Linnæus University

Programming for Digital Humanities
4ME501, Fall Term 2017

Assignment 3: *Data Visualization*
Deadline: October 29th, 2017, Moodle

Contact persons:
lir Jusufi: ilir.jusufi@lnu.se
Marcelo Milrad: marcelo.milrad@lnu.se

Description:

The purpose of this assignment is to demonstrate your skills and knowledge on processing the content of the specific chapters of two e-books, exporting the relevant data from these texts into Comma Separated Valued (CSV)[1] files and then visualizing these data sets. The assignment is a further elaboration and enhancement of assignment #2 in the sense that you can reuse/readapt the code (or parts of it) you have written before. In this assignment you will need to solve the following tasks: 1) find the top 10 longest sentences; 2) to identify and extract the 10 most frequently used words.

The last and third task will be to generate a number of visualizations using the RAWGraph tool. In summary, you will need to write a Python program that should perform tasks 1-2 outlined below and then visualize these data sets (task 3) and then write a final report. Please have also a look at the video available in the Vimeo channel of this course describing this assignment.

Acquire the content of the two e-books.

We have downloaded two freely available e-books from Gutenberg.org. The first one is entitled *Lectures on The Science of Language* by Max Müller from 1862 (http://www.gutenberg.org/files/32856/32856-0.txt) and the second one is entitled *Language: An Introduction to the Study of Speech* by Edward Sapir from 1921 (http://www.gutenberg.org/cache/epub/12629/pg12629.txt).

We then extracted the first chapter from each one of the books. From the first book, the first lecture (*Lecture I. The Science Of Language One Of The Physical Sciences*) and from the second book the first chapter (*I Introductory: Language Defined*). These files are available in Moodle, so please download these materials from there.

1. Extract the top ten longest sentences for each chapter

You will need to write a piece of Python code to find and write out the lists of the top ten longest sentences for each one of the chapters. Extract these sentences and append them to a CSV file, rather than simply printing them out on the screen. Furthermore, the program will need to calculate the number of words for each sentence and store that in the first column named *"length"* of the CSV file. Thereafter, you should store the actual content of the sentence in the second column named *"sentence"*. The third column will store the number of the *"chapter"* where the sentence has been extracted from. Finally, the last column should store the sentence *"ranking"*, where the longest sentence within the given chapter gets the ranking of #1, the second longest gets the ranking value of 2 and so on. As a result, you should have one CSV file for both chapters. Your final CSV file will have to look similar to the following example given below:

```
length, sentence, chapter, ranking
5, ”I remember everything about it”, 1, 10
6, ”The season is about to end”, 2, 9
...
```

[1] https://en.wikipedia.org/wiki/Comma-separated_values

2. Extract the top ten most frequent words and their length.

This part of the assignment is identical to the corresponding task of assignment 2 for this course (including the removal of stop words). The difference is that in this case you will extract the top ten most frequent words for each chapter, and information on the length of each word, and store this information into a CSV file. Thus, the most frequent words are stored in the first column *"keyword"*, the *"frequency"* is stored in the second column, the number of characters is stored in the next "*length"*. Finally the book chapter where the *"keyword"* is extracted is stored in the column *"chapter"*. As a result, you should have a CSV file that looks similar to the one described below:

```
keyword, frequency, length, chapter
speaking, 50 , 8, 1
without, 32, 7, 2
...
```

3. Generate visualization using RAWGraph

Use the http://app.rawgraphs.io/ tool in order to generate a set of visualizations using the exported CSV data you have created from tasks 1 and 2. You should generate **three** visualization images in total, for example, you can usa two *bar charts* for tasks 1 and 2 and an additional *circel packing* for task 2 or any other combination you deem useful. You should use at least two different visualization metaphors/graphs (example *Barchart* and *Circle Packing*) that would give insights into all of the given variables (dimensions) in the generated CSV files. Export the generated visualization as images, so you can use them later in your report. Describe your decisions for designing the visualizations (see below for details).

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. You should implement several *Python functions* in your code to perform the tasks described above. Remember to include relevant comments and explanations in your code where appropriate.

Deliverable 2: A written report (between 1200-1500 words at the most) in which you discuss and present the ideas and approaches that described how you have solved the tasks above. Remember to include the images representing the visualizations in the report. Make sure you explain your design decisions, i.e., explain your choices of visualization processes and why do you think they are good for the given data.

Your work should be reported following the publication format available at: http://goo.gl/OtPQ5. Please upload a ZIP file named 'lastname_assig3.zip' to the corresponding Moodle folder including all the materials you have produced (source Python code (csvgenerator.py), and your report (in PDF format).

This assignment is conducted on individual basis. The deadline for submitting your assignment is by October 29th, 2017 at 23:55. Good luck!