Session 1.2: Collecting, reading and cleaning text data

Contents

Session overview

In this session we will learn how to read texts into R. We will see how to handle files in .doc, .html, and machine-readable .pdf files. These are the easiest case scenarios. We will then move on to .pdf files that are challenging because of their layouts and then discuss strategies for digitizing PDFs that are not yet machine-readable. We will finally motion towards web scrapping of text data and point you towards what is possible and tools that can be used to gather text data from the web.

The second major goal of this session is to consider how to handle text in R. We will go further into regular expressions (those of you who have done the tutorial will have already encountered some of this) and then learn common techniques for cleaning up text in R and formatting it so that it is in a condition we can begin to do some text analysis on. Using tools of data frame manipulation we will then learn one method for creating Google n-gram-style graphs.

Download the following for this session:

Script for this session.

Four Charles Dickens novels can be downloaded here, here, here, and here. Put these in a separate folder within your working drive, preferably with a name that is easy to type.

We will also be working with some examples of historical documents. Download these (probably easiest to save them to your working directory):

- Sample Stortings melding from 1960-61.
- Statoil's 2001 Sustainability Report (also here)
- King Haakons 1925 speech to the Norwegian parliament