Spring Semester 2018
Iowa State University

LING 410X - Language as Data

Assignment 5
Submission Deadline: 31 March 2018, end of the day

Instructions: This assignment is for 15 marks and has only one question. Upload your submission as a PDF report in your_Lastname_A5.pdf format. Late submissions are allowed, but will not be awarded full credit.

Question 1

This assignment comes with a corpus of 15 most popular books on Project Gutenberg. I am also providing code in R markdown to create topic models, based on an online tutorial. Take that corpus and the R markdown code, and answer the following questions:

1. How long did it take for you to run the program as is? Also mention what machine are you using (processor speed, RAM etc)
2. What are the five most frequent words and their frequencies in this corpus?
3. Looking at the topics and the terms associated with them, what do you think the code is learning? Do the word clusters per topic make sense to you?
4. Change the settings and/or pre-processing so that you will eventually end up seeing at least partially meaningful topic-word clusters (one thing you can change: create a custom stop word list because words such as the, said etc are not meaningful topic descriptions. Figure out how).
5. If you run it twice, are the topic clusters showing up as the same or different?
6. What happens if you increase or decrease the number of topics?
7. Try adding a few additional texts (5 more from Gutenberg.org, for example), and see if you see different topic distributions
8. What happens if we have a huge amount of texts (not 10-20 as we are doing now?)

9. Do you think we can do the same procedure with bigrams and trigrams instead of words?

10. What makes topic models more useful than word clouds?

11. Think of 3 scenarios where this is a useful exercise to do to draw some insights.

Write a report (up to 5 pages) discussing these issues.

Note: The reason why I am referring to Gutenberg so frequently is not because I want everyone to work on literary texts. It is just because it is a huge corpus of publicly and freely accessible texts.