

Disruptive Data Summer School: Project 4

Fabbi Francesco

1 Project Documentation

In **Project 4**, from a given .csv file we need to take the list of words within and:

1. order the list from A to Z and write it down in a new .csv file;
2. evaluate the number of words;
3. evaluate the average word length for the list;

To solve these three problems, a Python script has been wrote.

First of all, we open the given file (in reading mode) and read all the words within by using the function **readlines()**, which returns a list with the words, return character included except for the last one. The given file is then closed.

We now sort this list because we want to reorder the words from A to Z (crescent alphabetical way). However, we want to do it in an insensitive case way! That's why we add the "key" value in the function "sort()": the **key** value allows us to apply a function to the elements of the list before doing comparisons, and in this case we apply the function **lower()** to them, to put the strings in lower case. The result is that the original list is now modified and the words within are ordered from A to Z, and we can write them in another file, which we create accordingly. Before that, the number of words in the original file is evaluated by counting how many words are contained in the modified list, by using the function **len()** applied to the modified list.

Next, we enter a "**for**" **cycle** where each element of the modified list is written into another file, which maintains the same structure of the previous one, since each element of the list is a string containing also the **special return character**, except for one (in this particular case, the return character is added manually). While doing so, we evaluate the length of each word we write in the file, but the return character IS NOT CONSIDERED. After writing the last word, we exit the cycle.

In the end, we close the second file because we've finished writing on it, and return the number of words and the average word length (exact result and approximated one).

1.1 Usage

To evaluate the script correctness, the following procedure can be followed:

1. First of all, a Python version should be installed on the computer (not necessarily the latest one, the only thing that matters is that it belongs to the family of version 3.0);
2. Put the script file (named **Project 4 script.py**) and the .csv source file **Words - Foglio1.csv** both in the same directory (the Desktop should be fine);
3. Right-click the script file and select "**Edit with IDLE**";
4. A window showing the code of the script should appear. Move to "Run" and then select "Run Module";
5. The Python Shell window should appear showing the script execution and in the same directory where the script and the source file have been moved, another file should be created, named "**Words reorganized.csv**", where the words have been wrote;