Master BeNeFri in Computer Science

Course: Digital Humanities
Fall 2022

Exercise #4: RegExp, statistics and tests

Instructions

Download from the ILIAS website the French Theater data (path: "/french-theater/") and the GSS7214_R5 data (filename: "GSS7214_R5.DAT"). This last dataset contains information about the education level for various regions in the US (see slides 40-42 and 83).
This exercise series consists of 8 practical questions. Upload your answers and the source code used for your computations to the ILIAS website. You can submit either a .pdf file or comment the source .py file. IPython/Jupyter notebook files (.ipynb) are allowed as well.

Practical Questions (RegExp)

1) Consider the sentence "Ann plays the role with Mary and Annie". Use RegExp to replace "Ann" by "Alice" and obtain "Alice plays the role with Mary and Annie".

2) Consider the sentence "$99.99 to $87.80 or Fr. 75.50". Use RegExp to remove the decimal part of prices and obtain "$99 to $87 or Fr. 75".

Practical Questions (French Theater)

3) For the genre "Comédie", extract (in a list) the number of word-tokens per play. Do you obtain the same mean as the one indicated in the lecture slides (namely 9934.91 for 310 plays in this category)?

4) Apply the t-test to verify the hypothesis that, in mean, a French comedy contains 10,000 word-tokens.

5) Apply the t-test to verify the hypothesis that, in mean, a French tragedy contains 14,000 word-tokens.

6) Apply the t-test to verify the hypothesis that, in mean, a French tragedy contains 15,000 word-tokens.

7) Describe the results you obtained from the previous tests.

Practical Questions (GSS7214_R5)

8) Apply the χ^2 test to verify the hypothesis that the distribution of the education level in the New England region is similar to the the one appearing in the Pacific area.