T-61.184 Informaatiotekniikan erikoiskurssi IV

Harjoitukset artikkelista *Topic-Based Language Models Using EM*
Teemu Hirsimäki, 9.11.2001

1. Pohdi lyhyesti trigrammien ja artikkelissa esitetyn topic-mallin soveltumista suomen ja englannin kieleen.

2. Olet juuri estimoinut EM-algoritmilla seuraavat sanajakaumat aiheille t_1 ja t_2: Testivaiheessa

Sanat	$P(w\|t_1)$	$P(w\|t_2)$
Avaruus	0.4	0.0
Tähti	0.3	0.2
Konsertti	0.0	0.6
Olla	0.3	0.2

todennäköisyyksien $P(t|h)$ estimointiin käytetään artikkelissa esitettyä online-algoritmia. Aiheet oletetaan etukäteen yhtä todennäköisiksi. Huomaa, että kaavassa paperin kaavassa (7) pitäisi olla $P(t|h_0)$ eikä $P(t|h_1)$.

   (a) Millaiset todennäköisyydet $P(t_1|h)$ ja $P(t_2|h)$ saadaan, kun historia koostuu pelkästään sanasta "tähti"?

   (b) Millaisen jakauman malli antaa tällöin toiselle sanalle?

   (c) Saadaanko aiheen 1 todennäköisyys nollaan, jos historia koostuu pelkästään "konsertti"-sanoista?