NYNORSK (1) – BOKMÅL (2) – ENGLISH (3)

Diskusjonstekster frå Wikipedia

Dette korpuset inneheld ein dump av diskusjonstrådar frå Wikipedia, der forfattarar diskuterer ulike problemstillingar i samband med publisering av bestemde artiklar på Wikipedia. Artiklane er fordelte på to filer i JSON-format, ei for høvesvis bokmål (nb.wikipedia.json) og nynorsk (nn.wikipedia.json). Det er 31364 diskusjonar på bokmål (om lag 17 millionar ord), og 5500 diskusjonar på nynorsk (om lag 1,4 millionar ord). Kvar diskusjon er eit element i JSON-arrayet, med eitt nivå som inneheld tekst og diverse metadata. Det er åtte datafelt per diskusjon:

title:	tittel på artikkelen som vert diskutert
pageid:	identifikator for artikkelen
revid:	revisjonsinformasjon
wikidata:	ev. andre data
contentcategories:	metadata
hiddencategories:	metadata
text:	diskusjonstekst
bytelength:	lengde på teksten i bytes

Døme
Elementet med indeksverdi 229 i nynorsksfila (det 230. elementet i arrayet) ser slik ut:

```
title:             "Diskusjon:Adeccoligaen"
pageid:            67119
revid:             2874510
wikidata:          false
contentcategories:[]
hiddencategories: []
text:              "Adeccoligaen eller 1. divisjon fotball for menn?\nKva er
                   det best at denne artikkelen heiter. Er du for eller i
                   mot flytting?\n\n For Eg er for flytting av denne
                   artikkelen. Sjå kommentar under. --Cato 14:38, 6 august
                   2007 (CEST)\n For --Ekko 14:49, 6 august 2007 (CEST)\n
                   For --Anders 14:51, 6 august 2007 (CEST)--\n For --Frokor
                   15:05, 6 august 2007 (CEST)\n\nKommentarar\n\nSynest
                   verkeleg at denne sida bør flyttast til Adeccoliagen. Det
                   er dette som er det offisielle namnet på ligaen.
                   Artikkelen om eliteserien heiter jo Tippeligaen og dette
                   blir akkurat det same. Dersom ingen har nokon
                   innvendingar mot dette kjem eg til å flytte artikkelen. -
                   -Cato 14:27, 6 august 2007 (CEST).\n\nSjølv om ligaen
                   heiter Adeccoligaen er det 1. divisjon. Kor ofte kan dei
                   skifte sponsor og namn? Men så lengje det går klart fram
                   av artikkelen kor adeccoligaen er plassert i hierarkiet,
                   og så lenge Cato eller andre passar på at artikkelen er
                   oppdatert, er det vel ikkje noko gale i å flytte til
                   sponsornamnet. - Knut 16:42, 6 august 2007 (CEST)\nHei.
                   Eg flytta artikkelen Adeccoligaen tilbake til 1. divisjon
                   fotball for menn. I 2015 endra denne ligaen namn til
                   OBOS-ligaen. I staden for å oppretta ein ny artikkel med
                   namn OBOS-ligaen, synest eg dette var mest rettmessig. -
                   Sandve 23. april 2016, 15:31"
bytelength:        1283
```

NYNORSK (1) – BOKMÅL (2) – ENGLISH (3)

Diskusjonstekster fra Wikipedia

Dette korpuset inneholder en dump av diskusjonstråder fra Wikipedia, der forfattere diskuterer ulike problemstillinger i forbindelse med publisering av bestemte artikler på Wikipedia. Artiklene er fordelt på to filer i JSON-format, en for henholdsvis bokmål (nb.wikipedia.json) og nynorsk (nn.wikipedia.json). Det er 31364 diskusjoner på bokmål (om lag 17 millioner ord), og 5500 diskusjoner på nynorsk (om lag 1,4 millioner ord). Hver diskusjon er et element i JSON-arrayet, med ett nivå som inneholder tekst og diverse metadata. Det er åtte datafelt per diskusjon:

title:	tittel på artikkelen som blir diskutert
pageid:	identifikator for artikkelen
revid:	revisjonsinformasjon
wikidata:	ev. andre data
contentcategories:	metadata
hiddencategories:	metadata
text:	diskusjonstekst
bytelength:	lengde på teksten i bytes

Eksempel
Elementet med indeksverdi 229 i nynorsksfilen (det 230. elementet i arrayet) ser slik ut:

```
title:              "Diskusjon:Adeccoligaen"
pageid:             67119
revid:              2874510
wikidata:           false
contentcategories:[]
hiddencategories: []
text:               "Adeccoligaen eller 1. divisjon fotball for menn?\nKva er
                    det best at denne artikkelen heiter. Er du for eller i
                    mot flytting?\n\n For Eg er for flytting av denne
                    artikkelen. Sjå kommentar under. --Cato 14:38, 6 august
                    2007 (CEST)\n For --Ekko 14:49, 6 august 2007 (CEST)\n
                    For --Anders 14:51, 6 august 2007 (CEST)--\n For --Frokor
                    15:05, 6 august 2007 (CEST)\n\nKommentarar\n\nSynest
                    verkeleg at denne sida bør flyttast til Adeccoliagen. Det
                    er dette som er det offisielle namnet på ligaen.
                    Artikkelen om eliteserien heiter jo Tippeligaen og dette
                    blir akkurat det same. Dersom ingen har nokon
                    innvendingar mot dette kjem eg til å flytte artikkelen. -
                    -Cato 14:27, 6 august 2007 (CEST).\n\nSjølv om ligaen
                    heiter Adeccoligaen er det 1. divisjon. Kor ofte kan dei
                    skifte sponsor og namn? Men så lengje det går klart fram
                    av artikkelen kor adeccoligaen er plassert i hierarkiet,
                    og så lenge Cato eller andre passar på at artikkelen er
                    oppdatert, er det vel ikkje noko gale i å flytte til
                    sponsornamnet. - Knut 16:42, 6 august 2007 (CEST)\nHei.
                    Eg flytta artikkelen Adeccoligaen tilbake til 1. divisjon
                    fotball for menn. I 2015 endra denne ligaen namn til
                    OBOS-ligaen. I staden for å oppretta ein ny artikkel med
                    namn OBOS-ligaen, synest eg dette var mest rettmessig. -
                    Sandve 23. april 2016, 15:31"
bytelength:         1283
```

Discussions from Wikipedia

This corpus contains a dump of discussion threads from the Norwegian part of Wikipedia, where authors discuss various issues regarding specific Wikipedia articles. The material is split into two files (JSON-arrays), one each for Bokmål (nb.wikipedia.json) and Nynorsk (nn.wikipedia.json). There are 31364 discussions in Bokmål (approx. 17 mio. words), and 5500 discussions in Nynorsk (approx. 1,4 mio. Words). Each discussion is an element in the JSON array, with one level containing text and metadata. There are eight key/value pairs per discussion:

title:	title of article under discussion
pageid:	text identifier
revid:	audit information
wikidata:	other data
contentcategories:	metadata
hiddencategories:	metadata
text:	discussion text
bytelength:	length of text in number of bytes

Example
The elementet with index value 229 in the Nynorsk array (the 230th element) looks as follows:

```
title:             "Diskusjon:Adeccoligaen"
pageid:            67119
revid:             2874510
wikidata:          false
contentcategories:[]
hiddencategories: []
text:              "Adeccoligaen eller 1. divisjon fotball for menn?\nKva er
                   det best at denne artikkelen heiter. Er du for eller i
                   mot flytting?\n\n For Eg er for flytting av denne
                   artikkelen. Sjå kommentar under. --Cato 14:38, 6 august
                   2007 (CEST)\n For --Ekko 14:49, 6 august 2007 (CEST)\n
                   For --Anders 14:51, 6 august 2007 (CEST)--\n For --Frokor
                   15:05, 6 august 2007 (CEST)\n\nKommentarar\n\nSynest
                   verkeleg at denne sida bør flyttast til Adeccoliagen. Det
                   er dette som er det offisielle namnet på ligaen.
                   Artikkelen om eliteserien heiter jo Tippeligaen og dette
                   blir akkurat det same. Dersom ingen har nokon
                   innvendingar mot dette kjem eg til å flytte artikkelen. -
                   -Cato 14:27, 6 august 2007 (CEST).\n\nSjølv om ligaen
                   heiter Adeccoligaen er det 1. divisjon. Kor ofte kan dei
                   skifte sponsor og namn? Men så lengje det går klart fram
                   av artikkelen kor adeccoligaen er plassert i hierarkiet,
                   og så lenge Cato eller andre passar på at artikkelen er
                   oppdatert, er det vel ikkje noko gale i å flytte til
                   sponsornamnet. - Knut 16:42, 6 august 2007 (CEST)\nHei.
                   Eg flytta artikkelen Adeccoligaen tilbake til 1. divisjon
                   fotball for menn. I 2015 endra denne ligaen namn til
                   OBOS-ligaen. I staden for å oppretta ein ny artikkel med
                   namn OBOS-ligaen, synest eg dette var mest rettmessig. -
                   Sandve 23. april 2016, 15:31"
bytelength:        1283
```