4ME501 Assignment 3

George Akritidis
Department of Media Technology
Linnaeus University
ga222ey@student.lnu.se

Introduction

This assignment is the third in a row for the course Programming for Digital Humanities and it is a combination of the two previous but with increased difficulty. My tasks were to extract two texts, to analyze them and finally visualize my findings. Specifically, I had to find the 10 longest sentences and their length for each chapter; and the ten most used words for each chapter similarly. By chapter, I mean the two texts that I downloaded. For this assignment, I consider the book **Language: An Introduction to the Study of Speech** by Edward Sapir as chapter 1; and the book **Lectures on The Science of Language** by Max Müller as chapter 2. It is important for the reader to be aware of this, because one of the goals was to extract sentences and words from each book and present them together.

Approach to solving the given problem and tasks

My approach to solve the tasks was similar to the previous assignments. I tried to keep the structure of the code as simple as possible. The trick was to separate the solution of the problems in small steps, and proceed to the final solution step by step. The data processing of Max Muller book (chapter 2) was the most difficult part of this assignment. The reason for that was that, it contained a lot of unwanted text, such as slashes encoding etc, which had to be removed. To do that, I used several regular expressions. Because, I didn't have sufficient knowledge and experience of the regular expressions, I was in trouble in the beginning but after studying several tutorials and online examples I finally comprehended the syntax. My strategy was to create several empty lists and modify the text step by step. Every time, I made one modification I transferred all the data to a new list, so in the last list the text was rather clean and ready to apply the functions on it.

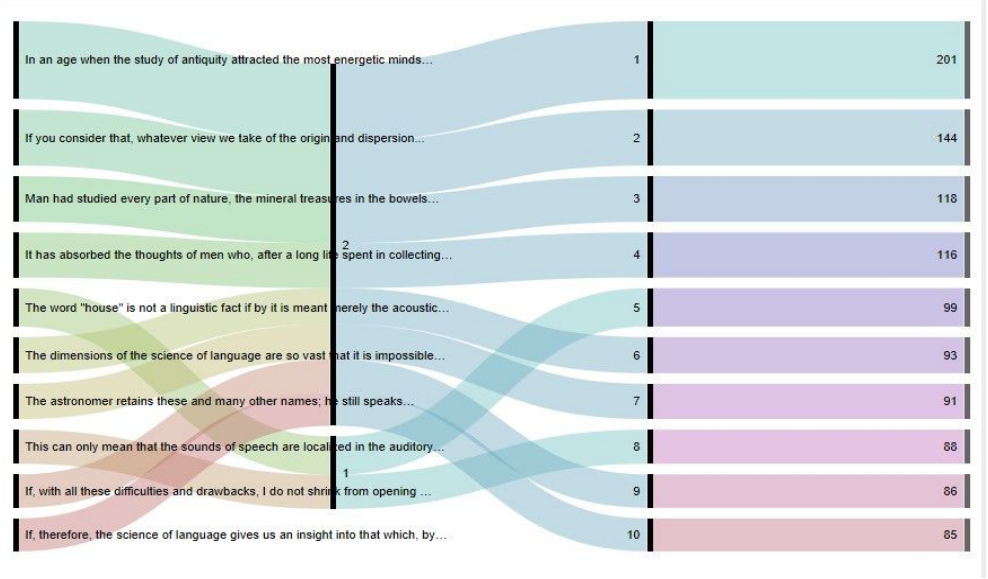
For this assignment, I made use of tuples as well and not only lists as I

did in the previous assignments. The reason for that is that I was able to gather the data(length, ranking, sentence) in pairs easier with the use of tuples. Eventually, I had to transfer the data again in a list of lists and finally all the data from the two merged and sorted lists in a string. I followed the examples that were provided to us in vimeo, regarding the creation of the csv files. But for the other part of the assignment, I mostly searched online for similar cases where other programmers had the same questions as I did. These questions involve problems, such as how to convert tuples in lists or how to add items in tuples and lists, etc.

Outcomes/Analysis of results

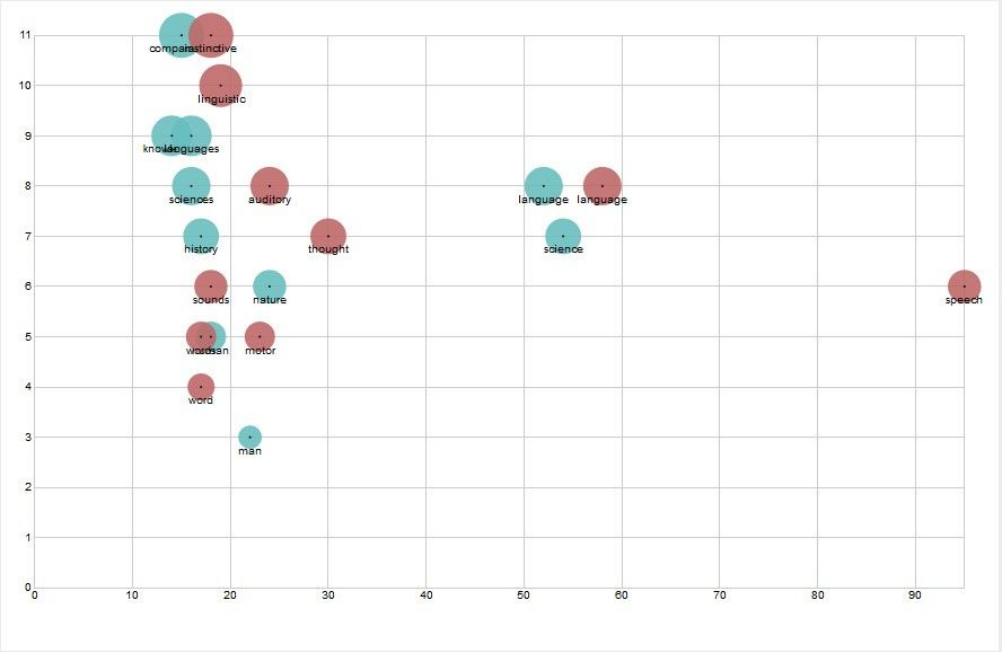
The outcome of the results were stored into two csv files. Basically, the files were two arrays with four columns. There was not anything surprising to the results. As in the previous assignments the results concerning the most used words were more or less the same. Part of the assignment was to visualize the results with visualization means such as bar chart, pie charts, etc. To do so, I used the website that was suggested to the students (http://app.rawgraphs.io), that offers a great variety of charts. For the implementation of this assignment, I selected an alluvial diagram for task 1 and a circular dendrogram and a scatter plot for task 2. The reasons for doing so, was mainly because of the formation of data; and these specific charts give the best overview of the findings.

As far as task 1 is concerned, the alluvial chart is a very good option because it gives me the capability to present the sentences in the left side of the chart nicely; the ranking in the middle of the chart and the length in the right. Every lane represents a sentence and it is easy for the eye to distinguish each

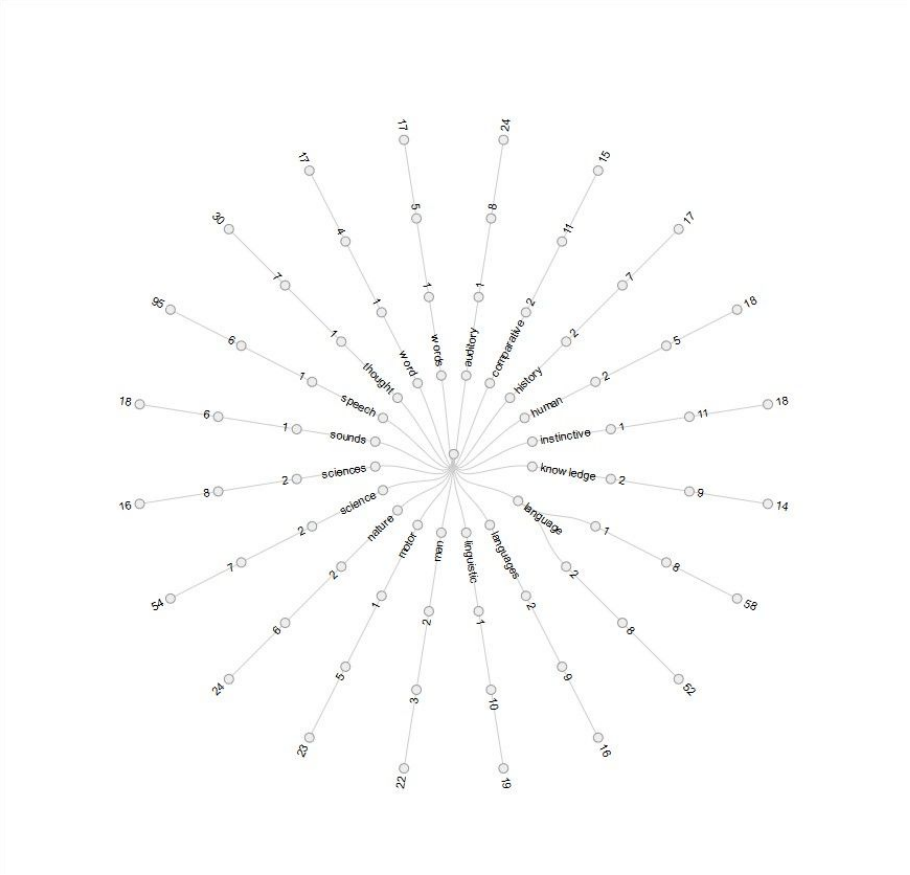


group because of the small gap between the two groups and the numbers 1 and 2 situated on top of the lanes. Thus, in my case the reader can see that the sentences which belong in chapter 1 are listed in the ranking at 5 and 8 position.

Now regarding task 2, I thought that a scatter plot and a dentrogram are the ideal charts to present the data. To begin with, I have to mention that in this task there were two parameters which had to be visualized; and those where the number of words and number of characters for each word. So, a chart with two axis x,y were a good option for the visualization. Specifically, in the x axis I cite the occurences of the words and in the y axis the number of characters. The color indicates the chapter from which each word was extracted. Red symbolizes chapter 1 and cyan the chapter 2. Lastly, the size of each circle indicates the length of the word in characters; bigger the circle, longer the word.



About the second chart, I consider the dentrogram to be also a very good decision because it presents the data very clear; spherically and in a cohesive way. The words are listed in the center alphabetically, then the first parameter is the chapter where the word derives from, followed by the length of the word and finally the frequency of each word. It is very easy for the reader to have a full image of the findings without having to bend his head to read the data, in contrast to other charts.



Conclusions and Reflections

This assignment was apparently the most demanding from this series of assignments for the course Programming for Digital Humanities, yet it was the most challenging and interesting. I consider the ability to perform such actions with programming languages very useful, because we can infer a lot of things from the statistics that are generated. I am looking forward for the continuation of this course, which will allow me to enhance my knowledge in Python. This kind of assignments help you to sharpen your mind, because you put yourself in a position to think of solutions on accomplishing something. This action is very beneficial for the brain, and the idea of creating something with joy is something that encourages me to carry on in the future.

REFERENCES

https://stackoverflow.com/questions/12883376/remove-the-first-word-in-a-python-string
https://www.decalage.info/en/python/print_list
https://stackoverflow.com/questions/4481724/convert-a-list-of-characters-into-a-string
https://stackoverflow.com/questions/4174941/how-to-sort-a-list-of-lists-by-a-specific-index-of-the-inner-list
https://stackoverflow.com/questions/2407398/how-to-merge-lists-into-a-list-of-tuples-in-python
https://stackoverflow.com/questions/2663391/using-a-loop-to-add-objects-to-a-listpython
https://stackoverflow.com/questions/5618878/how-to-convert-list-to-string
https://stackoverflow.com/questions/4481724/convert-a-list-of-characters-into-a-string