Justīne Jaudzema

014824887

justine.jaudzema@helsinki.fi

Introduction to methods for digital humanities

For my research, I used Latvian folk legends collected and published by researcher Peteris Šmits. The data that I used is freely available in http://valoda.ailab.lv/folklora/pasakas/.

Legends that I used are published in three volumes and are divided into numerous themes. By counting words and seeing which ones are used more often I might get a new perspective on characters and themes that are viewed in these legends.

Since the data that I used is a book text I needed to clean it. Since I need only the text from the legends and not any other additional information I used OpenRefine to delete all the metadata and notes that were connected to the legends. I also used OpenRefine to make all the text in lower case and get rid of all the punctuation marks. After that, I tried to clean all the data from all the stop words but as it can be seen in the end result I was not really successful with that.

Then I used Python to count the word frequency.

To visualize and make the end result more attractive I used https://voyant-tools.org/ all though this does not show some characters that are found in Latvian alphabet.

As I mentioned before I did not clean data as much as I should have and the first thing to do would be that. I also think that I should have used less amount of data and looked at different words used in legends that are connected to one theme not cluster many of them together.