Nasjonalbiblioteket
National Library of Norway

Freely available documents from Norwegian state institutions (Målfrid 2022)

Description

The corpus consists of documents from 571 domains of Norwegian state institutions and comprises totally approx. 4.5 billion tokens, which makes it one of the largest freely available resources for Norwegian Bokmål and Nynorsk. In addition to Norwegian, the corpus contains texts in Northern Sami, Lule Sami, Southern Sami and English:

Number of tokens per language:

Norwegian Bokmål	'nob'	3 284 142 329
English	'eng'	910 993 105
Norwegian Nynorsk	'nno'	300 210 030
Northern Sami	'sme'	5 566 229
Southern Sami	'sma'	377 795
Lule Sami	'smj'	220 287

The data were collected as part of the so-called Målfrid project, where the National Library of Norway on behalf of the Ministry of Culture and in collaboration with the The Language Council of Norway collects and aggregates data for mapping the usage of Norwegian Bokmål and Norwegian Nynorsk on the domains of Norwegian state institutions.

The corpus is the result of a focused crawl conducted between December 2021 and January 2022, recursively downloading text documents (HTML, DOC(X)/ODT and PDF) from a set of domains (down to and including level 12), while obeying robots.txt and politeness restrictions. The crawled documents were further processed according to their format: Natural language was extracted from HTML using the boilerplate removal system Justext (http://corpus.tools/wiki/Justext), from the Word/ODT documents using Textract (https://textract.readthedocs.io/en/stable/) and from the PDFs using Google Cloud Vision OCR.

The extracted text was classified using TextCat language identification (cf. https://www.let.rug.nl/~vannoord/TextCat/) on document-level, provided as part of the metadata. The documents were deduplicated on domain level (exact duplicates).

Format

The corpus is provided as gzipped JSON lines (jsonl), one document per line. There is one JSONL file per combination of domain, language and content type. The files are encoded as UTF-8, with ASCII escape sequences. Each dictionary contains the following keys:

- doch_hash: a sha256 checksum of the fulltext
- lang: language of the document (detected using TextCat)
- url: the url of the document at crawl time
- date: crawl date
- mimetype: media type of the document (simplified): HTML, DOC or PDF
- fulltext: an array of strings, where each string represents one paragraph. An empty string denotes a new page in the PDF documents.

License

Contact

If you have questions regarding the material, please contact us at sprakbanken@nb.no.