Task 12 TAG_it

Paolo Renzi
Sapienza Università di Roma

1 The description of the dataset

The TAG-it dataset, originates from the EVALITA 2020 competition, aims to explore the relationships among gender, age, and various topics across different blog authors. Originally, the dataset was designed to capture these interactions, with a focus on longer texts from blog posts. However, for the evaluation task, a modified version of the dataset is employed, concentrating solely on shorter posts to accurately classify the topics of grouped posts. This approach allows for a more focused examination of the relationship between author characteristics and topic classification.

2 Methodology to reframe the dataset

I exploited the fact that each row of data ends with the name of the dataset and used that to separate each line. In the test set the data is taken from two datasets so the name changes accordingly but i still used the same approch by first dividing for political_test and then I iterated in each line to divide it again by Religious_test and put it all in a list. I then iterated over the resulting list and created a dictionary where the keys are the id of the tweet and the values are themselfs a dictionary with as key the type of data (Ex. "text", "label", "created at") and the values is the data. I differentiate between test and training and textual data and conte data from the filepath string.

3 Prompts

I created 4 prompts for each task, and i added in task 2 the following metadata, created at and user_created_at beacause the age of the profile at the time of writing might have a correlation with the task. For the same reason I added status, friend e followers count.

4 How to run

To run the parser you should download all the required libraries (Gitpython zipfie and pandas) set the folder contating task_12.py as CWD and then run

```
python task_12.py
```

You will need to download the dataset manually from this link https://live.european-language-grid.eu/catalogue/corpus/8112/download/, and put it in the same directory as the .py file but .

Table 1: Distractor Relationships

Topic	Related Topics
CELEBRITÀ	MEDICINA-ESTETICA, INTRATTENIMENTO
ANIME	GIOCHI, GIOCHI_DI_RUOLO, INTRATTENIMENTO
FUMO	TECNOLOGIA, OROLOGI
AUTO-MOTO	SPORT, MOTO
SPORT	AUTO-MOTO, MOTO, GIOCHI
MOTO	SPORT, AUTO-MOTO
METAL-DETECTING	NATURA, TECNOLOGIA, AUTO-MOTO
TECNOLOGIA	INTRATTENIMENTO, GIOCHI, AUTO-MOTO
MEDICINA-ESTETICA	CELEBRITÀ
INTRATTENIMENTO	GIOCHI, GIOCHI_DI_RUOLO, SPORT, ANIME
NATURA	METAL-DETECTING
GIOCHI	GIOCHI_DI_RUOLO, INTRATTENIMENTO
GIOCHI_DI_RUOLO	GIOCHI, INTRATTENIMENTO
OROLOGI	CELEBRITÀ

Table 2: Example commands for accented characters, to be used in, *e.g.*, BibTeX entries.