Nasjonalbiblioteket

N-grammer fra NBdigital 2022

Dette korpuset inneholder n-grammer – unigrammer, bigrammer og trigrammer – fra alle bøker og aviser som var blitt digitalisert ved Nasjonalbiblioteket per 15. juli 2022. De er laget på basis av et material på om lag 610.000 bøker og 4.000.000 aviser. N-grammene finnes på CSV-format (UTF-8-kodert).

Innhold

Kolonnene i CSV-filene med n-grammer er som følger:

- first - det første ordet i n-grammet (i unigram, bigram og trigram)
- second - det andre ordet i n-grammet (i bigram og trigram)
- third - det tredje ordet i n-grammet (i trigram)
- lang - språkkode for n-grammet (bare i bøker, aviser har ingen språkklassifikasjon per nå)
- freq - den totale frekvensen for n-grammet i samlingen av bøker eller aviser
- json - et dictionary med råfrekvens per år

totals.json inneholder totalfrekvenser innenfor årganger i bok- og aviskorpuset. Med disse kan man lett regne ut relativfrekvenser for sammenlikning på tvers av år som i NB N-gram.

metadata-digibok.csv og *metadata-digavis.csv* inneholder enkle metadata for alle bøkene og avisene som inngår i bok- og aviskorpuset. Hvis du er interessert i mer utførlige metadata, henviser vi til Oria eller NBs APIer under https://api.nb.no/.

Tilrettelegging

N-grammene ble ekstrahert fra fulltekstbasene til DH-labben ved Nasjonalbiblioteket (https://www.nb.no/dh-lab/). Følgende frekvenskutt ble gjennomført:

- unigram (bøker): totalfrekvens på mindre enn fem i delkorpuset for norsk bokmål, mindre enn to i alle andre språk. I tillegg må n-grammet være brukt i mer enn ett år.
- unigram (aviser): totalfrekvens på mindre enn fem. I tillegg må n-grammet være brukt i mer enn ett år.
- bigram (bøker og aviser): begge unigrammene i bigrammet må være et unigram etter kuttreglene ovenfor. Ellers gjelder de samme reglene som for unigrammene.
- trigram (bøker og aviser): begge bigrammene i trigrammet må være et bigram etter kuttreglene ovenfor. I tillegg ble alle hapax-trigrammer innenfor ett år fjernet. Ellers gjelder de samme reglene som for unigrammene.

Lisens

Nasjonalbiblioteket, 2022-12-21

National Library of Norway

N-grams from NBdigital 2022

Description

This resource contains n-grams - i.e. unigrams, bigrams and trigrams - from all books and newspapers that had been digitized at the National Library of Norway up to 15 July 2022. The n-grams have been extracted from a material consisting of approximately 610,000 books and 4,000,000 newspapers. The n-grams are offered as CSV files (UTF-8-encoded).

Contents

Columns in the n-gram CSV files:

- first - the first word (in unigrams, bigrams and trigrams)
- second - the second word (in bigrams and trigrams)
- third - the third word (in trigrams)
- lang - the language of the n-gram (only in books, newspapers have no language classification as for now)
- freq - the total frequency of the n-gram in the collection of books or newspapers
- json - a dictionary with raw frequency for each year

totals.json contains aggregated frequencies per year in the book and newspaper corpora. Using them, you can calculate relative frequencies in order to compare frequencies over time as in NB N-gram.

metadata-digibok.csv and *metadata-digavis.csv* contain simple metadata for the books and newspapers. If you need more extensive metadata, you could use Oria or the APIs at https://api.nb.no/.

Data preparation

The n-grams were extracted from the fulltext databases provided by the DH-lab at the National Library of Norway (https://www.nb.no/dh-lab/). The following frequency cuts were applied:

- unigram (books): total frequency of less than five in the corpus for Norwegian Bokmål, less than two in all other languages. Additionally, the ngram must appear in more than one year.
- unigram (newspapers): total frequency of less than five in the corpus. Additionally, the ngram must appear in more than one year.
- bigram (books and newspapers): both parts of the bigram must appear as a unigram after the cuts above were applied. Apart from that, the same rules as for the unigrams were applied.
- trigram (books and newspapers): the two bigrams of the trigram must appear as a bigram after the cuts above were applied. Additionally, hapax trigrams within one year were removed. Apart from that, the same rules as for the unigrams were applied.

License

National Library of Norway, 2022-12-21