Data Science Lab: Process and methods
Politecnico di Torino

Project description
Winter calls, A.Y. 2019/2020

Last update: January 11, 2020

1 Competition dates

Start date: January 12, 2020 at 00:01 AM
Due date: January 26, 2020 at 00:01 AM

Due date is a **strict deadline**.

2 Problem description

In this competition, you have to perform a sentiment analysis task, analyzing user's textual reviews, to understand if a comment includes a positive or negative mood.
In practice, you are required to build a robust classification model that is able to predict the sentiment contained in a text.

2.1 Dataset

The dataset for this competition has been specifically scraped from the tripadvisor.it Italian web site. It contains 41077 textual reviews written in the Italian language.
The dataset is provided as textual files with multiple lines. Each line is composed of two fields: `text` and `class`. The `text` field contains the review written by the user, while the `class` field contains a label that can get the following values:

- `pos`: if the review shows a positive sentiment.
- `neg`: if the review shows a negative sentiment.

Dataset tree hierarchy The data have been distributed in two separate collections. Each collection is in a different file.
The dataset archive is organized as follows:

- `development.csv` (Development set): a collection of reviews **with** the class column. This collection of data has to be used during the development of the regression model.
- `evaluation.csv` (Evaluation set): a collection of reviews **without** the class column. This collection of data has to be used to produce the submission file.
- `sample_submission.csv`: a sample submission file.

The dataset is located at:

`http://dbdmg.polito.it/wordpress/wp-content/uploads/2020/01/dataset_winter_2020.zip`

2.2 Task

You are required to build a classification pipeline to assign a label to each record in the Evaluation set. The label specifies the sentiment of the review.

2.3 Evaluation metric

Your submissions will be evaluated exploiting the f1_score with the following configuration:

```
from sklearn.metrics import f1_score
f1_score(y_true, y_pred, average='weighted')
```

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,pos
123,pos
21,neg
345,pos
42,neg
...
```

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 corrisponding 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 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_winter_2020**.

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