1. Data Science Homework

Background

The core of Realeyes' business is interpreting the behavioural connection between consumers and products by analysing their emotional reaction to video advertisements. Thus, we record subjects while they are watching video advertisements and extract emotional and behavioural data at different levels of granularity. Then, the main goal of the Data Science team is to analyse this data to produce insights and predictions about video advertisements that subjects were watching.

This exercise presents you with a real use case of subjects' emotional data. As subjects are being video recorded, data provided is per each captured frame of their face while they watch an advertisement.

Data

The files **realeyes_dsc_homework_video_X.csv** contain all required data for this exercise. Each file contains information for a different video:

- *video_id* and s*ubject_id*, which allow to pair every subject's emotional data with the ad video they were watching.
- *frame_no*, number of the webcam recorded frame for that person.
- *millesecond_from_start*, time since the video started playing, in milliseconds, for each frame.
- The corresponding collection of emotional data assigned to the frame, with one emotion per column. Emotions are classified into positive or negative.
  Note: bear in mind that subjects are recorded with variable recording framerates and consequently the frames are not aligned along time between subjects.

It is important to note that the dataset provided is given on three different levels: frame level, subject level and video level.
The source data is given on the frame level. It contains a *subject_id*, which allows it to be grouped into a subject level. Source data also has a *video_id*, which allows it to be also grouped to a video level.

Task

The goal is to answer the following 3 questions, which are a simplification of a typical Data Science task.

Be aware that **Realeyes' existing production code is in Python**, so that is the preferred coding language for the task, although it is not mandatory. You are free to use external open-source libraries if you wish.

- Question 1: understanding the data

Although we try to keep our data as clean as possible, some issues can arise during data storage or data loading. Besides, the intrinsic nature of the data poses some challenges. For instance, emotion information is automatically produced by an algorithm from facial recordings obtained in live conditions. Therefore, some recordings may be more challenging to "read" for the algorithm.

Load and process the data as you consider and present a high-level overview of the data provided. Please explain the data transformations you made, if any.

- Question 2: characterizing the audience

The Data Science team is trying to decide between three difference ways of summarizing the emotional reactions of viewers to the videos provided. They propose to compute the features for each available frame, and then aggregate them by taking the average for each person.

The proposed definitions to compute the features per frame are:

1. *positive_1 + positive_2 - (negative_1 + negative_2 + negative_3)*
2. *any(positive_1, positive_2) - any(negative_1, negative_2, negative_3)*
3. *any(positive_1, positive_2, negative_1, negative_2, negative_3)*

Which are the pros and cons of each definition? Which one would you pick to characterize the audience of the provided videos and why? You are also free to propose your own definition if you wish.

- Question 3: differentiating between videos

Once you have selected which feature seems the most promising to summarize emotional reactions per person, we want to check whether that feature allows us to distinguish between the three videos.

You can aggregate the feature per video by taking the average of all its viewers. Can you state that the videos are different to each other regarding their emotional reactions?

Deliverables

- Provide explanations, charts and numbers of your findings through the three previous questions.
- For all analysis that you conduct please provide us with code and results.

Recommendations

General

Consider this homework as the opportunity to show us your rigour, coding style, technical skills, knowledge and reasoning, your reporting and explanation skills.

Please also note that your submission will be judged solely based on depth and quality, without considering the time that you actually invest in this task.

Coding style guidelines

- Python is preferred
- Build modular and optimised code if possible
- Try to use an interactive plotting library
- Document all your functions

2. Python Coding Homework (Optional)

This task is not mandatory, although as it was already mentioned Realeyes' production code is written in Python, so Python programming skills are good to have.

Task

The following files, available inside the *"python coding tasks"* folder, contain explanations on the functionality of each script:

1. Warm-up task: *helloworld.py*
2. Small game: *orcs_vs_goblins.py*
3. Extra task: *orcs_vs_goblins_extra.py*

Please start with the warm-up task to understand the dynamic of this homework, and follow the given instructions in the files. Note that you can test your work by running the python file.

Deliverables

Please send back to us the same input Python files containing your code in the designated area.