knowit

# Søknad om penger til:
# Stoppordlister for nordsamisk, lulesamisk og sørsamisk for bedre digital samisk språkforståelse

Espen Klem

**Godkjent av**
Petter Hellevik

**Dato**
15.06.2022

# Innholdsfortegnelse

# 1 Prosjektets målsetting/faglig innhold med delmål

Prosjektets målsetting er å kunne generere gode stoppordlister for nordsamisk, lulesamisk og sørsamisk under en åpen kildekode-lisens for alle som vil til å benytte seg av.

Dette så hvem som helst skal kunne lage/utvikle gode teknologiløsninger hvor du trenger en form for språkforståelse for de samiske språkene nordsamisk, lulesamisk og sørsamisk.

## 1.1 Hva er en stoppordliste og hvordan kan den bedre digitale løsninger?

`stopword-sami` blir enkle ordbøker for dataprogrammer. Stoppord er en svartelisting av ord. Dette er ord du ikke ønsker å bruke i digital analyse av en tekst og gir dataprogrammer enkel men kraftig språkforståelse.

Stoppord er ord som brukes ofte og har lite meningsbærende innhold. De er derfor lite egnet til bruk i digital tekstanalyse. Eksempler fra bokmål kan være “og”, “eller”, “men”, “for”, “å”, “en”, “ei”, “et”, men også veldig mange andre ord. Det finnes ingen universell eller definitiv liste over stoppord for et gitt språk. Ei heller gitte regler for hvordan identifisere dem. Hovedregelen er uansett at substantiver og de fleste verb ikke skal med i en stoppordsliste.

En stoppordliste er en liste med typiske stoppord for et gitt formål. Formålet kan være:

- En søkemotor
- En chatbot
- En eller annen form for maskinlæring basert på tekstlig innhold
- Duplikat-/plagiat-identifisering
- Automatisk finne mulige nøkkelord i et dataset hvor hver tekst har en tittel og en brødtekst

## 1.2 Hvorfor er det viktig å kunne identifisere stoppord?

For å ta en søkemotor som eksempel. Det er to hovedårsaker til at fjerning av stoppord er viktige for en søkemotor.

1. Ved å fjerne stoppord vil du få færre ord som skal lagres. Du lagrer bare de som er egnet til å identifisere enkelte dokumenter istedenfor å lagre alle. Dette gir mindre behov for lagring og en raskere søkemotor.
2. Når du søker i en søkemotor vil du ha tilbake de mest relevante treffene. Vanlig er å tillate OR-søk. Resultatet blir da dokumenter med ett eller flere søkeord i seg. Er ett av søkeordene et typisk stoppord og dette ikke er fjernet vil du få veldig mange unødvendige svar fra søkemotoren. Det gir en dårlig brukeropplevelse.

Et annet eksempel er en chatbot. For en chatbot er det viktig å skjønne hovedessensen av det som sies/skrives til den. Da hjelper det å fjerne alle ord som inneholder lite informasjon. Mye av det samme gjelder for maskinlæring basert på tekstlig innhold.

### 1.3 Hvordan genereres stoppordslistene

Hoveddelen av jobben har vært å sette opp en crawler som henter innhold fra kortartiklene til NRK Sápmi.

Innholdet blir så kjørt gjennom et program som kalkulerer hvor stoppords-aktig hvert ord er. Grunnlaget for denne kalkuleringen er hvor mange ganger et ord er brukt i et sett med dokumenter, kombinert med hvor mange dokumenter ordet finnes i.

Den spesifikke beregningen er:

```
stopWordiness =
  (termInCorpus / totDocs) * (1 / (Math.log(totDocs/(termInDocs - 1))))
```

- **termInCorpus** - Antall ganger et ord finnes i en samling av dokumenter
- **totDocs** - Totalt antall dokumenter i dokumentsamlingen
- **termInDocs** - Antall dokumenter et ord finnes i

### 1.4 Kvalitetssikring og -heving - Manuelt arbeid vi søker om penger til

Ikke alt innholdet er samsik

Som nevnt tidligere bruker vi kortartiklene til NRK Sápmi som grunnlag for kalkulering. Vi har tidligere vurdert å bruke Wikipedia som grunnlag for i hvert fall en stoppordsliste for nordsamisk, men for mye av innholdet består av norske ord til at analysen blir spesielt bra. Uansett har også en liten del av innholdet til NRK Sápmi norske ord og setninger. Å identifisere disse og luke dem ut er en del av oppgaven.

Rødlisting av ord

Den automatiske genereringen av stoppord gir oss en rangering av ord fra mest stoppords-aktig til minst. Kanskje prosjektet ser at de 200 første ordene i hovedsak kan brukes som stoppord, men at 20-30 ord innimellom de andre helt klart ikke er stoppord. Disse kan legges til en rødliste, en liste over ord vi helt sikkert ikke vil svarteliste. Vi legger altså til ord til en rødliste fordi vi ikke vil ha dem i stoppordslistene.

Fordelene med å legge disse ordene til en rødliste, istedenfor å bare fjerne de på slutten av prosjektet er at du kan fortsette den automatiske innhentingen av nytt innhold og forbedre stoppordslistene og samtidig holde de rødlistede ordene ute av sluttresultatet.

Rødlister gjør at vi kan øke kvaliteten på stoppordslistene ut over prosjektets tidsrammer. Dette er særlig viktig for lulesamisk- og sørsamisk stoppordsliste hvor mengden tekstinnhold er lite.

Hva er grensen for stoppord og ikke-stoppord - Hvor setter du grensa?

Det automatisk genererte resultatet er en liste fra mest til minst stoppords-aktig. Siste delen av arbeidet er å definere hvor grensen går for hver enkelt liste. Stoppordslistene kan godt gjøres ganske lange siden de er sortert etter mest til minst stoppords-aktige. Du kan da velge å bruke bare en liten del (toppen) av en stoppordsliste hvis du kun er ute etter å fjerne veldig mye brukte ord. Eller du kan velge en mer aggressiv tilnærming og fjerne mange ord for å sørge for f.eks. en mindre søkeindeks (søkemotorens database) og en raskere søkemotor.

## 1.5 Delmål

1. Rødlister over ord som helt sikkert ikke skal være med i stoppordslistene. En liste for hvert av språkene - nordsamisk, lulesamisk og sørsamisk.
2. Publisere stoppordslister for nordsamisk, lulesamisk og sørsamisk som kan tas i bruk. Det viktigste her er å definere hvor mange ord som skal tas med i listen og blir en diskusjon mellom prosjektansvarlig og språkfaglig ansvarlige.
3. Inkludering av stoppordslistene for nordsamisk, lulesamisk og sørsamisk i stoppordsbiblioteket - stopword.
4. (utenfor søknadens prosjektmandat) - Fortsette innhenting av mer innhold i 5 - 10 år til for å se om det gir enda høyere kvalitet på stoppordslistene.

## 1.6 Begrunnelse for søknaden

Mesteparten av programmeringen er allerede gjort. Dette gjelder:

- Innhenting av URLer til tekstdokumenter på nordsamisk, lulesamisk og sørsamisk.
- Innhenting av den faktiske teksten.
- Programatisk analyse av tekstinnholdet for å kalkulere hvor stoppords-aktig hvert ord er.

Koden til dette finnes her:

- stopword-sami - Biblioteket som kommer til å inneholde de ferdige stoppordslistene, samt tekstgrunnlaget.
- nrk-sapmi-crawler - Programmet for å hente innhold fra NRK Sápmi.
- stopword-trainer - Programmet som kalkulerer hvor stoppords-aktig hvert ord er.
- words-n-numbers - Henter ut ord av tekststrenger.

Arbeidet som gjenstår:

- Gjennomgang av innhold hentet fra NRK og luke ut kort-artikler med norsk tekst.
- Finne mennesker som kan lulesamisk og sørsamisk. For nordsamisk har vi allerede en internt som kan bistå - Levi Sørum.
- Koordinere arbeidet med disse menneskene. Forklare hva stoppord er og få dem til å gå gjennom listene.
- Selve arbeidet med å gå gjennom automatisk genererte lister og melde tilbake ord som burde være rødlistede og hvor lange de ferdige listene kan være.
- Legge ord til rødlistene.
- Generere ferdige lister, klare til publisering.
- En siste kvalitetskontroll av listene.
- Publisere stoppordslistene på stopword-sami
- Generere tester og publisere stoppordslistene på stopword
- Bloggpost om prosjektet og hva det ferdige resultatet betyr på Knowit blogg.

# 2 Prosjektbeskrivelse med fremdriftsplan

Ressurser

- Prosjektleder og utvikler: Espen Klem
- Kvalitetssikring av nordsamisk stoppordliste: Levi Sørum
- Kvalitetssikring av lulesamisk og sørsamisk stoppordsliste: [vil bli avgjort etterhvert]

## 2.1 Fremdriftsplan

### 2.1.1 September / oktober

- Gjennomgå innholdet og luke ut norsk tekst - Espen Klem
- Finne mennesker som kan bidra på kvalitetssikring av lulesamisk og sørsamisk - Espen Klem
- Forberede lyntale, NDC Oslo

### 2.1.2 Oktober / desember

- Starte koordinering av arbeidet Levi Sørum og to andre skal gjøre - Espen Klem
- Gjennomgang av ordlistene, definere ord som skal rødlistes og melde tilbake til prosjektet. Levi Sørum + to andre ressurser.
- Legge inn ordene i rødlistene - Espen Klem
- Definere hvor vi skal sette grensene for hver enkelt stoppordsliste. Kort diskusjon med hver enkelt av de ansvarlige for språkforståelse - Alle
- Generere ferdige lister, klare for publisering - Espen Klem
- Siste kvalitetskontroll - Levi Sørum + to andre ressurser.
- Publisere lister til modulene/kode-bibliotekene stopword-sami og stopword - Espen Klem
- Skrive bloggpost om prosjektet og spre på LinkedIn, Reddit, Facebook og Twitter - Espen Klem + ansvarlig for innholdsmarkedsføring i Knowit.

## 2.2 Prosjektets gjennomførbarhet / risikovurdering

Prosjektet har lav risiko. Espen Klem har tidligere gjennomført lignende prosjekter med stoppordslister for finsk og punjabi gurmukhi. Mye av eksisterende kode er

programmert for disse prosjektene. Men tre faktorer kan fremdeles forlenge prosjektet og/eller forringe kvaliteten.

1. Det gjenstår å finne to ressurser, en som kan lulesamisk og en som kan sørsamisk. Nivået trenger ikke være høyere enn for eksempel de fleste nordmenns engelsknivå. Vi kommer til å finne dem, men det kan ta lengre tid en planlagt.
2. Kvaliteten på tekstkilden kan være for lav. Korte tekster er ikke ideelt fordi den ofte får et høyere andel av meningsbærende ord. For å kompensere for dette bruker vi rødlister, og det er sannsynligvis nok. At ikke alle ord som kan defineres som stoppord ikke skulle komme med er ikke et problem så lenge mange nok faktisk gjør det.
3. Mengden tekstinnhold kan være for lav. Dette gjelder særlig lulesamisk og sørsamisk fordi antall artikler vi har tilgjengelig per dato er henholdsvis rundt 600 og 400. For nordsamisk er tallet ca. 1800 og vil være over 4000 i løpet av året, og rundt 8000 i løpet av neste år. Dette er grunnen til at vi legger opp til såpass mange timer brukt på å finne fram til ord som vi rødlister. Tallet på artikler som trengs har grunnlag i tidligere analyse av bruk av Wikipedia for å generere en liste over norske stoppord, men kan kompenseres med å ha laget store rødlister.
4. NRK kan tenke seg å endre oppbygging av siden vi i dag henter tekst-innhold fra. Skjer dette må nrk-sapmi-crawler skrives noe om, men vil ikke utgjøre en stor utfordring.

# 3 Hvordan prosjektet får effekt utover prosjektperiode

Ved å publisere stoppordslistene med en av de mest åpne kildekode-lisensene, MIT License, sikrer vi at alle som vil kan bruke stoppordlisten til hva de vil i all framtid: stopword-sami sin MIT lisens. Alle står fritt til å komme med bidrag til prosjektet, men kanskje aller viktigst: Alle står fritt til å bruke koden i sine egne prosjekter eller å kopiere prosjektet for å fortsette utvikling i egen, ønsket retning. Dette gjelder både åpen og lukket kildekode, vederlagsfritt.

Alle språk i bruk endrer seg over tid. Derfor har prosjektet også kode for å oppdatere listene etter som nytt innhold blir skrevet på NRK Sápmi. Dette gjør at

vi ganske enkelt kan vedlikeholde og forbedre stoppordslistene i de neste 5 - 10 årene.

Prosjektet har liten verdi i prosjektperioden, men vil kunne ha stor verdi etter at prosjektet er over, i lang tid framover.

Espen Klem har siden 2017 vedlikeholdt og utviklet kildekoden til programvare-modulen stopword, som i skrivende stund inneholder stoppordslister for 62 språk. Den har cirka 1 million nedlastinger siste året på NPM - Node Package Manager, i tillegg til tilgjengelighet via andre kanaler. Bruken her er som en del av søkemotorer, chatbot'er, generell maskinlæring, sentimentanalyser, duplikatanalyser, spam-filter, nøkkelord-uthenting mm.

# 4 Hvordan likestillingsperspektivet ivaretas i prosjektet

Vi jobber hardt for være vennlig og inkluderende mot alle som tar kontakt med prosjektet. Vi er her fordi vi synes innholdet er interessant. I tillegg har vi en fellesskapsstandard "Code of conduct" som vi følger strengt. Vi er hjelpsomme mot alle uansett kjønn, legning, alder, kroppsform, synlig og usynlig handicap, etnisitet, nivå på ferdigheter, utdanning, sosioøkonomiske forhold, nasjonalitet, personlig utseende eller religion. Dette er noe vi har erfaring med fra andre åpne kildekode-prosjekter og innad i Knowit som Knowit Amende AS er en del av. Knowits likestillings- og mangfoldspolicy.

# 5 Hvordan prosjektresultatet skal synliggjøres/publiseres

To bloggposter på blogg.knowit.no. En bloggpost om prosjektet før NDC, og en ved lansering. Bloggpostene blir også postet på LinkedIn, Reddit/JavaScript, Facebook og Twitter.

I tillegg til at stoppordslistene blir publisert på stopword-sami, vil de også bli lagt til på kodebiblioteket stopword som har ca. 1 million nedlastinger siste 12 månedene.

Og så håper vi at stopwords-iso vil være interessert i ihvertfall nordsamisk stoppordsliste. De samiske stoppordslistene vil bli foreslått som nytt innhold til dette kodebiblioteket, men det er ikke opp til prosjektet om de blir inkludert. Grunnen til at det er sannsynlig at de bare er interessert i nordsamisk er at de følger en eldre ISO-standard for språkkoder. Denne har bare kode for nordsamisk. Modulen `stopword` har blitt endret tidligere i år for å kunne støtte 3-bokstavs språkkoder sånn at både nordsamisk, lulesamisk og sørsamisk skal kunne legges til.

Espen Klem har meldt inn og fått akseptert en lyntale om prosjektet til NDC: Sami stopwords - How far have we gotten and why does it matter?, september 2022. Tanken er å få fortelle hvorfor vi kjører prosjektet og vise raskt hvordan det kan gi samisk språkforståelse, samt en kjapp demo.

# 6 Budsjett som viser alle kostnadene og finansieringsplan

## 6.1 Finansieringsplan

- 73500 kroner i tilskudd fra Sametinget.

Dette er ment å dekke resterende kostnader i prosjektet som utgjør 49 timer. Det er ingen andre kostnader tilknyttet prosjektet.

Det er ikke søkt midler fra andre steder enn Sametinget.

## 6.2 Gjenstående tidsbruk

- Gjennomgang av innhold hentet fra NRK og luke ut kort-artikler med norsk tekst. **(4 timer)**

- Finne mennesker som kan lulesamisk og sørsamisk - ansvarlige for språkforståelsen. For nordsamisk har vi allerede en internt som kan bistå - Levi Sørum. **(8 timer)**
- Koordinere arbeidet med disse menneskene. Forklare hva stoppord er og få dem til å gå gjennom listene. **(6 timer)**
- Selve arbeidet med å gå gjennom automatisk genererte lister og melde tilbake ord som burde være rødlistede og hvor lange de ferdige listene kan være. 3 timer per person. Levi Sørum + 2 eksterne ressurser. **(9 timer)**
- Legge ord til rødlistene. **(2 timer)**
- Definere hvor vi skal sette grensene for hver enkelt stoppordsliste. Kort diskusjon med hver enkelt av de ansvarlige for språkforståelse. **(3 timer)**
- Generere ferdige lister, klare til publisering. **(1 time)**
- En siste kvalitetskontroll av listene. **(3 timer)**
- Publisere stoppordslistene på stopword-sami **(1 time)**
- Generere tester og publisere stoppordslistene på stopword **(3 timer)**
- Forberede lyntale til NDC Oslo. **(3 timer)**
- Bloggpost om prosjektet og hva det ferdige resultatet betyr. **(6 timer)**

**Totalt: 49 timer * 1200 kr/t + MVA = 73500 kr.**

## 6.3 Tidligere investering i prosjektet

- Fra før av har Knowit finansiert utvikling av nrk-sapmi-crawler og stopword-sami i tidsrommet desember 2021 - april 2022. Dette ved at Espen Klem har brukt av fagtid for å utvikle løsningen. **(44 timer)**
- I tillegg kommer egeninnsats fra Espen Klem. **(50 – 100 timer)**