N-gram 1-6, dansk (NST)

Disse n-grammene er laget ved Uni Research AS av Knut Hofland. Utgangspunkt er tekstene i det danske tekstkorpuset til Nordisk språkteknologi holding AS, som ble overført til Nasjonalbiblioteket i 2011.

N-grammene kan benyttes fritt. Brukere oppfordres til å informere Språkbanken om hvilken nytte de gjør av denne ressursen.

Eventuelle spørsmål og tilbakemeldinger kan rettes til sprakbanken@nb.no.

1. Innhold

N-grammene er basert på følgende materiale av dansk nyhetstekst fra siste halvdel av 1990-tallet:

Kilde	Tidsperiode	Antall ord*
Berlingske Tidende	1995-1999	132
Ekstrabladet	1995-1999	51
Politiken	1995-1999	107
Totalt		290

* Millioner ord

- <s> og </s> markerer setningsgrenser.
- Skilletegn er separert med blanke.
- Setninger som ikke er avsluttet med stort skilletegn (.!?;:) er filtrert vekk (overskifter ol.).
- Setninger med mye tall er filtrert vekk. Dette er ofte resultatlister, oppramsinger o.l.

2. Antall linjer

1-gram..... 3.274.757
2-gram..... 35.423.015
3-gram..... 114.114.098
4-gram..... 194.789.520
5-gram..... 241.732.764
6-gram..... 256.862.606

3. Filoversikt

- Filarkivet **ngram_dan_1000.zip** (45 KB, utpakket 141 KB) inneholder de 1000 mest frekvente 1-gram, 2-gram, 3-gram, 4-gram, 5-gram og 6-gram, til sammen seks tekstfiler:

  ngram1-1-topp1000.txt ngram4-1-topp1000.txt
  ngram2-1-topp1000.txt ngram5-1-topp1000.txt
  ngram3-1-topp1000.txt ngram6-1-topp1000.txt

- Filarkivet **ngram_dan.tar** (8 GB, utpakket 35 GB) inneholder følgende filer, alle rene tekstfiler:

Filnavn	Innhold
ngram1.srt	Alle 1-gram, alfabetisk liste
ngram1-1.frk	1-gram, frekvenssortert (fallende), frekvens > 1
ngram1-1.srt	1-gram, alfabetisk liste, frekvens > 1
ngram1-1-topp1000.txt	De 1000 mest frekvente 1-gram
ngram2.srt	Alle 2-gram, alfabetisk liste
ngram2-1.frk	2-gram, frekvenssortert (fallende), frekvens > 1
ngram2-1.srt	2-gram, alfabetisk liste, frekvens > 1
ngram2-1-topp1000.txt	De 1000 mest frekvente 2-gram
ngram3.srt	Alle 3-gram, alfabetisk liste
ngram3-1.frk	3-gram, frekvenssortert (fallende), frekvens > 1
ngram3-1.srt	3-gram, alfabetisk liste, frekvens > 1
ngram3-1-topp1000.txt	De 1000 mest frekvente 3-gram
ngram4.srt	Alle 4-gram, alfabetisk liste
ngram4-1.frk	4-gram, frekvenssortert (fallende), frekvens > 1
ngram4-1.srt	4-gram, alfabetisk liste, frekvens > 1
ngram4-1-topp1000.txt	De 1000 mest frekvente 4-gram
ngram5.srt	Alle 5-gram, alfabetisk liste
ngram5-1.frk	5-gram, frekvenssortert (fallende), frekvens > 1
ngram5-1.srt	5-gram, alfabetisk liste, frekvens > 1
ngram5-1-topp1000.txt	De 1000 mest frekvente 5-gram
ngram6.srt	Alle 6-gram, alfabetisk liste
ngram6-1.frk	6-gram, frekvenssortert (fallende), frekvens > 1
ngram6-1.srt	6-gram, alfabetisk liste, frekvens > 1
ngram6-1-topp1000.txt	De 1000 mest frekvente 6-gram