Freely available documents from Norwegian state institutions (Målfrid 2025)

Description

The corpus consists of documents from 493 domains of Norwegian state institutions and comprises totally approx. 2.4 billion tokens, which makes it one of the largest freely available resources for Norwegian Bokmål (nob) and Nynorsk (nno). In addition to Norwegian, the corpus contains texts in Northern Sami (sme), Lule Sami (smj), Southern Sami (sma) and English (eng).

Number of tokens per language:

Norwegian Bokmål	'nob'	1 731 882 315
English	'eng'	520 138 241
Norwegian Nynorsk	'nno'	164 754 934
Northern Sami	'sme'	2 208 507
Southern Sami	'sma'	368 489
Lule Sami	'smj'	282 859

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 2024 and January 2025, 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. Crawling and processing were carried out using the freely available Python package maalfrid_toolkit (https://github.com/NationalLibraryOfNorway/maalfrid_toolkit). The package extracts natural language from HTML using the boilerplate removal system Justext (http://corpus.tools/wiki/Justext), from Word documents using docx2txt/antiword and from PDFs using a custom text extractor created by the AI-lab at the National Library of Norway.

The extracted text was classified using Gielladetect/pytextcat (https://github.com/NationalLibraryOfNorway/gielladetect) 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

The data is provided under the Norwegian Licence for Open Government Data (NLOD) 2.0: https://data.norge.no/nlod/en/2.0/

Contact

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