Master Big Data Analytics and Social Science 2024/2025

Giuseppe Prencipe, Roberto Pellungrini

December 10, 2024

Abstract

Solve the two exercises below, reading carefully all the instructions before starting to code. The exam requires access to two files. They are linked in the Moodle. You can also find the links here:

- DivinaCommedia.txt
- Ted.csv

Upload all your work in a single zip file, named nomeCognome.zip (with nome and Cognome your Name and Last Name, respectively).

Deadline: 31st of December 2024.

1 Exercise 1

One way to recognize the language in which a text written in an unknown language might be is to calculate the frequency of various letters. Different languages have distinct profiles. For example, in Italian, the six most frequent letters are, in order, "eaionl" (thus, the most used letter in Italian is "e"); in French, they are "esaitn"; in Turkish, "aeinrl"; and in Finnish, "aintes", and so on. Of course, a specific text might not strictly follow this sequence, but in sufficiently long texts, the result will be at least very close to the expected one.

Define a Python program structured as follows:

- A Frequency class to handle a text file. Essentially, the class must have (at least) a property called fileToProcess, whose purpose is to store the text file managed by the class (you can obviously include any additional properties that you find useful for designing the class).
  - Assume the text file to be processed could be reasonably long (hence, big in size).
  - Use .txt as the standard suffix for file names. For example: pippo.txt or divinaC.txt
- A guess6 method that analyzes fileToProcess as follows:
  - Reads the file managed by the class (using an appropriate strategy for its length).
  - Returns the string composed of the six most frequent characters in fileToProcess, in descending order of frequency, considering the following rules:
    - Ignore the distinction between uppercase and lowercase letters.
    - Include accented letters.
    - Exclude punctuation and other symbols.
    - Include numbers.
  - In case of a tie in frequency, the characters should be sorted lexicographically.

Use the file divinaCommedia.txt as the reference file, which contains The Divine Comedy in plain text format. Clearly, the program must work with any text file given as input to the function.

- Provide the program in a .py file (not in a notebook).
- Name the file SurnameName.py, with Surname and Name being your last and first names, respectively.

- Provide the result obtained by running the program on the example file divinaC.txt.
- Provide the result obtained on another text file of your choice.
- Properly comment the code.
- Provide a brief report (max one page) explaining the solutions adopted.

2 Exercise 2

The file Ted.csv contains information about various presentations given for the TED platform by different speakers. Each row represents a presentation with all its associated details. The goal of this exercise is to explore the data contained in the file and solve the following tasks:

1. Open the file and display basic information for each variable, the number of null values and type. Each row's variables are separated by commas.
2. Each presentation is identified by a title. Assign a sequential numeric value to each title. Save the mapping in a dictionary and add a variable with this numeric identifier to the original dataframe.
3. One variable contains missing values. Print the possible values of this variable. Create a new dataframe containing only the rows with missing values in this variable and remove these rows from the original dataframe.
4. Generate descriptive statistics for the numerical variables present: Mean, median, mode, and standard deviation. Additionally, produce the correlation matrix for the numerical variables.
5. Add a new variable to the data representing the ratio between the number of comments on a presentation and its duration. Plot the distribution of comments, duration, and their ratio in three separate plots.
6. What occupation is most represented among the speakers? Identify the most frequent profession.
7. For each year, show the total number of presentations and the total number of views. Determine which year had the highest number of views per presentation.

Some of the exercises will require you to read and understand some functions of the pandas library. Use the basic logic that was provided during the lectures and read the documentation carefully. Some useful resources:

- Adding new variable
- Basic statistics
- Plotting in Pandas

Complete tasks 1 through 7 in Python, structuring the code in multiple .py files or in a single Python notebook. The code must clearly and concisely address each question, showing the procedures and results for each of them.