Tekster fra norsk Wikipedia

Dette korpuset inneholder en dump av samtlige Wikipediaartikler på bokmål, nynorsk og nordsamisk fra ca. 20. mars 2019. Det er 492 864 artikler for bokmål, 139 927 for nynorsk og 7 626 for nordsamisk.

Artiklene er fordelt på tre filer, en for henholdsvis bokmål (nob.wikipedia.json, 1,3 GB), nynorsk (nno.wikipedia.json, 300 MB) og nordsamisk (sme.wikipedia.json, 10 MB). Hver fil er strukturert som et JSON-array over artiklene slik de foreligger på nettet. Hver artikkel er et strukturert element, med ett nivå av "nøkkel:verdi", som inneholder tekst og metadata. Det er åtte slike nøkkel/verdi-par i artiklene:

bytelength: lengde på teksten i bytes
pageid: identifikator for teksten.
title: tittel som i Wikipedia
hiddencategories: metadata
text: teksten som i Wikipedia
revid: revisjonsinformasjon
contentcategories: metadata
wikidata: andre data

Eksempel
Elementet med indeksverdi 1957 i bokmålsfilen (det 1958. elementet i arrayet) ser slik ut i JSON-format:


```
{
'bytelength': 397,
'pageid': 1348455,
'title': '12 Monkeys (fjernsynsserie)',
'hiddencategories': ['Artikler_med_filmlenker_fra_Wikidata',
'Artikler_med_offisielle_lenker_fra_Wikidata'],
'text': '12 Monkeys er en amerikansk science fiction- og thrillerserie som
har gått på Syfy siden 2015. Serien er science fiction, basert på
spillefilmen 12 Monkeys fra 1995.\n\nHovedrollene innehas av Aaron
Stanford, Amanda Schull, Kirk Acevedo og Noah Bean.\n\nEksterne lenker\n*
Offisielt nettsted\n*(en) 12 Monkeys på Internet Movie Database\n*(en) 12
Monkeys på Rotten Tomatoes\n*(en) 12 Monkeys på Metacritic',
'revid': 17416772,
'contentcategories': ['Science_fiction-TV-serier_fra_USA',
'TV-serier_fra_2010-årene,_fra_USA',
'Thrillerserier_fra_USA',
'TV-produksjoner_på_SyFy'],
'wikidata': 'Q18421127'
}
```

Tekster frå norsk Wikipedia

Dette korpuset innheld ein dump av alle Wikipediaartiklar på bokmål, nynorsk og nordsamisk frå ca. 20. mars 2019. Det er 492 864 artiklar for bokmål, 139 927 for nynorsk og 7 626 for nordsamisk.

Artiklane er fordelte på tre filer, ei for høvesvis bokmål (nob.wikipedia.json, 1,3 GB), nynorsk (nno.wikipedia.json, 300 MB) og nordsamisk (sme.wikipedia.json, 10 MB). Kvar fil er strukturert som eit JSON-array over artiklane slik dei ligg føre på nettet. Kvar artikkel er eit strukturert element, med eitt nivå av "nøkkel:verdi", som inneheld tekst og metadata. Det er åtte slike nøkkel/verdi-par per artikkel:

bytelength: lengd på teksta i bytes
pageid: identifikator for teksta
title: tittel som i Wikipedia
hiddencategories: metadata
text: teksta som i Wikipedia
revid: revisjonsinformasjon
contentcategories: metadata
wikidata: andre data

Døme
Elementet med indeksverdi 1957 i bokmålsfila (det 1958. elementet i arrayet) ser slik ut i JSON-format:


```
{
'bytelength': 397,
'pageid': 1348455,
'title': '12 Monkeys (fjernsynsserie)',
'hiddencategories': ['Artikler_med_filmlenker_fra_Wikidata',
'Artikler_med_offisielle_lenker_fra_Wikidata'],
'text': '12 Monkeys er en amerikansk science fiction- og thrillerserie som
har gått på Syfy siden 2015. Serien er science fiction, basert på
spillefilmen 12 Monkeys fra 1995.\n\nHovedrollene innehas av Aaron
Stanford, Amanda Schull, Kirk Acevedo og Noah Bean.\n\nEksterne lenker\n*
Offisielt nettsted\n*(en) 12 Monkeys på Internet Movie Database\n*(en) 12
Monkeys på Rotten Tomatoes\n*(en) 12 Monkeys på Metacritic',
'revid': 17416772,
'contentcategories': ['Science_fiction-TV-serier_fra_USA',
'TV-serier_fra_2010-årene,_fra_USA',
'Thrillerserier_fra_USA',
'TV-produksjoner_på_SyFy'],
'wikidata': 'Q18421127'
}
```

Texts from Norwegian Wikipedia

This corpus is a dump of all Wikipedia articles written in Norwegian Bokmål, Norwegian Nynorsk and Northern Sami from approx. March 20, 2019. There are 492,864 articles for Bokmål, 139,927 for Nynorsk and 7,626 for Northern Sami, respectively.

The articles are split into three files, one each for Bokmål (nob.wikipedia.json, 1,3 GB), Nynorsk (nno.wikipedia.json, 300 MB) and Northern Sami (sme.wikipedia.json, 10 MB). Each file is structured as a JSON Array of all the articles as they appear on the web. Each article is a structured element, with one level of "key:value" pairs containing text and metadata. There are eight such key/value pairs per article:

bytelength: length of text in number of bytes
pageid: text identifier
title: title as in Wikipedia
hiddencategories: metadata
text: text as in Wikipedia
revised: audit information
contentcategories: metadata
wikidata: other data

Example
The element with index value 1957 in the Bokmål file (the 1958th element of the array) looks as follows in JSON format:


```
{
'bytelength': 397,
'pageid': 1348455,
'title': '12 Monkeys (fjernsynsserie)',
'hiddencategories': ['Artikler_med_filmlenker_fra_Wikidata',
'Artikler_med_offisielle_lenker_fra_Wikidata'],
'text': '12 Monkeys er en amerikansk science fiction- og thrillerserie som
har gått på Syfy siden 2015. Serien er science fiction, basert på
spillefilmen 12 Monkeys fra 1995.\n\nHovedrollene innehas av Aaron
Stanford, Amanda Schull, Kirk Acevedo og Noah Bean.\n\nEksterne lenker\n*
Offisielt nettsted\n*(en) 12 Monkeys på Internet Movie Database\n*(en) 12
Monkeys på Rotten Tomatoes\n*(en) 12 Monkeys på Metacritic',
'revid': 17416772,
'contentcategories': ['Science_fiction-TV-serier_fra_USA',
'TV-serier_fra_2010-årene,_fra_USA',
'Thrillerserier_fra_USA',
'TV-produksjoner_på_SyFy'],
'wikidata': 'Q18421127'
}
```