Codebook for LaroplanOCR

21 September, 2022

Contents

Overview

Swedish primary school curricula (Läroplaner för grundskolan) in digital format. We use optical character recognition (OCR) to transform curricula in image format into text. For each curriculum we construct datasets at the paragraph, sentence, and word levels.

We include the following Läroplaner in the data:

Year	Title
1962	Läroplan för grundskolan 1962
1969	Läroplan för grundskolan 1969
1980	Läroplan för grundskolan. Allmän del: mål och riktlinjer, kursplaner, timplaner
1994a	Kursplaner för grundskolan 1994
1994b	Läroplaner för det obligatoriska skolväsendet och de frivilliga skolformerna : Lpo 94
2011	Läroplan för grundskolan, förskoleklassen och fritidshemmet 2011

The files were obtained from the Gothenburg University Publications Electronic Archive.

The full code that processes the raw pdf files into the datasets available in this package can be found in `github.com/JMSLab/LaroplanOCR`.

Using the Datasets

In `./datasets/` you will find the following for the Läroplan for year `YYYY`:

- `lgrYYYY_counts.csv`: counts of individual words
- `lgrYYYY_paragraphs.csv`: individual paragraphs with page of appearance
- `lgrYYYY_sentences.csv`: individual sentences with page and paragraph of appearance

In the folder `./example/` we provide an illustration on how to use the data to search for, and plot, counts of a desired set of words.

Counts

The “counts” files include two variables: `word`, `n`.

Some summary statistics of these variables:

Year	Unique word	Mean n
1962	19,085	9.79
1969	10,578	9.65
1980	5,603	6.81
1994a	3,403	5.99
1994b	1,674	5.97
2011	15,398	5.94

Citations

- Hermo, Santiago, Christian Lundqvist, Miika Päällysaho, David Seim, Jesse M. Shapiro, and Stina Trollbäck. 2022. LaroplanOCR. Code and data repository at https://github.com/JMSLab/LaroplanOCR.
- Hermo, Santiago, Miika Päällysaho, David Seim, and Jesse M. Shapiro. 2022. Labor Market Returns and the Evolution of Cognitive Skills: Theory and Evidence. The Quarterly Journal of Economics, Volume 137, Issue 4, November 2022, Pages 2309–2361. DOI: https://doi.org/10.1093/qje/qjac022.