T-61.281 Luonnollisten kielten tilastollinen käsittely

Harjoitus 3, ti 11.2.2003, 16:15-18:00 – Kollokaatiot, Versio 1.0

Voit valita kuhunkin tehtävään haluamasi alijoukon sanoja, koko taulukon laskemiseen kuluu runsaahkosti aikaa.

1. Taulukkoon 1 on laskettu sanaparin esiintymistiheyksiä erikseen ja toistensa suhteen. Laske annettujen sanojen bigrammitiedoista frekvenssimenetelmällä kollokaatioehdokkaat. Paraneeko tulos jos normalisoit bigrammien määrän sen komponenttien esiintymistiheyksien tulolla ?

2. Laske taulukosta 1 valitsemillesi sanapareille esiintymispaikan keskiarvo ja varianssi. Arvio, miten hyvin näillä tilastotiedoilla voidaan päätellä, ovatko sanat kollokaatioita. Miten tarkasteluikkunan leveys vaikuttaa tuloksiin ? Annetut tilastothan on laskettu viiden sanan kokoisen ikkunan yli.

3. Järjestä vielä taulukon 1 sanat kollokaation todennäköisyyden mukaiseen järjestykseen käyttämällä Pearsonin khi toiseen -testiä ja Dunningin uskattavuussuhde -testiä (Pearson's chi-square test, Dunning's likelihood ratio test). Vertaa tuloksia.

4. Etsi taulukon 1 sanoista kollokaatioita yhteisinformaatiomenetelmällä (Mutual Information). Vertaa tuloksia muiden menetelmien kanssa.

Taulukko 1: *Sanojen esiintymistiheyksiä. $C(a)$ kertoo, kuinka monta kertaa tapahtuma a esiintyi testijoukossa. x:llä on merkitty jotain muuta, kuin sanapariin kuuluvaa sanaa. Aineistossa oli kaiken kaikkiaan 28 181 344 sanaa. Sanat on perusmuotoistettu ennen taulukon laskemista.*

s_1	s_2	$C(s_1)$	$C(s_2)$	$C(s_1, s_2)$	$C(s_1, x, s_2)$	$C(s_2, s_1)$	$C(s_2, x, s_1)$
hakea	työ	10435	26174	31	26	22	11
valkoinen	talo	3665	10767	710	2	1	6
herne	nenä	115	974	3	0	0	0
ja	olla	818046	1387476	7329	39979	3612	38162
venäjä	presidentti	27637	26855	717	216	10	24
vihainen	mielenosoittaja	589	1757	7	0	0	0
tuntematon	sotilas	1967	4806	154	4	0	0
aste	pakkanen	2879	1440	160	8	13	32
heittää	veivi	8126	21	5	0	0	1
kova	tuuli	20613	3916	279	16	9	12
liukas	keli	735	728	106	2	3	7
sekä	myös	50193	135637	138	124	34	244
oppia	lukea	2831	8952	21	4	7	1
olla	ula	1387476	44	3	2	1	2
ottaa	onki	38304	110	9	3	0	0