Oppgavesett 2, INF5820, H2010

Innleveringsfrist 8.10

Vi skal i dette oppgavesettet eksperimentere med ordmeningsentydiggjøring. Vi vil bruke "bag-of-word"-informasjon, og naive Bayes-klassifikator. Som materiale skal vi bruke det såkalte "line-" eller "line-hard-serve-"korpuset. Her har man valgt ut et substantiv, et adjektiv og et verb. For hver av dem har man plukket ut et antall setninger (fra Brown-korpuset mm.) som inneholder ordet. Og hver forekomst er merket med en Word-net betydning. Vi skal se på ordet "line".

Korpuset finner du på www.senseval.org. Velg data, deretter "line, hard, serve". På denne siden går vi ned til "line", leser "README" og vi bruker Senseval-1-format.

Filen line.cor består av en rekke oppslag ("entries"). Hvert oppslag består av en setning som inneholder målordet "line" og setningen umiddelbart før.

Hovedalgoritmen (oppgave 4 og 5) kan programmeres i Java, Lisp, Prolog, Python eller Scheme. For de første oppgavene kan du bruke samme språk eller andre hjelpemidler som unixkommandoer eller Perl.

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

1 Oppgave

Først må vi dele korpuset i et treningssett og et testsett. Ta ut 25% av line.cor til testsett. Ta ut hvert fjerde oppslag, start med oppslag 4. Kall fila test.cor og kall de 75% som blir igjen for trening.cor.

2 Oppgave

Vi vil bruke 30 nøkkelord som trekk fra omgivelsene, altså en vektor med 30 plasser som hver kan ta verdien 0 eller 1. Det finnes metoder for å prøve å velge best mulig nøkkelord automatisk. Vi skal nøye oss med å gjøre det manuelt. Nøkkelordene bør på den ene siden være frekvente — sjeldne ord får vi sjeldent bruk for. På den andre siden bør de skille mellom de ulike betydningene, dvs. forekomme hyppig med noen betydninger, sjeldent med andre. En mulighet er som følger. For hver betydning lager vi en liste over ord i omgivelsene ordnet etter frekvens. Så sammenligner vi listene manuelt og prøver å velge gode kandidater.

3 Oppgave

I fortsettelsen av oppgaven kan det være lurt å passe på å "tokenizere" korpuset noe, f.eks. gjøre om store bokstaver i begynnelsen av en setning og skille ut skilletegn som følger rett etter et ord.

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. Test først klassifikatoren på trening.cor. Selv om klassifikatoren er trent på dette materialet, er det ingen grunn til å tro at den ikke vil gjøre feil. Hvor mange prosent blir riktig? Hva er en rimelig "base-line" og hvordan gjør vi i forhold til den?

7 Oppgave

Test deretter klassifikatoren på test.txt og sammenlikn med en rimelig "baseline", og med resultatet på treningsmaterialet.

8 Oppgave

Så langt har vi brukt to setninger som kontekst. Se nå bort fra setningen før og bruk bare setningen som inneholder ordet som kontekst. Gjenta eksperimentet. Hvordan blir resultatet nå sammenlignet med når vi bruker to setninger som kontekst?

9 Oppgave, opsjonell

En betydning *product2* er mye mer frekvent enn de andre. Rykk tilbake til start (før delingen i test- og treningskorpus), fjern de 1800 første oppslagene med *product2* og gjenta resten av eksperimentet. Bruk valget av kontekst (1 eller 2 setninger) som ga best resultat i forrige eksperiment. Hvordan blir nå resultatet i forhold til når du brukte hele korpuset, og i forhold til en rimelig baseline?

Innlevering

Følgende skal leveres inn:

1. trening.cor og test.cor
2. Listen av nøkkelord du har valgt og forklaring på hvordan du valgte dem
3. En opplisting av hvilke normaliseringer du har gjort under tokenisering.
4. Koden til oppgave 4 og oppgave 5
5. Resultatene for oppgave 6–9. Her bør resultattallene presenteres i en tabellform. Og de bør tolkes. Spesielt bør spørsmålene i teksten besvares. En passende tabellform kan være

		Korrekte resultater					
		cord	div	form	phon	prod	text
Klassifikatoren	cord2						
	division2						
	formation2						
	phone2						
	product2						
	text2						