Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
January Call, A.Y. 2021/2022

Last update: December 29, 2021

1 Project dates

Start date: December 29, 2021 at 23:59 (CET)
Due date: January 20, 2022 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Different people express different sentiments using different words. This phenomenon is particularly evident on online social platforms. Further, as psychological studies reveal, sentiment and emotions vary over a broad spectrum and are characterized by taxonomies.

In this project, you are required to predict the sentiment of a given content on Twitter. In a simplified manner, the sentiment provided in this task is either positive or negative.

2.1 Dataset

Warning: For this project, you are not allowed to use external datasets other than the one provided. Adoption of external resources will result in failure of the exam.

The dataset consists in a collection of tweets in tabular format. Each record is characterized by several attributes. The following is a short description for each of them.

- *ids*: a numerical identifier of the tweet;
- *date*: the publication date;
- *flag*: the query used to collect the tweet;
- *user*: the username of the original poster;
- *text*: the text of the tweet.

The sentiment of the tweet is reported on the feature named ***sentiment*** and is equal to **1** for the Positive class and **0** for the Negative one.

The dataset is located at:
https://dbdmg.polito.it/dbdmg_web/wp-content/uploads/2021/12/DSL2122_january_dataset.zip

Within the archive, you will find the following files:

- **development.csv** (development set): a comma-separated values file containing the records from the development set. This portion does have the `Sentiment` column, which you should use to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the records from the evaluation set. This portion does not have the `Sentiment` column.
- **sample_submission.csv**: a sample submission file.

2.2 Task

You are required to build a classification pipeline to predict the sentiment of tweet in the Evaluation set.

2.3 Evaluation metric

Your submissions will be evaluated through F1 score (Macro).

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
10,2
123,0
21,23
345,12
42,3
...
```

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: 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.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `giuseppe.attanasio@polito.it`. Please refer to the guide on the course website, to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

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

Submission All the required files (i.e. for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_january_2022**.

Info: A ZIP archive is a ZIP archive, not a RAR, a 7z or, a tarball archive, nor any of those renamed with a trailing `.zip` extension.

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