Latentti semanttinen indeksointi - esimerkkitapaus

Tietojenkäsittelyteorian erikoistyö Tik-79.189

Eerika Savia, 39482a, Tf

28.8.1998

Sisältö

1 Johdanto

Tässä työssä kokeiltiin Latenttia semanttista indeksointia WSOY:n Yritystietoteoksen osiin esimerkkitapauksena. Latentti semanttinen indeksointi (LSI) on lineaarialgebraan perustuva menetelmä, joka on kehitetty tekstipohjaisten informaation suodatus- ja hakumenetelmien parantamiseksi [1], [2], [3]. Sen tarkoitus on kuvata dokumentteja tekstin merkityksen ja käsitteiden avulla kirjaimellisten sanaesiintymien sijasta. LSI:n pyrkimys on kuvata samalle käsitteelle samaa tarkoittavat sanat eli synonyymit ja toisaalta erottaa toisistaan muodoltaan samanlaiset sanat, joilla on eri merkitys.

Matemaattisesti LSI perustuu matriisin singulaariarvohajotelmaan, joka muistuttaa ominaisarvohajotelmaa.

Tarkoituksena oli tutkia, ovatko analyysin perusteella löytyvät termiavaruuden vektorit ymmärrettäviä ja löytyykö analyysin avulla termien välisiä riippuvuuksia, jotka implikoivat teoksen latenttia semanttista rakennetta.

2 Tausta

2.1 Informaation suodatus ja haku

LSI:n taustalla on informaation suodatuksessa (Information Filtering IF) ja haussa (Information Retrieval IR) yleisesti käytetty *vektoriavaruusmalli* (tai *termivektorimenetelmä*).

2.1.1 Vektoriavaruusmalli

Vektoriavaruusmalli perustuu sanojen esiintymisfrekvensseihin tekstissä, joten se soveltuu lähinnä tekstidokumenttien sisällön esittämiseen. Tekstissä esiintyvät sanat palautetaan perusmuotoihinsa ja liian yleiset sanat poistetaan. Jäljelle jääviä, tekstiä kuvaavia sanoja kutsutaan termeiksi. Dokumentin esityksenä käytetään vektoria, jossa kukin komponentti vastaa yhtä termiä *(termivektori)* . Kullakin termillä on oma esiintymistiheydestä riippuva painonsa vektorissa. [4], [5] Termien määrä voi nousta kymmeniin, jopa satoihin tuhansiin, vaikka useimmiten pienempikin avaruus riittäisi kuvaamaan kaikkia dokumentteja. Vektoriavaruusmallin perusoletuksiin kuuluu, että vektorin komponentit (termit) ovat ortogonaalisia, eli termien esiintymät ovat toisistaan riippumattomia, mutta se ei tyypillisesti pidä paikkaansa. Samanmuotoisella sanalla voi olla useita eri merkityksiä (homonyymit) ja useat sanat voivat tarkoittaa samaa asiaa (synonyymit). Lisäksi tiettyjen sanojen esiintyminen tekstissä voi tehdä muiden, niihin liittyvien sanojen esiintymisen selvästi todennäköisemmäksi.

Useimmissa sovelluksissa dokumentin termivektoria d verrataan käyttäjän kiinnostusta kuvaavaan käyttäjäprofiilin termivektoriin u (tai hakukyselyä esittävään termivektoriin) vektorien välisen kulman kosinia käyttäen [5]. Kosini saadaan kaavasta

$$\cos(u, d) = \frac{u \cdot d}{|u| \cdot |d|} = \frac{\sum_k u_k d_k}{\sqrt{\sum_k u_k^2} \cdot \sqrt{\sum_k d_k^2}}. \tag{1}$$

Jotta kosinimitta tuottaisi järkeviä tuloksia, vektorikomponenttien kuvaamien termien tulisi olla ortogonaalisia ja kantavektoreiden normeerattuja. Näin ollen redundantista tiedosta on muutakin haittaa kuin tarpeettoman suuri dokumenttiavaruuden dimensio.

2.2 Matemaattinen tausta

Matemaattisesti latentti semanttinen indeksointi perustuu matriisin *singulaariarvohajotelmaan* (Singular Value Decomposition, SVD). Mille tahansa äärellisulotteiselle vektoriavaruudelle on mahdollista muodostaa kanta, jonka avulla voidaan lausua kaikki avaruuden vektorit. Jos kyseessä on *sisätuloavaruus* (ks. Liite A), kanta voidaan myös ortonormeerata. Singulaariarvohajotelma etsii sekä matriisin pystyvektorien että vaakavektorien virittämälle avaruudelle ortonormaalin kannan.

2.2.1 Singulaariarvohajotelma (SVD)

Olkoon

$$\underset{t \times d}{A}$$

reaalinen $t \times d$-matriisi. Tällöin on olemassa matriisit U, V ja Σ siten että

$$\underset{t \times d}{A} = \underset{t \times \rho}{U} \quad \underset{\rho \times \rho}{\Sigma} \quad \underset{\rho \times d}{V}^T \tag{2}$$

missä $\rho = A$:n rangi , jolle pätee $\quad \rho \leq \min\{t, d\} \quad$ ja lisäksi $\quad U^T U = I_\rho$, $V^T V = I_\rho, \quad \Sigma = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_\rho\} \quad$ ja $\quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_\rho > 0$. [6]

Singulaariarvohajotelmassa singulaariarvot asettuvat diagonaalimatriisiin Σ suuruusjärjestykseen siten, että suurin singulaariarvo $\quad \sigma_{max} = \sigma_1 \quad$ on ensimmäisenä ja niitä vastaavat singulaarivektorit ovat samassa järjestyksessä. Matriisin U pystyvektorit ovat nk. vasemmanpuoleiset singulaarivektorit ja vastaavasti matriisin V pystyvektorit ovat oikeanpuoleiset singulaarivektorit. U:n pystyvektorit muodostavat kannan A:n pystyvektorien virittämälle avaruudelle ja V:n pystyvektorit vastaavasti A:n vaakavektorien virittämälle avaruudelle. Erikoistapauksessa $U = V$ singulaariarvohajotelma yhtyy ominaisarvohajotelmaan.

2.2.2 Vektorin esittäminen uudessa kannassa

Kunkin A:n pystyrivin esitys uudessa kannassa U on Σ kertaa vastaava V^T:n pystyrivi. Alkuperäisen t-ulotteisen avaruuden vektorin y esitys $\hat{y}$ uudessa ortonormaalissa ρ-ulotteisessa kannassa saadaan seuraavasti:

$$y = U \Sigma v^T = U \hat{y} \qquad \qquad || \cdot U^T \text{vas.}$$

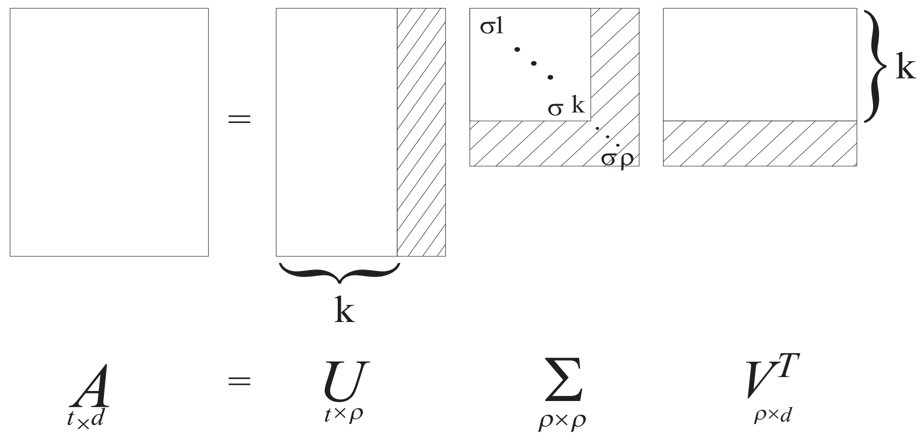
$$U^T y = (U^T U) \hat{y} \qquad \qquad || \ (U^T U = I)$$

$$\hat{y} = U^T y. \tag{3}$$

3 Latentti semanttinen indeksointi

LSI lähtee olettamuksesta, että termivektoriavaruuden todellinen dimensio on pienempi kuin termien lukumäärä ja että alla piilee latenttia semantiikkaa, joka saadaan esille faktorianalyysillä. Jos käytettävissä on riittävän suuri ja edustava dokumenttikokoelma (joukko termivektoreita), voidaan ajatella dokumenttijoukon edustavan koko dokumenttiavaruutta ja tutkia, millaisen avaruuden se virittää. Samalla saadaan selville, missä määrin kyseisen vektorijoukon termeissä on redundanssia.

LSI:ssä matriisin A pystyriveiksi asetetaan d kappaletta dokumenttivektoreita, joissa on t termiä. Syötteenä annettavien dokumenttien termivektoriesitysten joukko virittää ρ-ulotteisen avaruuden S (ρ on A:n rangi $\leq \min\{t, d\}$).Tälle termivektoriavaruudelle S saadaan ortonormaali kanta U tekemällä syötematriisille A singulaariarvohajotelma. Lisäksi alkuperäistä dokumenttiavaruutta voidaan approksimoida pienemmällä avaruudella poimimalla vain k kappaletta dominoivimpia kantavektoreita mukaan kantaan. Latentti semanttinen indeksointi pudottaa matriisista pois vähiten merkitsevät kantavektorit ja ottaa vain riittävän suurta singulaariarvoa vastaavat mukaan , kaava (4), (kuva 1).



Kuva 1: Approksimaation A_k muodostaminen: U:sta poimitaan k ensimmäistä pystyvektoria, V^T:sta k ensimmäistä vaakavektoria ja Σ:sta $k \times k$-lohko, joka käsittää k suurinta singulaariarvoa.

Singulaariarvot σ_i ovat vähenevässä järjestyksessä, joten k ensimmäistä U:n pystyvektoria ovat merkitsevimmät. Näin saadaan rank-k -approksimaatio A_k matriisille A [1], [2], [3].

$$\underset{t\times d}{A} \approx \underset{t\times d}{A_k} = \underset{t\times k}{U_k} \quad \underset{k\times k}{\Sigma_k} \quad \underset{k\times d}{V_k}^T \tag{4}$$

Tämä A_k on Frobenius-normin mielessä paras mahdollinen rank-k -approksimaatio A:lle [3] eli

$$\min_B \|A-B\|_F = \|A-A_k\|_F \quad , \quad \text{missä} \quad \|A\|_F = \sqrt{\sum_i \sum_j |a_{ij}|^2}.$$

U_k:n pystyvektorit kuvaavat dokumentti- avaruutta termien avulla, joten jokaisen vektorin pituus kattaa kaikki käytössä olevat termit. (Vastaavasti V_k:n pystyvektorit kuvaavat termiavaruutta dokumenttien avulla.)

Uusi profiili tai dokumentti q on LSI:n tuottaman kannan U_k avulla lausuttuna (kaavan (3) mukaan) :

$$\hat{q} = U_k^T q. \tag{5}$$

3.1 LSI:n käyttötarkoituksia

LSI on kehitetty parantamaan informaation suodatusmenetelmiä ja tekstipohjaisia hakumenetelmiä. Vaikka LSI:ssä syntyy ortogonaalinen kanta termivektoreille, tyypillisesti tavoitteena ei ole ymmärtää saatujen kantavektoreiden merkitystä. Sen sijaan halutaan lausua suodattimina toimivat käyttäjäprofiilit (IF) - tai hakujen tapauksessa kyselyt (IR) - uuden kannan avulla ja verrata niitä dokumentteihin, jotka on myös lausuttu uuden, ortogonaalisen kannan avulla. Tällöin jokaisen pois jätetyn kantavektorin mukana menetetään käyttökelposta informaatiota, joten vektoreita poimitaan tyypillisesti mukaan 100-300 eikä vain muutamia (kun dokumenttien lkm $\approx$ 1000-2000). Loppujen kantavektoreiden katsotaan edustavan kohinaa [1]. Jos halutaan tutkia ja ymmärtää kantavektoreiden merkitystä, on käytännössä rajoituttava pienempään määrään vektoreita.

4 LSI - esimerkkitapaus

4.1 WSOY Yritysjulkaisut: Yritystieto-sarja

WSOY:n teos nimeltä Yritystieto koostuu 8 kansiosta, jotka sisältävät erilaista yrityksiä koskevaa tietoa [7]. Näistä neljä saatiin mukaan kokeiluun, jossa kansioiden lukuihin ja alilukuihin sovellettiin eräänlaista termivektorimallia. Termeinä toimivat teoksen hakusanat ja paino kertoi yksinkertaisesti, montako kertaa hakusanaan oli viittaus ko. tekstissä. Näihin hakusanavektoreihin kokeiltiin latenttia semanttista indeksointia. Mukana olivat kansiot Arvonlisäverotus (AV), Ennakkoperintä (EP), Kirjanpito (KP) sekä Yhtiöoikeus (YO).

Tarkastelu tehtiin sekä koko neljän kansion kokonaisuudelle että kullekin kansiolle erikseen. Hakusanoja oli kaikkiaan 9498, joista vain 283 (alle 3 %) esiintyi useammassa kuin yhdessä kansiossa. Tästä syystä kansioiden tarkastelu myös erillisinä oli mielekästä.

Tällaisessa kokonaisuudessa dokumentin käsite ei ole yksiselitteinen vaan julkaisu voidaaan jakaa eri mittaisiin osiin, jotka katsotaan erillisiksi dokumenteiksi informaation suodatus/haku -mielessä. Sitä, miten pieniin paloihin teksti on jaettu, kutsutaan *granulariteetiksi*. Teksti jaettiin dokumenteiksi viidellä eri granulariteetilla : kansio, luku, 2. tason otsikko, 3. tason otsikko ja marginaaliotsikko. Taulukossa 1 on dokumenttien lukumäärät eri granulariteeteilla.

Dokumenttien lukumäärä					
	Kansioittain kansio ja termien lukumäärä				Koko data, termien lkm
granulariteetti	AV, 2903	EP, 2999	KP, 1498	YO, 2381	9498
kansio					4
luku	17	18	14	17	66
otsikko 2	129	206	101	131	567
otsikko 3	335	459	195	391	1380
marginaaliotsikko	688	848	418	788	2742

Taulukko 1: Dokumenttimatriisien dimensiot

4.2 LSI:n käyttötarkoitus tässä työssä

Tässä työssä oli tarkoitus tutkia, miten LSI toimii ortonormaalien kantavektoreiden etsinnässä edellä kuvatussa esimerkkitapauksessa eli millainen kanta dokumenttiavaruudelle saadaan ja kuinka paljon hakusanaesityksessä on redundanssia. Haluttiin myös selvittää, ovatko tuloksena saatavat kantavektorit ymmärrettäviä ja voidaanko ne tulkita jonkinlaisiksi käsitteiksi. Koska kyseistä materiaalia ei ole käytetty tavanomaiseen termivektoreiden pohjalta tapahtuvaan profiileihin vertaamiseen, mitään käyttäjäprofiileja ei ollut käytettävissä, eikä niin muodoin ollut mahdollista vertailla etäisyyslaskentaa LSI:n avulla ja ilman.

4.3 Laskennan käytännön toteutus

Kaikille kansioille erikseen sekä koko datalle laskettiin 5 eri granulariteetilla singulaariarvohajotelma, josta talteen otettiin vain singulaariarvot ja vasemmanpuoleiset singulaarivektorit (matriisi U). Matriisi U oli välttämätöntä tallettaa binäärisenä, koska tekstitiedostoista olisi tullut usean gigatavun kokoisia (Taulukko 2).

Avaruuden dimensiota karsittiin singulaariarvojen suhteellisten suuruuksien perusteella. Kynnysehto mukaan otettaville kantavektoreille oli $\sigma_i > \frac{3}{4}\sigma_{max}$ (Taulukko 3) ja kullekin mukaan otetulle vektorille tutkittiin, millaisilla termeillä painon itseisarvo oli suhteellisesti suuri. Näiden kantavektoreiden oletettiin kuvaavan joitakin käsitteitä ja kuhunkin vektoriin liittyvien termien tulkittiin kuvaavan käsitteen pääpiirteitä. Kunkin vektorin termit järjestettiin suurimmasta painosta alkaen ja käsitteitä kuvaavia termilistoja saatiin poimimalla vain itseisarvoltaan suurimpien painojen termit.

Laskenta toteutettiin CSC:n koneessa caper (DEC AlphaServer 8400) käyttäen LAPACK-aliohjelmakirjaston FORTRAN-rutiinia nimeltä DGESVD, joka laskee matriisin SVD:n kaksinkertaisella tarkkuudella. Laskenta-aika näyttäisi kasvavan syötematriisin rangin mukana $\mathcal{O}(\rho^2)$.

Binäärisen tulostiedoston koko, Mb					
	kansio ja termien lukumäärä				Koko data,
granulariteetti	AV, 2903	EP, 2999	KP, 1498	YO, 2381	9498 termiä
luku	0.098	0.11	0.045	0.081	5.1
otsikko 2	0.84	1.4	0.37	0.69	48.6
otsikko 3	2.4	3.2	0.72	2.2	126.4
marginaaliotsikko	7.4	8.4	2.3	7.0	375.4

Taulukko 2: Matriisin U sisältävä tiedosto oli tehokkuussyistä talletettava binäärisenä.

Suurimpien singulaariarvojen lukumäärä $\sigma_i > \frac{3}{4}\sigma_{max}$					
	Kansioittain				Koko data
granulariteetti	AV	EP	KP	YO	
kansio					3
luku	4	6	2	4	12
otsikko 2	3	8	5	6	4
otsikko 3	5	8	6	3	6
marginaaliotsikko	4	5	8	2	4

Taulukko 3: Koska tuloksena saatavia kantavektoreita oli tarkoitus tulkita ja ymmärtää, keskityttiin vain kohtuullisen pieneen määrään merkitsevimpiä kantavektoreita. Kynnysehtona oli, että singulaariarvoille päti $\sigma_i > \frac{3}{4}\sigma_{max}$.

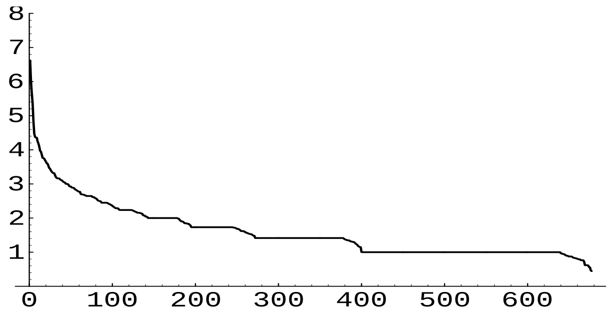
5 Tulokset

5.1 Esimerkki kantavektoreista: Arvonlisäverotus-kansion marginaaliotsikkotaso

Tarkastellaan esimerkin vuoksi Arvonlisäverotus-kansion marginaaliotsikkotason dokumenteille saatuja kantavektoreita ja singulaariarvoja.

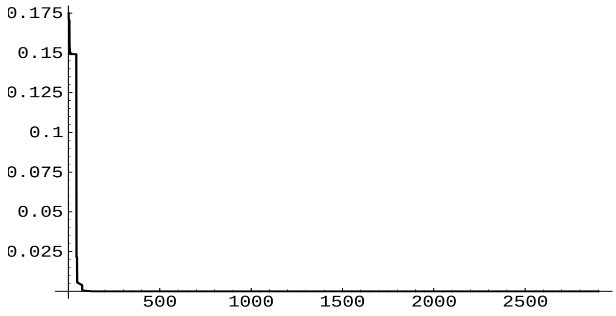
5.1.1 Singulaariarvot ja kantavektorit

Kuvassa 2 näkyy, miten singulaariarvot laskevat 6.6:sta noin 3:een nopeasti. Tämän jälkeen arvot vähenevät tasaisemmin häviämättä jyrkästi nollaan.

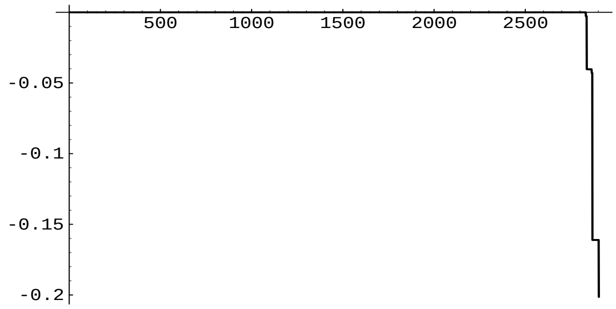


Kuva 2: Singulaariarvot (677 kpl): AV-kansio, marginaaliotsikkotaso.

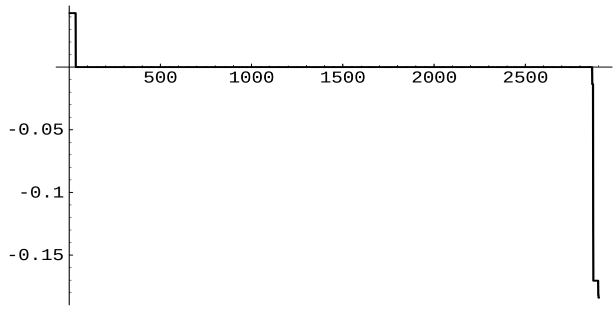
Neljä ensimmäistä singulaariarvoa täytti vaaditun ehdon ja siten neljä ensimmäistä kantavektoria otettiin lähempään tarkasteluun. Kun näiden vektoreiden termit järjestettiin painojen suuruusjärjestyksen mukaan, saatiin kantavektoreita kuvaavat termilistat. Kuvissa 3 - 6 näkyy, miten painot jakautuivat eri termeille. Selvästi kukin vektori liittyy vain tiettyyn, pieneen joukkoon termejä. Termit järjestettiin painon mukaiseen järjestykseen kullekin vektorille erikseen.



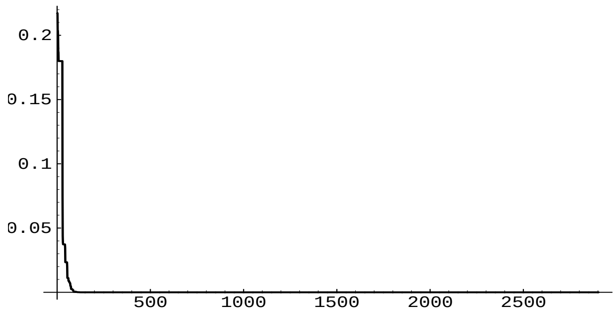
Kuva 3: 1. kantavektori, Arvonlisäverotus-kansio, marginaaliotsikkotaso, painot suuruusjärjestyksessä.



Kuva 4: 2. kantavektori, Arvonlisäverotus-kansio, marginaaliotsikkotaso, painot suuruusjärjestyksessä.



Kuva 5: 3. kantavektori, Arvonlisäverotus-kansio, marginaaliotsikkotaso, painot suuruusjärjestyksessä.



Kuva 6: 4. kantavektori, Arvonlisäverotus-kansio, marginaaliotsikkotaso, painot suuruusjärjestyksessä.

5.1.2 Termilistat

Kuvissa 3 - 6 näkyvät piikit poimittiin termilistaksi kullekin vektorille. Arvonlisäverotus-kansion 4 ensimmäisen vektorin termilistat ovat sivuilla 13 - 16.

- Ensimmäinen kantavektori liittyy vapaa-ajan ja urheilupalveluihin.
- Toinen kantavektori liittyy kaikenlaisiin raaka-aineisiin.
- Kolmannessa kantavektorissa tehdään ero ruoka-aineiden ja muiden raaka-aineiden välille. Yhdistämällä vektoreita 2 ja 3 saadaan esitettyä erikseen ruoka-aineet, muut raaka-aineet tai niiden yhdiste.
- Neljäs kantavektori käsittää yleishyödylliset palvelut ja yhdistykset.

(Vektoreiden suuntaa ei ole erityisesti valittu vaan niitä voi halutessaan kertoa (-1):llä ja kyseessä on olennaisesti sama kanta.)

1. kantavektori, itseisarvoltaan suurimmat painot Arvonlisäverotus-kansio, marginaaliotsikkotaso			
hakusana	paino	hakusana	paino
Kuntosalipalvelu	1.75E-01	Koirakoulutus	1.49E-01
Laskettelurinne	1.70E-01	Laskuvarjohyppykoulutus	1.49E-01
Keilahalli	1.70E-01	Kynsien hoito	1.49E-01
Squash-halli	1.70E-01	Kiropraktikkotoiminta	1.49E-01
Tennishalli	1.70E-01	Kissojen hoitola	1.49E-01
Saunapalvelu	1.54E-01	Naprapatiahoito	1.49E-01
Biljardisali	1.53E-01	Eläintäysihoitolan pitäminen	1.49E-01
Tanssikoulu	1.52E-01	Koirien hoitola	1.49E-01
Harmonikkakoulutoiminta	1.49E-01	Hiihtokeskustoiminta	1.49E-01
Autokoulutoiminta	1.49E-01		
Balettikoulutoiminta	1.49E-01		
Hevostallin pitäminen	1.49E-01		
Moottoripyöräkoulu	1.49E-01		
Liikuntakoulutus	1.49E-01		
Solariumpalvelu	1.49E-01		
Kauneudenhoito: verollinen palvelu	1.49E-01		
Hypnoosikäsittely	1.49E-01		
Jalkojenhoito	1.49E-01		
Luontaishoitola	1.49E-01		
Karatesalin pitäminen	1.49E-01		
Kielilaboratoriotoiminta	1.49E-01		
Mannekiinikoulu	1.49E-01		
Lentokoulutoiminta	1.49E-01		
Kieliopistotoiminta	1.49E-01		
Kuntokoulutoiminta	1.49E-01		
Parturitoiminta	1.49E-01		
Golfradan pitäminen	1.49E-01		
Kylpylä	1.49E-01		
Sukelluskoulutus	1.49E-01		
Kampaamo	1.49E-01		
Musiikkikoulutus	1.49E-01		
Kielikurssitoiminta	1.49E-01		
Judosali	1.49E-01		
Uimahallin pitäminen	1.49E-01		

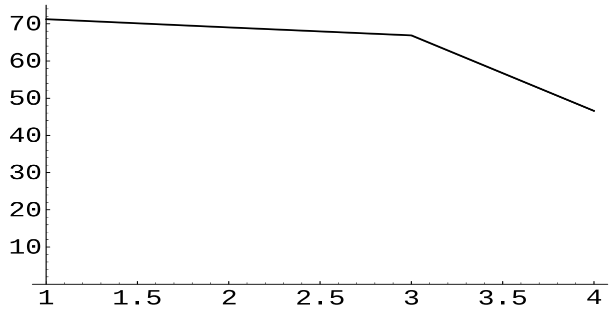
2. kantavektori, itseisarvoltaan suurimmat painot Arvonlisäverotus-kansio, marginaaliotsikkotaso			
hakusana	paino	hakusana	paino
Pähkinä	-2.01E-01	Kaakaopapu	-1.61E-01
Kemikaali	-1.61E-01	Kahvi:	
Jyvä	-1.61E-01	verovarastointimenettely	-1.61E-01
Kasvisrasva	-1.61E-01	Rapu	-4.30E-02
Indium	-1.61E-01	Kala	-4.30E-02
Kasvisöljy	-1.61E-01	Sieni	-4.26E-02
Öljyhedelmä	-1.61E-01	Marja	-4.26E-02
Palladium	-1.61E-01	Herne	-4.03E-02
Lyijy	-1.61E-01	Kumina	-4.03E-02
Raakasokeri	-1.61E-01	Jalosteet	-4.03E-02
Alumiini	-1.61E-01	Maito:verokanta	-4.03E-02
Rodium	-1.61E-01	Liha:verokanta	-4.03E-02
Vilja: verovarastointimenettely	-1.61E-01	Tattari	-4.03E-02
Nikkeli	-1.61E-01	Banaani	-4.03E-02
Kupari	-1.61E-01	Viiriäisen muna	-4.03E-02
Öljynsiemen	-1.61E-01	Katajanmarja	-4.03E-02
Platina	-1.61E-01	Rapsi	-4.03E-02
Petroleumiraakaöljy	-1.61E-01	Sinapinsiemen	-4.03E-02
Oliivi	-1.61E-01	Juures	-4.03E-02
Selluloosa	-1.61E-01	Appelsiini	-4.03E-02
Peruna: verovarastointimenettely	-1.61E-01	Rypsi	-4.03E-02
Kookospähkinä	-1.61E-01	Tee:verokanta	-4.03E-02
Tee: verovarastointimenettely	-1.61E-01	Maissi	-4.03E-02
Soijapapu	-1.61E-01	Peruna:verokanta	-4.03E-02
Parapähkinä	-1.61E-01	Vihannes	-4.03E-02
Kumi	-1.61E-01	Hedelmä	-4.03E-02
Propaani	-1.61E-01	Kaviaari	-4.03E-02
Butaani	-1.61E-01	Kaakao	-4.03E-02
Mineraaliöljy	-1.61E-01	Sokerijuurikas	-4.03E-02
Sinkki	-1.61E-01	Riisi	-4.03E-02
Hopea	-1.61E-01	Kahvi:verokanta	-4.03E-02
Cashew-pähkinä	-1.61E-01	Kananmuna	-4.03E-02
Siemen	-1.61E-01	Hunaja	-4.03E-02
Villa	-1.61E-01	Hirssi	-4.03E-02

3. kantavektori, itseisarvoltaan suurimmat painot Arvonlisäverotus-kansio, marginaaliotsikkotaso			
hakusana	(posit.)paino	hakusana	(negat.)paino
Alumiini	4.30E-02	Rapu	-1.84E-01
Selluloosa	4.30E-02	Kala	-1.84E-01
Sinkki	4.30E-02	Marja	-1.81E-01
Parapähkinä	4.30E-02	Sieni	-1.81E-01
Öljynsiemen	4.30E-02	Maito:verokanta	-1.70E-01
Nikkeli	4.30E-02	Jalosteet	-1.70E-01
Rodium	4.30E-02	Vihannes	-1.70E-01
Butaani	4.30E-02	Kaakao	-1.70E-01
Villa	4.30E-02	Riisi	-1.70E-01
Vilja: verovarastointimenettely	4.30E-02	Viiriäisen muna	-1.70E-01
Hopea	4.30E-02	Maissi	-1.70E-01
Palladium	4.30E-02	Sinapinsiemen	-1.70E-01
Mineraaliöljy	4.30E-02	Rapsi	-1.70E-01
Oliivi	4.30E-02	Rypsi	-1.70E-01
Öljyhedelmä	4.30E-02	Tattari	-1.70E-01
Petroleumiraakaöljy	4.30E-02	Sokerijuurikas	-1.70E-01
Tee: verovarastointimenettely	4.30E-02	Hedelmä	-1.70E-01
Raakasokeri	4.30E-02	Peruna:verokanta	-1.70E-01
Siemen	4.30E-02	Tee:verokanta	-1.70E-01
Peruna: verovarastointimenettely	4.30E-02	Kananmuna	-1.70E-01
Lyijy	4.30E-02	Kumina	-1.70E-01
Kupari	4.30E-02	Hunaja	-1.70E-01
Cashew-pähkinä	4.30E-02	Hirssi	-1.70E-01
Platina	4.30E-02	Kahvi:verokanta	-1.70E-01
Kookospähkinä	4.30E-02	Juures	-1.70E-01
Soijapapu	4.30E-02	Kaviaari	-1.70E-01
Propaani	4.30E-02	Katajanmarja	-1.70E-01
Kahvi: verovarastointimenettely	4.30E-02	Liha:verokanta	-1.70E-01
Kaakaopapu	4.30E-02	Herne	-1.70E-01
Indium	4.30E-02	Banaani	-1.70E-01
Kumi	4.30E-02	Appelsiini	-1.70E-01
Kasvisöljy	4.30E-02	Pähkinä	-1.27E-01
Jyvä	4.30E-02		
Kasvisrasva	4.30E-02		
Kemikaali	4.30E-02		

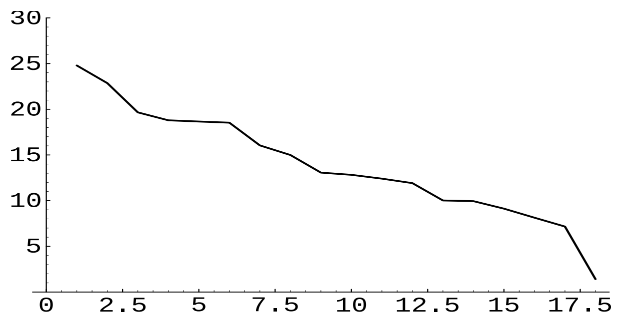
4. kantavektori, itseisarvoltaan suurimmat painot Arvonlisäverotus-kansio, marginaaliotsikkotaso	
hakusana	paino
Käsi- ja taideteollisuusyhdistys	2.17E-01
Matkamuisto	2.17E-01
Urheilukilpailu	2.03E-01
Tanssitilaisuus	2.03E-01
Urheilutilaisuus	1.99E-01
Ruoka-annos:tarjoilu	1.87E-01
Vanhainkoti	1.86E-01
Invalidihuoltolaitos	1.79E-01
Huvitilaisuus	1.79E-01
Huoltola	1.79E-01
Jäsenlehti:yleishyödyllinen yhteisö	1.79E-01
Mainosten jakelu	1.79E-01
Jäsenluettelo, yleishyödyllinen yhdistys	1.79E-01
Viiri	1.79E-01
Rakennusnäyttely	1.79E-01
Työlaitos	1.79E-01
Tavarankeräys	1.79E-01
Jäsenpalvelut	1.79E-01
Vajaamielislaitos	1.79E-01
Rangaistuslaitos	1.79E-01
Urheiluseura:talkootyö	1.79E-01
Suojatyökeskus	1.79E-01
Kortti	1.79E-01
Luostarin opastuskierros	1.79E-01
Sairaala:yleishyödyllinen yhteisö	1.79E-01
Bingopeli:yleishyödyllinen yhteisö	1.79E-01
Oppilaitos	1.79E-01
Adressi	1.79E-01
Messut	7.46E-02

5.2 Singulaariarvojen jakaantuminen

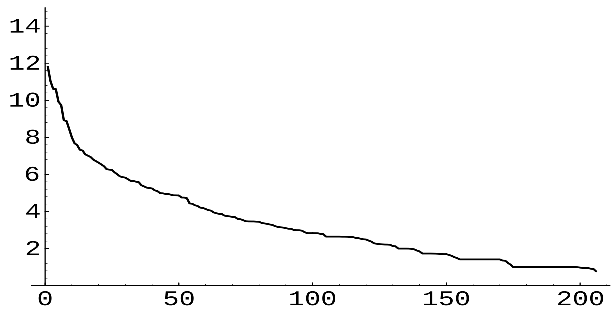
Seuraavassa kuvasarjassa näkyy singulaariarvojen jakaantuminen eri granulariteeteilla. Kuvassa 7 on koko datan singulaariarvot kansiotasolla (vain 4 dokumenttia). Koska dokumentteja on vain vähän, kaikki singulaariarvot ovat lähes samanarvoisia ja niitä vastaavat singulaarivektorit ovat hyvin sekavia, koska kaikki 9498 hakusanaa on edustettuna vain neljällä vektorilla. Kuvissa 8 - 11 on esitetty singulaariarvojen jakaantuminen Ennakkoperintä-kansiossa, kun jakoa dokumentteihin tihennetään. Mitä tiheämpi jako on, eli mitä suurempi dokumenttikokoelma on kyseessä, sitä selvemmin suurimmat singulaariarvot eroavat muista singulaariarvoista. Jaon tihentäminen vaikuttaa myös siten, että kukin dokumentti on yhä selvemmin vain yhdestä aiheesta (ks. myös Taulukko 4). Muille kansioille kuvat olivat samankaltaisia.



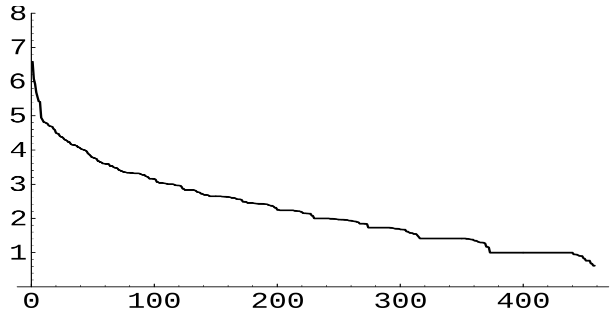
Kuva 7: Singulaariarvot koko datalle kansiotasolla, vain 4 singulaariarvoa.



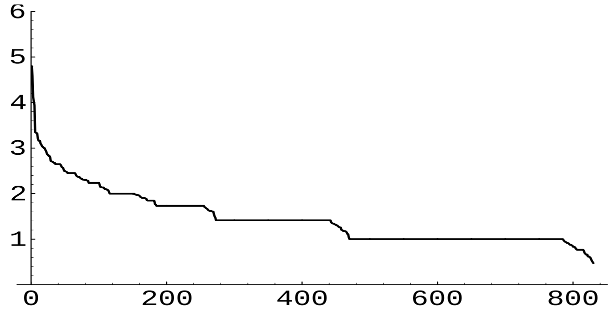
Kuva 8: Singulaariarvot: Ennakkoperintä-kansio, lukutaso, 18 singulaariarvoa.



Kuva 9: Singulaariarvot: Ennakkoperintä-kansio, otsikko2-taso, 206 singulaariarvoa.



Kuva 10: Singulaariarvot: Ennakkoperintä-kansio, otsikko3-taso, 458 singulaariarvoa.



Kuva 11: Singulaariarvot: Ennakkoperintä-kansio, marginaaliotsikkotaso, 830 singulaariarvoa.

5.3 Lähtödatan tilastollisia suureita

Taulukossa 4 näkyy, kuinka monta nollasta poikkeavaa painoa dokumenteissa oli keskimäärin.

Tilastollisia suureita			
Dokumentti-kokoelma	**Granulari-teetti**	**Keskimääräinen nollasta poikkeavien painojen lkm**	**Nollasta poikkeavien painojen lkm:n keskihajonta**
Koko data	kansio	2443.25	595.31
	luku	159.18	101.77
	otsikko2	18.58	26.56
	otsikko3	7.12	8.49
	marginaaliotsikko	2.43	3.37
AV	luku	186.12	94.67
	otsikko2	24.64	39.07
	otsikko3	8.97	12.01
	marginaaliotsikko	3.17	4.82
EP	luku	175.17	121.29
	otsikko2	15.17	20.41
	otsikko3	6.28	6.43
	marginaaliotsikko	2.40	2.90
KP	luku	110.43	84.86
	otsikko2	14.98	18.93
	otsikko3	7.34	8.12
	marginaaliotsikko	2.20	2.85
YO	luku	155.47	77.45
	otsikko2	20.76	23.43
	otsikko3	6.40	6.86
	marginaaliotsikko	1.94	2.43

Taulukko 4: Keskiarvo ja keskihajonta nollasta poikkeavien painojen lukumäärälle / dokumentti. Nollasta poikkeavat painot olivat enimmäkseen ykkösiä.

Lisäksi laskettiin kullekin dokumentille nollasta poikkeavien painojen keskiarvo ja niistä otettiin keskiarvo yli kunkin dokumenttikokoelman. Näin saadut keskiarvot olivat lähellä yhtä, välillä 1.0008 (EP, marginaaliotsikko) - 1.1863 (Koko data, kansio). Nollasta poikkeavat painot olivat siis enimmäkseen ykkösiä.

5.4 Matriisien rangit ja nollavektorit

Tällaisen hakusanoihin perustuvan esityksen vakava puute on se, että joissakin tekstiosuuksissa ei ole välttämättä yhtään hakusanaa ja silloin kyseinen dokumentti jää kokonaan vaille esitystä. Taulukossa 6 näkyy, miten suuri osa dokumenteista jäi nollavektoriksi kullakin granulariteetilla. LSI:n toimintaa ylimääräiset nollavektorit eivät haittaa, sillä hajotelmasta tulee olennaisesti sama ilman niitä ja niiden kanssa. (U ja Σ säilyvät ennallaan ja V^T:hen tulee nollasarakkeita.)

Dokumenttien lukumäärään otettiin huomioon vain ne dokumentit, joilla oli olemassa nollavektorista poikkeava esitys hakusanamallissa. Dokumenttimatriisien rangit olivat lähes täysiä näin laskettuun matriisin dimensioon nähden (Taulukko 5). Dokumenttivektorit eivät siis olleet lausuttavissa lineaarikombinaatioina toisistaan.

Dokumenttien lukumäärän ja rangin erotus					
	Kansioittain				
granulariteetti	AV	EP	KP	YO	Koko data
kansio					0
luku	0	0	0	0	0
otsikko 2	0	0	0	0	0
otsikko 3	0	1	5	0	7
marginaaliotsikko	11	18	7	16	63

Taulukko 5: Dokumenttimatriisien rangit poikkesivat alkuperäisestä dokumenttien lukumäärästä vain otsikko3-tasolla ja marginaaliotsikkotasolla. Dokumenttien lukumäärää laskettaessa otettiin huomioon vain nollavektorista poikkeavat vektorit.

Dokumentit joiden termivektori = $\overline{0}$			
Dokumentti-kokoelma	**Granulari-teetti**	**Nolladokumenttien lukumäärä**	**Nolladokumenttien osuus %**
Koko data	kansio	0	0
	luku	1	1.5
	otsikko2	72	11.3
	otsikko3	284	17.1
	marginaaliotsikko	2198	44.5
AV	luku	0	0
	otsikko2	15	10.4
	otsikko3	81	19.5
	marginaaliotsikko	591	46.2
EP	luku	0	0
	otsikko2	22	9.6
	otsikko3	75	14.0
	marginaaliotsikko	557	39.6
KP	luku	1	6.7
	otsikko2	21	17.2
	otsikko3	44	18.4
	marginaaliotsikko	359	46.2
YO	luku	0	0
	otsikko2	14	9.7
	otsikko3	84	17.7
	marginaaliotsikko	691	46.7

Taulukko 6: Hakusanaesityksen pahin ongelma on liian pieni termien lukumäärä/dokumentti. Kun jakoa hienonnettiin, suuri osa dokumenteista jäi kokonaan vaille esitystä hakusanoihin perustuvassa vektorimallissa.

6 Johtopäätökset

Latentin semanttisen indeksoinnin tuloksena saadut kantavektorit edustivat yllättävän selkeitä käsitteitä. Tutkittu dokumenttikokoelma on varsin rakenteinen, mikä on epätavallista LSI-aineistoille. Rakenne peilautuu LSI:n tuloksiin, mikä sinänsä on hyvä merkki, mutta koska rakenne tunnettiin ennalta, mitään uutta, ilmeistä tapaa luokitella aineistoa ei löytynyt. (Semantiikka ei siis ollut latenttia.)

Singulaariarvoissa ei ilmennyt äkillistä pudotusta, mikä tukisi kirjallisuudessa esitettyä käsitystä [1], jonka mukaan pieni määrä (≈ 10) kantavektoreita ei riitä esittämään dokumentteja. Toisaalta liian suuri kantavektorijoukko saattaa mallintaa dokumenttiotoksen yksityiskohtia, jotka eivät päde yleisemmin (kohinaa).

Dokumentit liittyivät dokumenttijaon tihentyessä yhä harvempiin hakusanoihin (Taulukko 4), mikä on luonnollista, sillä lyhyemmät tekstiosuudet liittyvät yleensä yhä tarkemmin vain yhteen aiheeseen.

Se, että jokaisella dokumentilla on esityksensä mallissa, on perusedellytys minkä tahansa dokumenttimallin käyttökelpoisuudelle. Dokumenttijaon tihentyessä syntyi kuitenkin dokumentteja, joilla ei ollut esitystä hakusanamallissa, koska dokumentissa ei esiintynyt yhtään hakusanaa. Tämän mallin soveltaminen käytäntöön muulla granulariteetilla kuin lukutasolla ei siis tulisi kyseeseen (Taulukko 6).

Jatkossa voisi kokeilla vastaavaa mallia oikeilla tekstissä esiintyvillä termeillä hakusanojen sijasta. Tällöin dokumenteilla olisi varmasti nollasta poikkeava esityksensä ja hakusanoja voisi lisäksi hyödyntää antamalla niille isommat painot kuin muille termeille.

Viitteet

[1] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A. (1990). Indexing by latent semantic analysis. Journal of the Society for Information Science, 41(6), 391-407. http://superbook.bellcore.com/ std/LSI.papers.html

[2] Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285. http://superbook.bellcore.com/ std/LSI.papers.html

[3] Berry, M. W., Dumais, S. T. (1995). Using linear algebra for intelligent information retrieval.
SIAM Review, 37(4), 1995, 573-595.
http://superbook.bellcore.com/ std/LSI.papers.html

[4] Salton, G., McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill

[5] Sheth, Beerud (1994). A Learning Approach to Personalized Information Filtering, M.I.T.
http://agents.www.media.mit.edu/groups/agents/papers/newt-thesis/main.html

[6] Golub, van Loan (1983). Matrix Computations.
The Johns Hopkins University Press.

[7] WSOY Yritysjulkaisut: Yritystieto-sarja
http://www.wsoy.fi/yjulk/kansiot

A Liite : Määritelmiä

Sisätuloavaruus

Sisätuloavaruus on vektoriavaruus S, jossa on määritelty sisätulo $(f, g) \in \mathbf{C}($ tai $\mathbf{R})$, jolle pätee

1. $(f, g) = \overline{(g, f)}$, reaalisessa tapauksessa symmetrisyys
2. $(\alpha \cdot f + \beta \cdot g, h) = (\alpha \cdot f, h) + (\beta \cdot g, h)$
3. $\forall f \ (f, f) \geq 0$ ja $(f, f) = 0 \iff f = \bar{0}$.

Ortogonaalisuus

Sisätuloavaruudessa voidaan määritellä ortogonaalisuus :
f ja g ovat ortogonaalisia eli $f \perp g \iff (f, g) = 0$.
(Nollavektori $\bar{0}$ on ortogonaalinen kaikkien vektoreiden kanssa.)

$\mathbf{R}^n$ tavanomaisella vektoriyhteenlaskulla ja reaaliluvulla kertomisella varustettuna on vektoriavaruus ja kun sisätuloksi valitaan $(f, g) = f^T \cdot g \in \mathbf{R}$, saadaan sisätuloavaruus.

Ortonormaalit vektorit

f ja g ovat ortonormaalit $\iff f \perp g \quad \text{ja} \quad (f, f) = (g, g) = 1.$

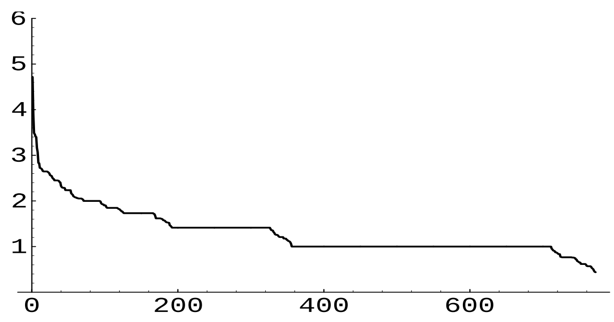
Kanta

Kanta on lineaarisesti riippumaton joukko vektoreita, jotka virittävät avaruuden eli kaikki $x \in S$ voidaan lausua kantavektoreiden lineaarikombinaatioina $(a_k \in \mathbf{R})$
$x = a_1 \cdot e_1 + a_2 \cdot e_2 + \cdots + a_n \cdot e_n$.

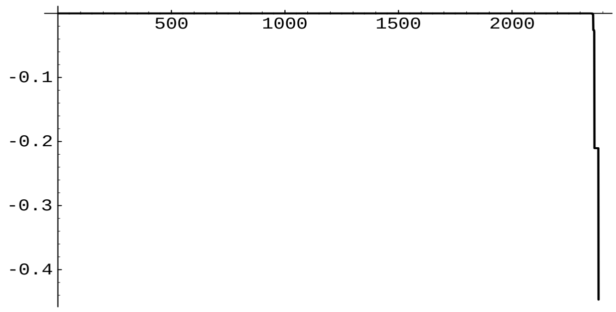
B Liite : Esimerkkejä kantavektoreista

B.1 Yhtiöoikeus-kansio: marginaaliotsikkotaso

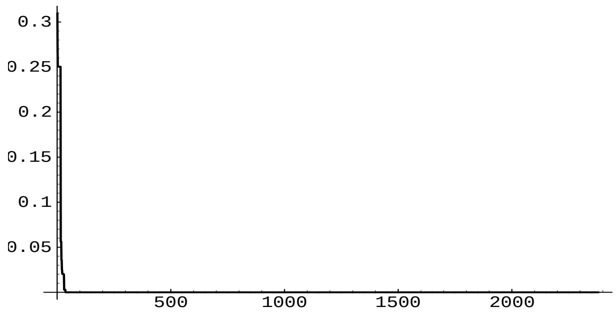
Singulaariarvot ovat kuvassa 12, 3 ensimmäisen kantavektorin painot ovat kuvissa 13 - 15 ja itseisarvoltaan suurimpia painoja vastaavien termien listat sivuilla 27 - 29.



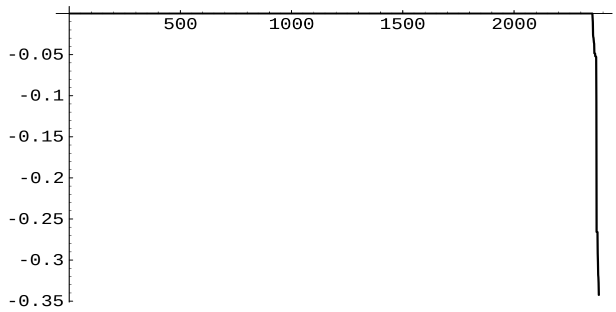
Kuva 12: Singulaariarvot: Yhtiöoikeus-kansio, marginaaliotsikkotaso, 772 singulaariarvoa.



Kuva 13: 1. kantavektori, Yhtiöoikeus-kansio, marginaaliotsikkotaso, painot suuruusjärjestyksessä.



Kuva 14: 2. kantavektori, Yhtiöoikeus-kansio, marginaaliotsikkotaso, painot suuruusjärjestyksessä.



Kuva 15: 3. kantavektori, Yhtiöoikeus-kansio, marginaaliotsikkotaso, painot suuruusjärjestyksessä.

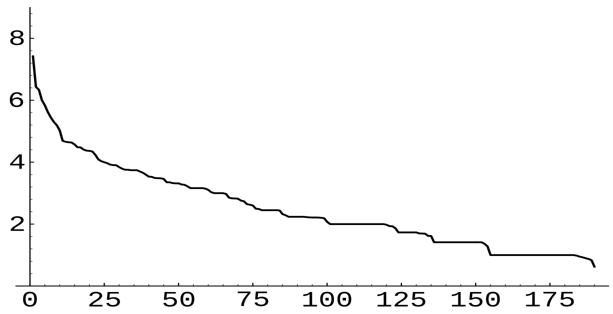
1. kantavektori, itseisarvoltaan suurimmat painot Yhtiöoikeus-kansio, marginaaliotsikkotaso	
hakusana	paino
Yhdistys:taloudellinen	-4.46E-01
Tiekunta	-2.10E-01
Osakaskunta	-2.10E-01
Yhteismetsä	-2.10E-01
Kauppakamari	-2.10E-01
Kalastuskunta	-2.10E-01
Viemäriyhtiö	-2.10E-01
Paliskunta	-2.10E-01
Henkilöstörahasto	-2.10E-01
Ojitusyhtiö	-2.10E-01
Avustuskassa	-2.10E-01
Työttömyyskassa	-2.10E-01
Järjestely-yhtiö	-2.10E-01
Vakuutusyhdistys	-2.10E-01
Hypoteekkiyhdistys	-2.10E-01
Riistanhoitoyhdistys	-2.10E-01
Säännöstely-yhtiö	-2.10E-01
Uittoyhdistys	-2.10E-01
Kalastusvakuutusyhdistys	-2.10E-01
Taloudellinen yhdistys	-2.88E-02
Julkisoikeudellinen yhdistys	-2.60E-02
Yhdistys:julkisoikeudellinen	-2.60E-02
Yhdistys:aatteellinen	-2.60E-02
Aatteellinen yhdistys	-2.60E-02

2. kantavektori, itseisarvoltaan suurimmat painot Yhtiöoikeus-kansio, marginaaliotsikkotaso	
hakusana	paino
Osakeyhtiö:perusilmoitus	3.09E-01
Kommandiittiyhtiö:perusilmoitus	2.69E-01
Aatteellinen yhdistys:perusilmoitus	2.50E-01
Osuuskunta:perusilmoitus	2.50E-01
Säästöpankki:perusilmoitus	2.50E-01
Keskinäinen vakuutusyhtiö:perusilmoitus	2.50E-01
Sivuliike:perusilmoitus	2.50E-01
Perusilmoitus:osakeyhtiö	2.50E-01
Valtion liikelaitos	2.50E-01
Perusilmoitus:kommandiittiyhtiö	2.50E-01
Perusilmoitus:avoin yhtiö	2.50E-01
Säätiö:perusilmoitus	2.50E-01
Asumisoikeusyhdistys:perusilmoitus	2.50E-01
Hypoteekkiyhdistys:perusilmoitus	2.50E-01
Vakuutusyhdistys:perusilmoitus	2.50E-01
Rinnakkaistoiminimi	5.62E-02
Toiminimi:aputoiminimi	5.62E-02
Aputoiminimi	5.62E-02
Toiminimi:rinnakkaistoiminimi	3.59E-02
Tilintarkastaja:todistus	3.59E-02
Osakeyhtiö:muutosilmoitus	2.63E-02
Muutosilmoitus:osakeyhtiö	2.34E-02
Jäsen:osuuskunta	2.02E-02
Ilmoitus kaupparekisteriin:osuuskunta	2.02E-02
Toissijainen tunnus:osuuskunta	2.02E-02
Toiminimi:osuuskunta	2.02E-02
Osuuskunta:toiminimi	2.02E-02
Vastuu:osuuskunta	2.02E-02
Perustamiskirja:osuuskunta	2.02E-02
Ilmoitus kaupparekisteriin:kommandiittiyhtiö	1.88E-02
Hypoteekkiyhdistys:muutosilmoitus	2.82E-03
Kommandiittiyhtiö:muutosilmoitus	2.82E-03
Vakuutusyhdistys:muutosilmoitus	2.82E-03
Vastuu:muutosilmoitus	2.82E-03
Säästöpankki:muutosilmoitus	2.82E-03
Osuuskunta:muutosilmoitus	2.82E-03

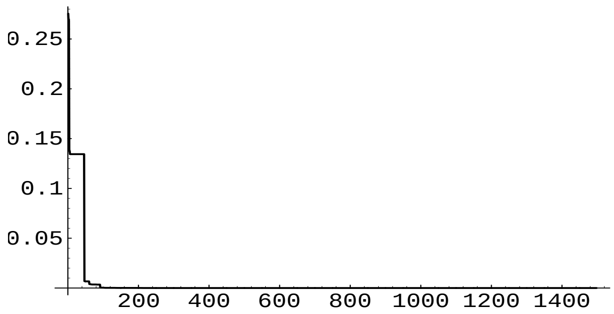
3. kantavektori, itseisarvoltaan suurimmat painot Yhtiöoikeus-kansio, marginaaliotsikkotaso	
hakusana	paino
Äänetön yhtiömies	-3.42E-01
Yhtiömies:äänetön	-3.27E-01
Vastuunalainen yhtiömies	-3.21E-01
Vastuu:avoin yhtiö	-3.17E-01
Vastuu:kommandiittiyhtiö	-3.02E-01
Kommandiittiyhtiö	-2.92E-01
Kommanditääri	-2.65E-01
Yhtiömies:vastuunalainen	-2.65E-01
Komplementääri	-2.65E-01
Avoin yhtiö:lainsäädäntö	-2.65E-01
Kommandiittiyhtiö:tunnusmerkit	-2.65E-01
Vastuu:yhtiömies	-1.07E-01
Yhtiömies:kommandiittiyhtiö	-5.33E-02
Avoin yhtiö:velkavastuu	-5.21E-02
Vastuu:keskinäinen	-5.21E-02
Velkavastuu:avoin yhtiö	-5.21E-02
Keskinäinen vastuu, yhtiösuhde	-5.21E-02
Kommandiittiyhtiö:äänetön yhtiömies	-4.85E-02
Vastuunalainen yhtiömies:kommandiittiyhtiö	-4.85E-02
Äänetön yhtiömies:kommandiittiyhtiö	-4.85E-02
Kommandiittiyhtiö:vastuunalainen yhtiömies	-4.85E-02
Panosvelka, äänetön yhtiömies	-3.67E-02
Kommandiittiosakeyhtiö	-3.54E-02
Kommandiittiyhtiö:yhtiöpanos	-3.25E-02
Kommandiittiyhtiö:kielto-oikeus	-2.93E-02
Yhtiöpanos:kommandiittiyhtiö	-2.87E-02
Vastuu:henkilökohtainen	-2.61E-02
Yhtiömies:vastuu	-1.05E-02
Perustaminen:kommandiittiyhtiö	-5.24E-03

B.2 Kirjanpito-kansio: otsikko3-taso

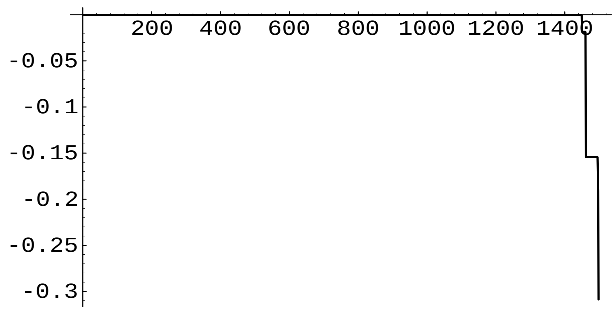
Singulaariarvot ovat kuvassa 16, 5 ensimmäisen kantavektorin painot ovat kuvissa 17 - 21 ja itseisarvoltaan suurimpia painoja vastaavien termien listat sivuilla 33 - 37.



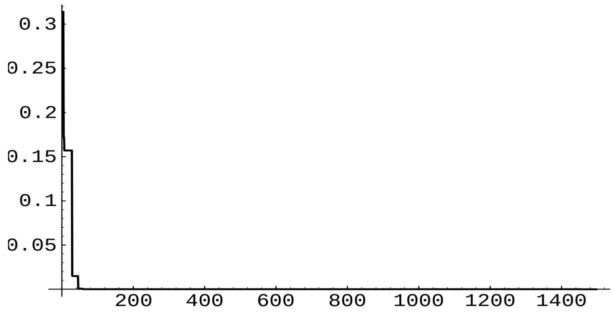
Kuva 16: Singulaariarvot: Kirjanpito-kansio, otsikko3-taso, 190 singulaariarvoa.



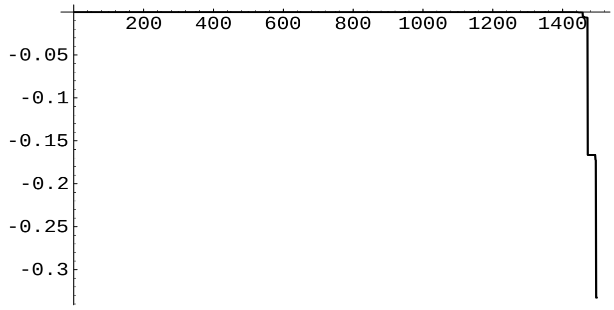
Kuva 17: 1. kantavektori, Kirjanpito-kansio, otsikko3-taso, painot suuruusjärjestyksessä.



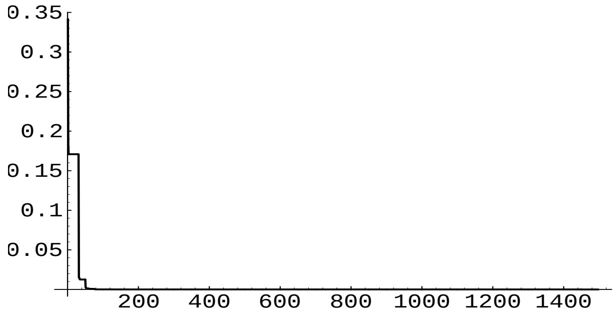
Kuva 18: 2. kantavektori, Kirjanpito-kansio, otsikko3-taso, painot suuruusjärjestyksessä.



Kuva 19: 3. kantavektori, Kirjanpito-kansio, otsikko3-taso, painot suuruusjärjestyksessä.



Kuva 20: 4. kantavektori, Kirjanpito-kansio, otsikko3-taso, painot suuruusjärjestyksessä.



Kuva 21: 5. kantavektori, Kirjanpito-kansio, otsikko3-taso, painot suuruusjärjestyksessä.

1. kantavektori, itseisarvoltaan suurimmat painot Kirjanpito-kansio, otsikko3-taso			
hakusana	paino	hakusana	paino
Poistosuunnitelma	2.75E-01	Poistosuunnitelma:korko	1.34E-01
Poistoero:yhdistäminen	2.68E-01	Poistoperiodi	1.34E-01
Suunnitelman mukaiset poistot:		Tasapoisto	1.34E-01
aloittaminen	2.68E-01	Tilikauden muutos:	
Pakollinen varaus	1.37E-01	pieni yritys	1.34E-01
Suunnitelman mukaiset poistot	1.37E-01	Perusparannusmeno	1.34E-01
Aineettomat menot	1.34E-01	Pienhankinta	1.34E-01
Alipoisto	1.34E-01	Verosidonnaisuus	1.34E-01
Varaukset:pakolliset	1.34E-01	Suunnitelman mukaiset	
Tulonodotus	1.34E-01	poistot:	
Varaus- ja poistojärjestelmän		verosidonnaisuus	1.34E-01
uudistamistyöryhmä	1.34E-01	Taannehtiva selvitys	1.34E-01
Romutus	1.34E-01	Käyttöomaisuuskirjanpito	1.34E-01
EVL-poisto	1.34E-01	Korko:poistosuunnitelma	1.34E-01
Pieni kirjanpitovelvollinen:		Koekäyttö	1.34E-01
suunnitelman mukainen poisto	1.34E-01	Jäännösarvo	1.34E-01
Pieni yritys	1.34E-01	Leasingsopimus,	
Aravavuokratalo, poistot	1.34E-01	suunnitelman	
Liikearvo	1.34E-01	mukaiset poistot	1.34E-01
Poistot:verosidonnaisuus	1.34E-01	Myyntivoitto:	
Poistojen aloittaminen	1.34E-01	suunnitelman	
Vähäinen hyödyke	1.34E-01	mukainen poisto	1.34E-01
Ylimääräinen poisto	1.34E-01	Taloudellinen pitoaika	1.34E-01
Suunnitelman mukaiset poistot:		Tilikauden muutos:	
käyttöomaisuuden luovutus	1.34E-01	suunnitelman mukainen	
Suunnitelman mukaiset poistot:		poisto	1.34E-01
leasingsopimus	1.34E-01	Myyntivoitto:	
Suunnitelman mukaiset poistot:		pieni yritys	1.34E-01
kertynyt poistoero	1.34E-01	Myynti	1.34E-01
Suunnitelman mukaiset poistot:			
aravavuokratalo	1.34E-01		
Kertynyt poistoero	1.34E-01		
Poistot:lyhennyssidonnaiset	1.34E-01		
Suunnitelman mukaiset poistot:			
alipoisto	1.34E-01		

2. kantavektori, itseisarvoltaan suurimmat painot Kirjanpito-kansio, otsikko3-taso			
hakusana	paino	hakusana	paino
Osuuskunta:		Osakaslaina:	
osuuspääoma	-3.08E-01	osakeyhtiössä	-1.54E-01
Osakepääoma	-1.91E-01	Oma pääoma:	
Sijoitusosuuspääoma	-1.72E-01	osakeyhtiö	-1.54E-01
Osakepääoma:		Apporttiehto	-1.54E-01
alentaminen	-1.54E-01	Osakeanti	-1.54E-01
Osuuskunta:		Oma pääoma:	
jäsenen eroaminen	-1.54E-01	osuuskunta	-1.54E-01
Osakeyhtiö:		Osakepääoma:	
johdon laina	-1.54E-01	korottaminen	-1.54E-01
Osuuspääomatili	-1.54E-01	Vararahasto:	
Osakeyhtiö:		emissiovoiton	
osakaslaina	-1.54E-01	kirjausajankohta	-1.54E-01
Emissiovoitto:		Palautettava osuuspääoma	-1.54E-01
kirjausajankohta,		Osuuspääoma:	
vararahasto	-1.54E-01	korottaminen	-1.54E-01
Osakkeet:merkintä	-1.54E-01	Maksullinen anti	-1.54E-01
Osuuspääoma	-1.54E-01	Osakeanti:	
Toimintakertomus:		apporttiehto	-1.54E-01
suunnattu anti	-1.54E-01	Osuuspääoma:	
Sidottu oma pääoma:		palauttaminen	-1.54E-01
osuuskunnasta eroaminen	-1.54E-01	Osuuskunta:	
Sidottu oma pääoma:		sidottu oma pääoma	-1.54E-01
osuuskunta	-1.54E-01	Ilmaisemissio	-1.54E-01
Osakeanti:ilmaisanti	-1.54E-01	Rahastoanti	-1.54E-01
Sidottu oma pääoma:		Johdon lainat	
osuuspääoma	-1.54E-01	osakeyhtiössä	-1.54E-01
Palautettavien		Apporttiomaisuus	-1.54E-01
osuuspääomien tili	-1.54E-01		
Tarjousesite,			
arvopaperimarkkinalaki	-1.54E-01		
Suunnattu anti,			
toimintakertomus	-1.54E-01		
Osakeanti:maksullinen anti	-1.54E-01		
Osakeanti:suunnattu	-1.54E-01		

3. kantavektori, itseisarvoltaan suurimmat painot Kirjanpito-kansio, otsikko3-taso	
hakusana	paino
Ammatinharjoittaja:tilinpäätös	3.14E-01
Tilinpäätös:ammatinharjoittaja	3.14E-01
Tuloslaskelma:ammatinharjoittaja	3.14E-01
Ammatinharjoittaja:kahdenkertainen kirjanpito	3.14E-01
Kirjanpito:säilyttäminen	1.71E-01
Tilinpäätös:valmistuminen	1.71E-01
Ammatinharjoittaja:tuloslaskelma	1.57E-01
Henkilöstörahasto	1.57E-01
Ammatinharjoittaja:liiketapahtuma	1.57E-01
Liiketapahtuma:ammatinharjoittaja	1.57E-01
Ammatinharjoittaja:yhdenkertainen kirjanpito	1.57E-01
Veroilmoitus, ammatinharjoittaja	1.57E-01
Tilikartta:ammatinharjoittaja	1.57E-01
Tilikausi:ammatinharjoittaja	1.57E-01
Tilintarkastus:ammatinharjoittaja	1.57E-01
Ammatinharjoittaja:oma käyttö	1.57E-01
Rahoitustapahtuma:ammatinharjoittaja	1.57E-01
Yhdenkertainen kirjanpito	1.57E-01
Ammatinharjoittaja:kirjanpidon säilyttäminen	1.57E-01
Ammatinharjoittaja:kirjanpitovelvollisuus	1.57E-01
Oma käyttö, ammatinharjoittaja	1.57E-01
Kahdenkertainen kirjanpito:ammatinharjoittaja	1.57E-01
Varallisuus, ammatinharjoittaja	1.57E-01
Ammatinharjoittaja:tilikausi	1.57E-01
Ammatinharjoittaja:veroilmoitus	1.57E-01
Ammatinharjoittaja:varallisuus	1.57E-01
Ammatinharjoittaja:jaksotus	1.57E-01
Ammatinharjoittaja:tilikartta	1.57E-01

4. kantavektori, itseisarvoltaan suurimmat painot Kirjanpito-kansio, otsikko3-taso	
hakusana	paino
Osavuosikatsaus	-3.32E-01
Vuosikatsaus	-3.32E-01
Yleinen tiedonantovelvollisuus	-3.32E-01
Tilinpäätös	-1.72E-01
Tilinpäätös:julkistaminen	-1.71E-01
Markkinayhtiö:osavuosikatsaus	-1.66E-01
Osavuosikatsaus:julkistaminen	-1.66E-01
Osavuosikatsaus:IAS-periaatteet	-1.66E-01
Valtiovarainministeriön päätös:tilinpäätöstiedote	-1.66E-01
Markkinayhtiö:tilinpäätöksen julkistaminen	-1.66E-01
Markkinayhtiö:tilinpäätös	-1.66E-01
Valtiovarainministeriön päätös: osakekohtaiset tunnusluvut	-1.66E-01
Rahoitustarkastus:määräykset	-1.66E-01
Markkinayhtiö:vuosikatsaus	-1.66E-01
Valtiovarainministeriön päätös: tunnusluvut, taloudellinen kehitys	-1.66E-01
Markkinayhtiö:tilinpäätöstiedote	-1.66E-01
Tunnusluvut:osakekohtaiset	-1.66E-01
Tunnusluvut:taloudellinen kehitys, VMp	-1.66E-01
Julkistaminen:osavuosikatsaus	-1.66E-01
Markkinayhtiö:joukkovelkakirjan liikkeeseenlaskija	-1.66E-01
Osakekohtainen tunnusluku, valtiovarainministeriön päätös	-1.66E-01
Markkinayhtiö:yleinen tiedonantovelvollisuus	-1.66E-01
IAS-periaatteet:osavuosikatsaus	-1.66E-01
Tilinpäätös:valtiovarainministeriön päätös	-1.66E-01
Valtiovarainministeriön päätös:tilinpäätös	-1.66E-01
Joukkovelkakirja, liikkeeseenlaskija	-1.66E-01
Tilinpäätöstiedote:valtiovarainministeriön päätös	-1.66E-01

5. kantavektori, itseisarvoltaan suurimmat painot Kirjanpito-kansio, otsikko3-taso	
hakusana	paino
Rahasto:asunto-osakeyhtiö	3.41E-01
Pääomavastike	1.83E-01
Asunto-osakeyhtiö:rahastot	1.70E-01
Asunto-osakeyhtiö:vastikerahoituslaskelma	1.70E-01
Vuokraus:kiinteistö	1.70E-01
Toimintakertomus:kiinteistö	1.70E-01
Poistot:kiinteistön hallinta	1.70E-01
Vararahasto:asunto-osakeyhtiö	1.70E-01
Vuokrankantokirja	1.70E-01
Palkat:kiinteistö	1.70E-01
Asunto-osakeyhtiö:vararahasto	1.70E-01
Kiinteistön hallinta:poistot	1.70E-01
Hoitovastike	1.70E-01
Kiinteistön hallinta:liitetiedot	1.70E-01
Kiinteistön hallinta:käyttökorvaukset	1.70E-01
Kiinteistön hallinta:erilliskirjanpito	1.70E-01
Suunnitelman mukaiset poistot:kiinteistön hallinta	1.70E-01
Kiinteistön hallinta:vuokrankantokirja	1.70E-01
Kiinteistön hallinta:vuokraus	1.70E-01
Kiinteistöosakeyhtiö:rakennusaikainen kirjanpito	1.70E-01
Rakennusrahasto	1.70E-01
Kiinteistön hallinta:tuloslaskelma	1.70E-01
Kiinteistön hallinta:toimintakertomus	1.70E-01
Kiinteistön hallinta:rakennusaikainen kirjanpito	1.70E-01
Kiinteistön hallinta:rahastot	1.70E-01
Käyttökorvaukset kiinteistössä	1.70E-01
Rakennusaikainen:kirjanpito	1.70E-01
Rakennus:kiinteistön hallinta	1.70E-01
Vastikerahoituslaskelma:asunto-osakeyhtiö	1.70E-01
Asunto-osuuskunta	1.70E-01
Asunto-osakeyhtiö:vuokrankantokirja	1.70E-01