T-61.281 Luonnollisten kielten tilastollinen käsittely

Harjoitus 11, ti 6.4.2004, 8:30-10:00 Klusterointi, Konekääntäminen. Versio 1.0

1. 2-ulotteiseen avaruuteen on projisoitu sanoja. Sanat, jotka esiintyvät samankaltaisissa lauseissa ovat lähekkäin tässä avaruudessa. Piirrä dendrogrammit taulukossa 1 olevien sanojen klusteroinnin muodostumiselle käyttäen

   a) yksinkertaisen linkin (single link) klusterointia
   b) kokonaisen linkin (complete link) klusterointia

koordinaatit	sana
(-4,2)	puukko
(-2,2)	tuppi
(-2,-1)	kaivuri
(-3,-2)	zetor
(1,-3)	kenraali
(2,2)	maija
(2.5,1)	matti
(4,2)	jens

Taulukko 1: *klusteroitavat sanat*

2. Etsi k-means –klusteroinnilla edellisen tehtävän sanoista 3 klusteria. Oletetaan, että klusterien lähtöarvoiksi on arvottu (2,3), (2,0.5) ja (4,1).

3. Taulukossa 2 on kaksiulotteiseen avaruuteen projisoituja sanoja.

   a) Käytetään knn-luokitinta 3:lla naapurilla. Arvioi sanoille, joille ei ole merkitty sanaluokkaa, paras sanaluokka.
   b) Korvaa samaan sanaluokkaan kuuluvien sanojen vektorit niiden keskiarvolla. Vertaa uusia sanoja kahteen muodostuneeseen prototyyppivektoriin. Minkälainen luokittelutulos nyt tulee ?
   c) Mitä etua ja haittoja b)-kohdan menetelmällä on verrattuna a)-kohdan menetelmään ?

4. Olet klusteroinut samankaltaiset sanat ryhmiin. Perinteisessä trigrammimallissa arvioidaan seuraavan sanan todennäköisyys perustuen edellisiin sanoihin ($P(w_n|w_{n-1}, w_{n-2})$). Mallin kokoa rajoittaaksesi haluat nyt kuitenkin arvioida seuraavan sanan todennäköisyyden niin, että käytät historiana vain edellisten sanojen klustereita, etkä sanoja sinällään. Johda tällaisen kieliopin matemaattinen muoto.

sana	luokka	koordinaatit
vetää	verbi	(2.5,3)
työntää	verbi	(1,2)
nostaa	verbi	(3,2)
moukari	substantiivi	(1,1)
naama	substantiivi	(-1,2)
hius	substantiivi	(-5,1)
heittää	?	(2.7,2.7)
kihartuminen	?	(-3,2)
kuula	?	(0.5,2)

Taulukko 2: *knn, data*

w	$P(w)$
it	0.18
becomes	0.05
clean	0.01
eats	0.1
the	0.12
seats	0.02
turns	0.07
into	0.11
a	0.21
reindeer	0.01
and	0.13
lichen	0.01

w_1	w_2	$P(w_1 \rightarrow w_2)$
it	den	1.0
becomes	blir	0.7
becomes	klär	0.3
turns	blir	0.7
turns	vänder	0.3
into	[]	1.0
clean	ren	0.9
clean	städa	0.1
a	[]	1.0
reindeer	ren	1.0
and	och	1.0
eats	äter	1.0
the	[]	1.0
seats	laven	0.1
seats	stolar	0.9
lichen	laven	1.0

Taulukko 3: *Vasemmalla unigrammikielimalli, oikealla käännöstodennäköisyydet.*

5. Etsit ratkaisua hevosmiehiä pitkään pohdituttaneeseen ongelmaan, "Varför får hästen inte gå i bastun?". Ratkaisun ongelmaan tuntevat vain ruotsalaiset ("Den blir ren och äter laven"). Osaat englantia ja käytössäsi on sekä taulukon 3 kielimalli ja käännöstiedot. Sinulla on kaksi vahvaa ehdokasta vastauksen käännökseksi:

   – It becomes clean and eats the seats
   – It turns into a reindeer and eats lichen

   Kumpi on todennäköisempi ?