N-gram 1-6, svensk (NST)

Disse n-grammene er laget ved Uni Research AS av Knut Hofland. N-grammene er laget med utgangspunkt i tekstene i det svenske 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 tekster fra ulike kilder, med en overvekt av nyhetstekst:

Del	Navn	Evt. beskrivelse	Antall ord[+]
01	datateknik30		0.9
02	merge.affarsvarlden		9.3
03	merge.annat (*)		3.3
04	merge.internetworld		2.6
05	merge.LexiLogik	Avistekst	40.5
06	merge.LexiRomaner		3.2
07	merge.ord90	Avistekst	8.1
08	merge.ord91		8.8
09	merge.ord92		27.1
10	merge.ord93		28.5
11	merge.ord94		29.4
12	merge.ord95		29.6
13	merge.ord96		29.2
14	merge.ord97		29.3
15	merge.ord99		29.7
16	merge.samhall	Romaner	26.9
17	ny_teknik		3.2
18	presstext		10.3
19	AdataData		13.5
20	AdataDI	Dagens Industri	23.6
21	AdataDiv		19.5
22	AdataEk		22.8
23	Annat (*)		1.8
24	FoF	Forskning och framsted	1.2
25	ica	Ica-kuriren, dameblad	1.7
26	internet1 (*)		11.6
27	internet2 (*)		19.5
28	internetguide		0.03
29	pdf		1.6
30	SvD	Avistekst	0.2
	Totalt (uten deler merket *)		400.7
	Totalt (inkl. deler merket *)		436.9

[+] Millioner ord

Det er laget n-grammer til hver enkelt del, men deler merket med (*) er ikke med i den sammenflettede delen pga. mye støy. Denne delen, ngram_swe.tar, se §3 under er altså basert på om lag 400 millioner ord. De enkelte delene er allikevel tatt vare på slik at en selv kan flette sammen de delene som en ønsker. Se avsnitt 4 under.

- <s> og </s> markerer setningsgrenser.
- Skilletegn er separert med blanke.
- Setninger som ikke er avsluttet med stort skilletegn (.!?;:) er filtrert vekk (overskifter o.l.).
- Setninger med mye tall (ofte resultatlister), oppramsinger o.l. er filtrert vekk.
- Første ord i en setning er gjort lower-case dersom ordet forekommer i lower-case i den delen som er blitt behandlet før sammenfletting (vanligvis en årgang av hver avis).
- I en del av materialet var det manglende linjeskift, orddeling og sperrete ord (ord med blank mellom hvert tegn). Det er gjort forsøk på å rydde opp i disse forhold.

2. Antall linjer

1-gram:..... 4.238.495
2-gram:..... 50.478.732
3-gram:.... 166.094.677
4-gram:.... 275.472.624
5-gram:.... 327.405.506
6-gram:.... 334.409.426

Filene er sortert på Linux med LC_ALL=POSIX.

3. Filoversikt

- Filarkivet **ngram_swe_1000.zip** (48 KB, utpakket 146 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_swe.tar** (11 GB, utpakket 50 GB) inneholder følgende filer, alle rene tekstfiler:

Filnavn	**Innhold**
ngram1.srt	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	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

Filnavn	Innhold
ngram3.srt	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	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	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	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

4. Fletting av ngram (under Linux)

I filarkivet **ngram_swe_deler.tar** (13GB) ligger hver enkelt del i tabellen i §1 som en egen fil, slik at en selv kan flette sammen de delene en ønsker. Dette kan gjøres med noen skript (**skript.tar.gz**) som også er lagt ved. Alt dette er utarbeidet av Knut Hofland ved Uni Research AS. Fremgangsmåten for hvordan man gjør dette er beskrevet i filen **lesmeg.txt**, og følger også her:

- Pakk ut alle filene i samme katalog.
- Det vil da bli 32 kataloger. "31" og "32" inneholder tomme ngramfiler for flettingen skyld.
- I filene tar frekvens 7 tegn, det er så en blank og deretter ngram.
- I resultatfilene fra fletting tar frekvens 9 tegn.
- For å flette alle katalogene kjører man skriptet jobb-samflett.sh
- Om man ønsker å utelate kataloger ved fletting:
  - Døp om katalogen
  - Kopier en av katalogene med de tomme filene til den aktuelle katalogen
- Eks: man ønsker ikke med katalog 20
  - mv 20 20-old
  - cp -R 31 20
- Man kan så kjøre skriptet.
- Fletteprogrammene er skrevet i Free Pascal: http://www.freepascal.org/