeline on contents from the Twitter social network. In practice, you are provided with a set of tweets posted by several real users, and your goal is to predict whether they support the Black Lives Matter movement or not.

2.1 Dataset

Considering the week between May 27 and June 5, we collected every tweet matching **all** the following criteria:

- the tweet contains only plain text (i.e. there is no media content attached);
- the text of the tweet is written in English;
- the text of the tweet is not offensive under the social network's policies;
- the text of the tweet contains at least one of the keywords related the Black Lives Matter movement.

The dataset for this competition consists of 100,000 tweets sampled from the original collection. Specifically, the tweets are organized into textual files with multiple lines. Each line contains a tweet encoded as a JSON string. The tweets are characterized by multiple attributes. Some of the most important are:

- created_at: UTC time when this Tweet was created, as a string.
- id: The string representation of the unique identifier for this Tweet.
- full_text: the actual UTF-8 text of the status update, as a string.
- user: the user who posted this Tweet, as a JSON string.
- retweet_count: number of times this Tweet has been retweeted.
- favorite_count: *Nullable*. Indicates approximately how many times this Tweet has been liked by Twitter users.

- coordinates: *Nullable.* Represents the geographic location of this Tweet as reported by the user or client application, as a JSON string.
- place: *Nullable.* When present, indicates that the tweet is associated (but not necessarily originating from) a Place, as a JSON string.

For the complete list of features, please refer to the official Tweet Object documentation.

The last attribute of each tweet corresponds to the class label. It can get one of the following values:

- **1**: if the tweet shows a positive sentiment towards the movement.
- **0**: if tweet shows a negative sentiment towards the movement.

Note on loading datasets organized in JSON lines Whenever a dataset is provided, like in this case, as a textual file with one JSON string per line, it can be easily loaded with pandas as follows.

```
import pandas as pd
data = pd.read_json('dataset.jsonl', lines=True)
```

Also, note that this loading method parses every nested JSON string into a Python dictionary for you.

Dataset tree hierarchy The dataset for this competition is split into separate collections. Each collection is in a different file. The dataset archive is organized as follows:

- `development.jsonl`: (Development set) a collection of tweets **with** the class attribute. This collection must be used during the development of the classification model.
- `evaluation.jsonl`: (Development set) a collection of tweets **without** the class attribute. This collection must be used to produce the submission file.
- `sample_submission.csv`: a sample submission file.

Download The dataset is located at:

`https://bit.ly/DSL1920_exam_summer_dataset`

2.2 Task

You are required to build a classification pipeline to assign a label to each record in the Evaluation set.

2.3 Evaluation metric

Your submissions will be evaluated on the accuracy_score with the following configuration:

```
from sklearn.metrics import accuracy_score
accuracy_score(y_true, y_pred)
```

3 Submit your result

Submission file In order to get your results evaluated, you have to upload a result file on our submission competition platform. The submission file has to be a `.csv` file formatted as follow:

```
Id,Predicted
10,0
123,0
21,1
345,1
42,0
...
```

The submission file must contains a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set, as an integer number.

  Info: Note that the `Ids` in the submission file must correspond to the positions of the records in the Evaluation set. **The first record in the Evaluation set has Id=0, the second has Id=1 ans so on.**

- the `Predicted` label for the corresponding record.

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**.

Submission Thanks to the novel technologies supported by the Politecnico di Torino, we are evaluating the possibility to use Exam/Exercise to redact the report and upload software. All the submission guidelines will be available soon on the course website. By that time, this document will be updated, and you will be notified by email.