Lab 6 - due 4 June 2017

Your Name

References in your solutions

Please add references to any external material that you base your answers on.

Typesetting

Please put your answers in a LaTeX document and submit both a pdf and the LaTeX source. Try to keep a similar template to the one given. You may find it helpful to use sharelatex or install a LaTeX distribution and editor, such as texworks. Some documentation on LaTeX can be found at https://en.wikibooks.org/wiki/LaTeX.

Study material

- Sissejuhatus hajasaruvtustesesse kasutades mpi4py https://github.com/Abercus/hajasaruvtusesse mpi4py
- Jockers Text Analysis with R for students of Literature Springer (2014), http://dx.doi.org/10.1007/978-3-319-03164-4
- Veillard, Libxml2 https://github.com/GNOME/libxml2
- Zhao http://www.rdatamining.com
- LaTeX https://en.wikibooks.org/wiki/LaTeX

Text processing

Homework question 1

ANSWER
COMMENT
GRADE

Homework question 2
What interesting topics do you find in Jockers (2014).

ANSWER
COMMENT
GRADE

Homework question 3
Different authors and different literary genres typically have statistically different distributions of words. Write a program to count words in either two long texts or two different corpora (collections of similar texts). The texts should be freely available online, and in the same language, but the language need not be English. Some examples sources include
Depending on your source of text, you may wish to remove markup language, for example by using Beautiful Soup (https://www.crummy.com/software/BeautifulSoup/) or Nokogiri (http://www.nokogiri.org/). Indicate possible choices of classifiers for distinguishing between your two different texts or corpora.

**Homework question 4**

What is libxml2 and what language is it written in?

**ANSWER**
Kirjandus

[6] LaTeX website http://www.latex-project.org/