

T-61.281 Luonnollisten kielten tilastollinen käsittely

Harjoitus 9, ti 25.3.2003, 16:15-18:00 Samankaltaisuusmitat, Versio 1.0

1. Kevätflunssaa odotellessa Teemu T. Teekkari testaili flunssalääkkeitä. Kokeiltavana olivat Tintus-yskänlääke, Koskisen Korvalääke ja Otaniemen Termiitti. Kutakin lääketä tarkkaan maistellessaan hän samalla kuvaili makutuntemuksiaan. Paikalla ollut virallinen tarkkailija kirjasi 5 valitun adjektiivin kohdalta ylös, kuinka usein Teemu lääketä kuvaillessaan käytti tätä adjektiivia.

	raikas	hapokas	makea	hedelmäinen	pehmeä
Tintus	0	0	5	1	4
Korvalääke	10	6	2	1	0
Termiitti	1	4	3	3	3

Taulukko 1: Dokumentti-sana –matriisi

Laske kunkin lääkkeen etäisyydet toisistaan käyttäen kaikkia allalistattuja mittoja:

- a) Euklidinen etäisyys
- b) L_1 -normi
- c) Kosini
- d) Informaatiosäde

Miksi Kullback-Leibler –divergenssin käyttö olisi epäkäytännöllistä tässä tehtävässä ?

2. Tarkastellaan seuraavia mittoja

- a) Kullback-Leibler –divergenssi
- b) Informaatiosäde
- c) L_1 -normi

Jos yhden mitan mukainen etäisyys on nolla, tarkoittaako se, että myös muiden mittojen mukaan etäisyys on nolla ?

3. Tarkastellaan edelleen toisessa tehtävässä annettuja mittoja. Etsi kullekin mitalle jakaumat, jotka antavat suurimman mahdollisen etäisyyden. *Vinkki: Informaatiosädeelle suurin mahdollinen etäisyys on $2 \log 2$.*