Text Mining for Social Scientists

Felix Lennert

2022-04-07

Contents

Chapter 1

Introduction

Dear student,

if you read this script, you are either participating in one of my courses on digital methods for the social sciences, or at least interested in this topic. If you have any questions or remarks regarding this script, hit me up at felix.lennert@ensae.fr.

1.1 Outline

This script will introduce you to the quantitative analysis of text using R. Through the last decades, more and more text has been readily available. Think for example of Social Networking Platforms, online fora, Google Books, newspaper articles, or the fact that YouTube can generate subtitles for all of its videos or the fact that administrative records are increasingly digitized. Social scientists of course have decades of experience analyzing these things, yet they used to be constrained by data availability and, hence, their data sets used to be way smaller and could be processed by humans. In order to make the most out of the newly available data sets I mentioned above and to repurpose them for social scientific research, we need to use tools from the information and computer sciences. Some are fairly old such as basic dictionary-based sentiment analysis whose precursors were introduced in the 1960s, others as recent as early 2000s (LDA) or even 2014 (word2vec).

In specific, this script will cover the pre-processing of text, the implementation of supervised and unsupervised approaches to text, and in the end I will briefly touch upon word embeddings and how social science can use them for inquiry.

The following chapters draw heavily on packages from the `tidyverse` (**?**), `tidytext` (**?**), and `tidymodels` (**?**), as well as the two excellent books "Text Mining with R: A Tidy Approach" (**?**) and "Supervised Machine Learning for