Exercises 3
10.03.2014

Rules: The document contains a set of 5 exercises, each of them worth 2 points. You need to provide for every exercise a Python script. All Python scripts must be put in a ZIP archive named FirstName_LastName_ex*n*.zip, where n is the number of the exercise session (see ex_set_1.pdf). The Python scripts must be named exercise_m.py, where m is the number of the exercise. The ZIP archive must be uploaded on ILIAS until the specified deadline.
For each exercise session you will get either one point or zero. One point is given if 4 out of 5 exercises in that particular session are correctly solved.
Good luck!

Exercise 1. Write a Python script that parses the input_ex1.txt file and identifies all XML tags. Write in output_ex1.txt file the root tag, an empty line and then each tag name only once.

Exercise 2. Write a Python script that decides if a XML file is well formed. If it is, output *filename: YES* on the standard output; otherwise output *filename: NO*, *reason*, where reason indicates the fault in the XML file. As input, you can use the following samples: input_ex2_1.txt, input_ex2_2.txt and input_ex2_3.txt.
Please note that the script should handle various types of faults in a XML file, not only those observed in the 3 samples. For grading, other files might be tested. See XML slides and the comparison with HTML.

Exercise 3. Write a Python script that outputs to the standard output all trigrams (where the grams are represented by characters) in the following sentence:
“Keep it short and simple!”
E.g.: K e e
e e p
etc.

Exercise 4. Write a Python script that parses the input_ex4.txt file and counts all bigrams (where the words are treated as grams) and outputs to a file those that appear more often than a certain threshold (given as a command line argument). The output format will be as pairs (bigram, number_of_occurrences).
Elements inside “<” “>” and numbers will not be considered as words.
Pay attention to the new line characters. For example in the first sentence of the file:

“AFTER an unequivocal experience of the inefficacy of the subsisting
federal government, you are called upon to deliberate on a new”

the group “subsisting federal” is also a bigram.

Exercise 5. Use the same input file as for the previous exercise. Create an output file containing all contexts of the word “France” (case insensitive), having *n* words before and after (*n* is given as a command line argument). Consider that the context ends at dot (i.e., if there are $m < n$ words between the required word and the dot, then output only these m) and starts at the beginning of the sentence (i.e., if there are $m < n$ words between the beginning of the sentence and the required word, then output only these m). Do not stop at the new line characters if they are in the middle of a sentence.