Linnæus University

Programming for Digital Humanities,
4ME501: Fall Term 2017

Assignment 2: *Intermediate Text Processing with Python*
Deadline: October 14^{h}, 2017, Moodle

Contact Persons:
Marcelo Milrad: marcelo.milrad@lnu.se
Dan Kohen: dan.kohen@lnu.se

Description:

The purpose of this assignment is to demonstrate your skills and knowledge on intermediate text processing with Python. The specific aim is to analyse the content of an e-book at the by: 1) determining longest sentences and words; 2) calculating most frequently used words; and, 3) extracting named entities. You will need to write a Python program that should perform the tasks 2-5 as described below. The assignment needs to be completed using the following steps as outlined below:

1. ***Acquire the content of the e-book.***
   We have downloaded one freely available e-book from www.gutenberg.org. The book is titled *Lectures on The Language: An Introduction to the Study of Speech* by Edward Sapir from 1921 (available at http://www.gutenberg.org/cache/epub/12629/pg12629.txt).

   We then extracted the first chapter (*I Introductory: Language Defined*), and made it available for you to download from Moodle course (as a ZIP file). Please download this file.

2. ***Extract longest sentence.***
   For the text file above, write a piece of Python code to find and write out on the screen the longest sentence.

3. ***Extract longest word.***
   For the text file above, write a piece of Python code to find and write out on the screen the longest word in this given file.

4. ***Extract top ten most frequent words.***
   This is a more complex and elaborated task which will comprise several sub-steps. First, identify tokens. Then, remove stop words using the list of "stop words" available in Moodle (included in the same ZIP file as mentioned above). Further, determine frequencies of occurrences for each word after the stop-word removal. Print out (on the screen) the top ten most frequent ones, together with their frequency counts.

5. ***Extract named entities.***
   While ignoring first words in sentences (which are capitalized by default), extract named entities in other parts of the sentences. Create two lists, one with entities starting with letters from A to L in the English alphabet, and the other from M-Z. Print out the results on the screen. In order to perform this task, you should refer to the "filtered" text you have generated in step four after removing all "Stop-Words".

6. ***Assignment's requirements***
   You should implement at least two *Python functions* in your code to perform the tasks described above. Note also that you should not use external functions available for instance via external natural language processing tools such as Natural Language Toolkit (http://www.nltk.org/#natural-language-toolkit)

Expected outcomes and final results:

You are expected to generate two deliverables as described below:

Deliverable 1: A program in Python with the code you have written and created for solving the tasks above. Remember to include comments in your code where appropriate.

Deliverable 2 :
You should produce a short report (500 to 700 words) in which you discuss and present the ideas and approaches on how you solved the tasks above. You work should be reported following the publication format available at: http://goo.gl/OtPQ5.

This assignment is conducted on individual basis. Please upload a ZIP file named 'lastname_assig1.zip' to the corresponding Moodle folder including all the materials you have produced (source Python code (Assignment2.py), and your report (in PDF format). The deadline for submitting your assignment is by October 14th, 2017 at **23:55**. Good luck!