Assignment 2

Text as Data

2023-09-25

Parsing XML text data

In this assignment we will access and work with German Parliamentary data, which is available in XML format here (scroll down) for the last two parliamentary periods. Remember XML format is very like HTML format, and we can parse it using a scraper and CSS selectors. Speeches are contained in `<rede>` elements, which each contain a paragraph element describing the speaker, and paragraph elements recording what they said.

1.1

Choose one of the sessions, and retrieve it using R or Python.

1.2

Using a scraper, get a list of all the elements.

1.3

For each element, get the name of the speaker, and a single string containing everything that they said. Put this into a dataframe.

2.1

Choose a politician, and print the number of speeches they made in this session

2.2

Print the content of the first speech by the politician you choose.

2.3

Process the list of speeches into a TFIDF matrix. What are the highest scoring terms in this matrix for the first speech by the politician you have chosen?

2.4

Using the resource "Stammdaten aller Abgeordneten seit 1949 im XML-Format", retrieve the records pertaining to your chosen politician and print the information they contain.