Oppgavesett 2, INF5820, H2008

Innleveringsfrist 1.10

Vi skal i dette oppgavesettet eksperimentere med ordmeningsentydiggjøring. Vi vil se på norsk språk. Vi vil bruke "bag-of-word"-informasjon, og naive Bayes-klassifikator. Som kontekst vil vi bruke den setningen ordet står i — og altså ikke et vindu av fast størrelse. Siden vi ikke har et korpus med oppmerket mening, vil vi bruke teknikken med pseudoord.

Som utgangspunkt bruker vi en del av LOGON-korpuset som ble utviklet i forbindelse med et prosjekt for oversettelse av turbeskrivelser fra norsk til engelsk. Vi trenger bare de norske setningene, som du finner her: ~jtl/norsk.txt. Korpuset er pent merket opp med en setning per linje.

Som pseudoord velger vi *topp* og *tur*, altså vi tenker oss at vi erstatter dem med et nytt ord *aaa* som har to betydninger — betydning 1 de stedene det har stått *topp* og betydning 2 de stedene det har stått *tur*.

Vi vil se på alle formene av de to ordene, *topp, toppen, topper, toppene* og tilsvarende for *tur*. Husk også begynnelse av setninger med stor bokstav (*Toppen*).

Les gjennom hele settet før du begynner. Det kan ha betydning for valg av strategi, osv.

1 Oppgave

For å gjøre det hele mer håndterbart i fortsettelsen, kan vi plukke ut de setningene som har minst en forekomst av en form av pseudoordet, og bruke dette for resten av eksperimentet. Kall filen med disse setningene for eksempler.txt.

2 Oppgave

Ta ut 25% av eksempler.txt til et test korpus, test.txt, og kall de resterende setningene trening.txt. Legg test.txt til side. Det beste er om test.txt er hver 4. setning.

3 Oppgave

Vi trenger nå nøkkelord til vektoren. En mulighet er å plukke ut de mest frekvente ordene i eksempler.txt. Vi må sortere bort såkalte stoppord (*og*,

er, osv.). I mangel av en stoppordliste kan det gjøres manuelt. Når vi ser på nøkkelord, bør vi se på dem som leksemer, dvs. at alle former av et nøkkelord (*fjell, fjellet, fjellene*) skal telle som en forekomst av det samme nøkkelordet. Uten en lemmatiserer kan det være litt bry å finne eksakt leksemfrekvens. Det er derfor i orden å ta utgangspunkt i de hyppigste ordformene og så ta de tilhørene leksemene (og i neste omgang alle formene av disse leksemene). Vi vil bruke 12 nøkkelord.

4 Oppgave

Vi er nå klar til å trene klassifikatoren vår, altså (modifisert) formel (20.7) og (20.8) i J&M. Ikke glem glatting.

5 Oppgave

Så kan vi programmere klassifikatoren. Husk å bruke logaritmer.

6 Oppgave

Vi kan så teste klassifikatoren. Vi kjører den da på et testmateriale og sammenlikner med fasiten. Hvor mange prosent blir riktig? Hva er en rimelig "base-line" og hvordan gjør vi i forhold til den. Test først klassifikatoren på treningsmaterialet? Selv om den er trent på dette materialet, er det ingen grunn til å tro at den ikke vil gjøre feil.

7 Oppgave

Test til slutt klassifikatoren på test.txt og sammenlikn med en rimelig "base-line", og med resultatet på treningsmaterialet.

Innlevering

Følgende skal leveres inn: eksempler.txt, trenings.txt, test.txt med kort forklaring av hvordan du gikk frem. Nøkkelordene med forklaring av hvordan du fant dem. Resultatene fra oppgave 4 med forklaring om hvordan du fant dem. Koden fra oppgave 5. Resultatene fra oppgave 6 og 7 med forklaring av hvordan du fant dem.