Course title: Digital Humanities Prof. J. Savoy

Serie 4. Statistics and tests

With Regexp, you can search and modify strings in a flexible way. Write regexp expression in Python to modify the string as follow:

1. Replace “Ann” by “Alice” in “Ann plays the role with Mary and Annie” to obtain “Alice plays the role with Mary and Annie”
2. Remove the decimal part of prices “$99.99 to $87.80 or Fr. 75.50” to obtain “$99. to $87. or Fr. 75.”

With the code available and the folder /data/theatre-classique/, you can extract many plays (stored in XML format).

Focus on the size (number of tokens) per play and text genre. For the genre ‘Comédie’, extract (in a list) the number of word-tokens per play. Do you obtain the same mean as the value indicated in the slides of the lecture (namely: 9934.91, for 310 plays in this category)?

1. Apply the *t*-test to verify whether we can specify that, in mean, a French comedy contains 10,000 word-token?
2. Apply the *t*-test to verify whether we can specify that, in mean, a French tragedy contains 14,000 word-token?
3. Apply the *t*-test to verify whether we can specify that, in mean, a French tragedy contains 15,000 word-token?

Explain the result of the previous test in plain English.