Task 12 TAG_it

Paolo Renzi
Sapienza Università di Roma

Leonardo Colosi
Sapienza Università di Roma

Lukasz Sztukiewicz
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 the data was formatted with <user... to signal the line where there was informations about the user I used this regularity to find the topic with a regualar expression and <post>, </post> to divide each post and </user> to indicate when the posts from a single user end.
I put in a dictionary the posts as values and the topics as keys and if the topics where not divisible by 5 I would add empty strings to fill missing samples

3 Methodology and rationale behind the distractors

We created a relational graph that associate to each topic a list of most similar topics based on human evaluation 1. Then we used a mix of topics from this graph and random ones to generate distractors, making sure to avoid replication that could arise from the random selection. We then shuffled the list of distractors to prevent the creation of a pattern in the indexes of the choice list.

4 Prompts

We created 6 prompts of 2 kinds 3 in free generation style and 3 where the answers were constrained by a list of provided choices.

5 How to run

To run the parser you should set the folder contating task_12.py as CWD and then run

```
python task_12.py
```

You will also need to download the dataset manually, from this link https://live.european-language-grid.eu/catalogue/corpus/8112/download/,, and put it either zipped or unzipped in the same folder as the .py file like so /task_12.py /final_package_train_test/...

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.