Luokitteluohje täysistuntokeskustelujen LDA-mallien laatuanalyysille
Versio 1.2

Kimmo Makkonen
5.9.2017

Tämän luokituksen tarkoituksena on yksittäisten LDA-mallien laadun arviointi, jota voidaan hyödyntää vertailtaessa eri LDA-malleja ja arvioitaessa mielekästä mallinnettavien puheenaiheiden määrää sekä muita mallinnusparametreja. LDA-mallit luokitellaan kahdella tavalla:

1) Top 50 -sanat
2) pyLDAvis-visualisoinnin erottamiskyvyn ja relevanssin avulla

Kolmas laadunarviointitapa olisi lukea puheita, jotka LDA-malli osoittaa kuuluvaksi puheenaiheeseen suurimmalla todennäköisyydellä, ja arvioida niiden kuuluvuutta yhtenäiseen teemaan. Tässä yhteydessä sitä ei tehdä ajankäytöllisistä syistä, mutta alkuperäiset puheet pidetään ihmislukijoiden saatavilla, jos jossakin kohdassa niistä halutaan tukea päättelytyölle.

Luokittelun tekijöiden on hyvä perehtyä LDA-mallin toiminnan yleisiin periaatteisiin (Blei 2012, Blei ym. 2003) ennen analyysia, mutta mallien laskennan ymmärtäminen ei ole ihmislukijan analyysille välttämätöntä.

LUOKAT

Ihmislukijan tehtävänä on määritellä jokainen puheenaihe johonkin seuraavista luokista:

Laadukas: puheenaihe vastaa yhtä ihmislukijalle mielekästä teemaa.
Läheiset: erilliset aiheet, jotka liittyvät toisiinsa jollain tapaa. Esimerkiksi toimeentulotuki ja lapsilisä ovat yhteiskunnan tukia, mutta kuitenkin erilaisia tukia. Luokan käyttö riippuu myös puheenaiheiden määrästä: jos kokonaisaihemäärä on pieni, voidaan tulkita, että kaikenlainen puhe yhteiskunnan tukimuodoista sopii yhteen aiheeseen.
Erilliset: puheenaihe vastaa kahta tai useampaa ihmislukijalle mielekästä teemaa.
Tunkeutunut: puheenaihe vastaa pääpiirteissään yhtä ihmislukijalle mielekästä teemaa, mutta todennäköisimpien sanojen joukossa esiintyy useita sanoja, jotka eivät sovi asiayhteyteen.
Satunnainen: puheenaiheen sanasto ei muodosta ihmislukijalle mielekästä kokonaisuutta.
Ketjuuntunut: puheenaiheessa esiintyy kaksi erillistä teemaa, jotka liittyvät toisiinsa jonkin yhdistävän sanajoukon kautta. Esimerkiksi puheet terrorismista ja lasten kotihoidosta saattavat luokittua samaan aiheeseen, koska molemmissa puhutaan tukemisesta. Tällaisia puheenaiheita esiintyy yleensä varsin niukasti.
Vinoutunut: puheenaiheessa yksi tai muutama sana saavat erittäin suuren todennäköisyyden, ja muilta osin todennäköisyydet ovat pieniä. Esimerkiksi sanat ’kannattaa’ ja ’ehdotus’ esiintyvät todennäköisyydellä > 0,1, ja muuten sanasto liittyy vain löyhästi johonkin

teemaan tai on satunnaista. Jos sanasto liittyy johdonmukaisesti samaan teemaan, kyseessä on laadukas puheenaihe.

Yleissanoja: puheenaihe sisältää kielessä yleisesti esiintyviä sanoja, esimerkiksi ’paljon’, ’mennä’, ’silloin’, ’puhua’, ’ihminen’, ’tietää’ jne., eikä joukossa ole merkittävässä määrin erityisalasanastoa.

Näiden lisäksi luokituksen yhteydessä kerätään muita luokittelijan tekemiä huomioita puheenaiheesta – esimerkiksi toistuuko sama teema mallin useissa puheenaiheissa, minkälaisia yksittäisiä tunkeutujasanoja esiintyy, vaikka koko aihetta ei luokiteltaisi tunkeutuneeksi, sekä mitä tahansa luokittelijaa kummastuttavia havaintoja.

Edellisten määritelmien lisäksi luokitusta laadittaessa kannattaa kiinnittää huomiota myös siihen, kuinka yleinen puheenaihe on kyseessä. Mikäli yhteen puheenaiheeseen malliintuu erityisalasanastoa, kyseessä saattaa olla erittäin käyttökelpoinen puheenaihe, vaikka aineistossamme ei olisi paljoakaan puheita, joissa aihe on keskeisimpänä.

Tämä luokitus on kehitetty Chuangin ym. (2013) ja Mimnon ym. (2011) luokituksen perusteella.

PUHEENAIHEIDEN NIMEÄMINEN

Tulosten raportoimiseksi on yleensä tarpeen nimetä laskennalliset puheenaiheet ihmislukijoille helposti ymmärrettävällä tavalla. Esimerkiksi jos aiheessa esiintyvät sanat ’lapsi’, ’äiti’, ’isä’, ’perhe’, ’vanhemmuus’ yms., on luontevaa nimetä puheenaihe perhepolitiikaksi, vaikka sana ’perhepolitiikka’ ei olisikaan merkitsevimpien sanojen joukossa.

Periaatteena on nimetä puheenaihe käyttäen teemaa yhdistävää abstraktia kokoavaa käsitettä, mikäli sellainen on suomen kielessä olemassa, tai ellei sopivaa käsitettä ole, kuvailemalla puheenaihetta muutamilla aiheen merkitsevimmillä sanoilla.

Jos aihe on ihmislukijalle satunnainen, on kuvauksena käytetty sanaa ”sekalaista”. Erittäin usein kaikenlaisissa teemoissa esiintyviä sanoja sisältävät puheenaiheet on luokiteltu nimikkeellä ”tavallisia sanoja”.

MALLITULOSTEIDEN LUKUOHJE

Kuvassa on tuloste yhdestä laadukkaasta LDA-puheenaiheesta, johon on malliintunut koulutuspolitiikkaa käsittelevää sanastoa.

Top 50 -sanat

Ylimmältä riviltä voidaan lukea, että puheenaiheen sanastoa esiintyy 22,57 prosentissa analysoiduista puheista, ja merkitsevin aihe se on 12,12 prosentissa puheenvuoroista. Sanakohtaiset todennäköisyydet kyseisen puheenaiheen todennäköisimmille sanoille esiintyvät kunkin sanan jälkeen.

topic #6
documents 40928 (22.57)%, top documents 4959 (12.12%)

koulutus	0.0385	perus#opetus	0.0064	mahdollisuus	0.0053	valmistua	0.0035
nuori	0.0371	rahoitus	0.0061	aste	0.0051	järjestää	0.0033
yli#opisto	0.0279	osaaminen	0.0060	laatu	0.0051	esimerkiksi	0.0033
opiskelija	0.0224	syrjäytyä	0.0059	lukio	0.0050	vahvistaa	0.0033
koulu	0.0176	sivistys#valio#kunta	0.0059	tutkinto	0.0050	tukea	0.0033
opetus	0.0131	opiskelu	0.0059	tavoite	0.0049	huomio	0.0032
ammatillinen	0.0117	lisätä	0.0058	opiskella	0.0048	työelämä	0.0032
tärkeä	0.0114	opinto	0.0058	opinto#tuki	0.0048	erityisesti	0.0032
kehittää	0.0108	opetus#ministeriö	0.0057	suorittaa	0.0047	yhteis#työ	0.0032
opettaja	0.0105	perus#koulu	0.0056	haluta	0.0043	jatko	0.0031
ammatti#korkeakoulu	0.0091	oppia	0.0054	tarve	0.0042	riittävä	0.0031
tutkimus	0.0083	tulevaisuus	0.0054	oppi#laitos	0.0038		
oppilas	0.0083	tarvita	0.0053	esi#opetus	0.0037		

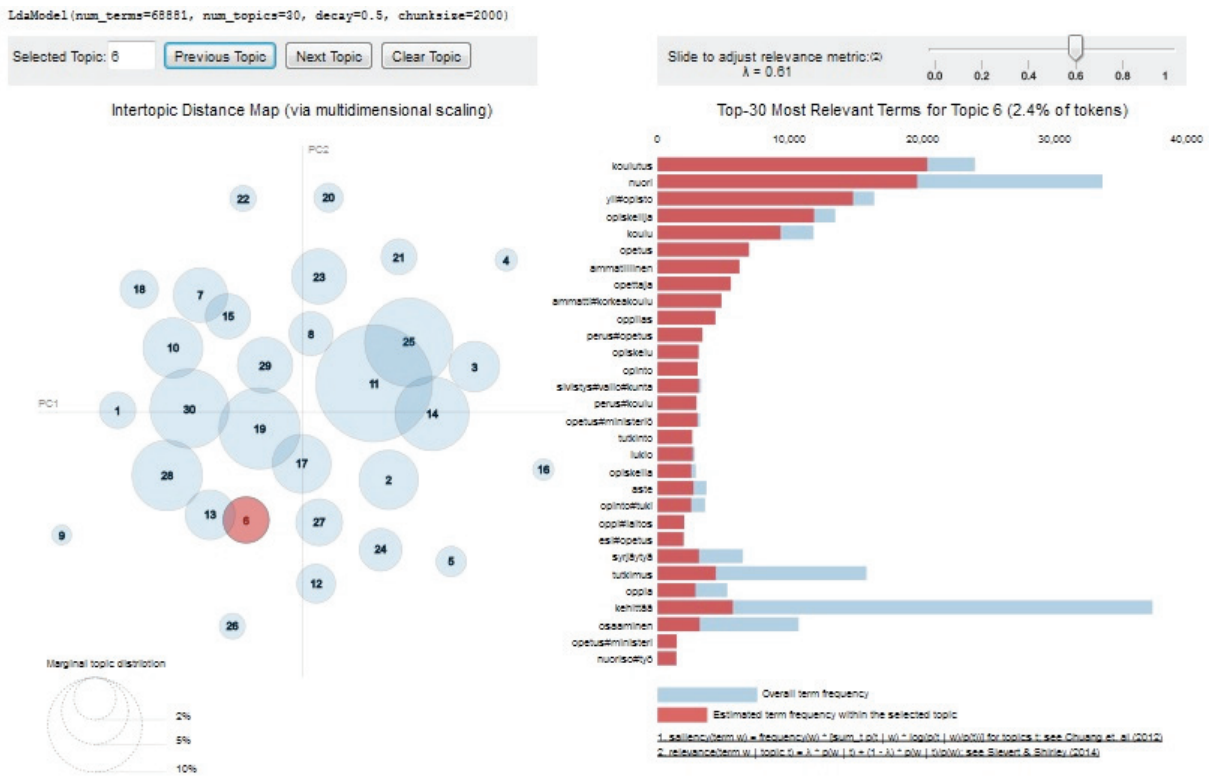
pyLDAvis-visualisointi

LDA-vis-visualisoinnissa ja relevanssin arviossa Sievert ja Shirley (2014) ovat määritelleet sanan *w relevanssin* käyttämällä Taddyn (2011) kehittämää suuretta *noste* (lift). He merkitsevät φ_{kw} sanan *w* todennäköisyys puheenaiheessa *k*, ja p_w sanan todennäköisyys korpuksessa. Noste on termin todennäköisyys puheenaiheen sisällä jaettuna todennäköisyydellä koko korpuksessa. Sievert ja Shirley pitävät nosteen etuna sitä, että koko korpuksessa tavallisten sanojen esiintyminen sanalistojen kärjessä vähenee, mutta toisaalta erittäin harvinaiset sanat saattavat nostetta käytettäessä saada liiankin suuren painoarvon, mikä saattaa vaikeuttaa puheenaiheen tulkintaa. Sievertin ja Shirleyn ratkaisuna on muodostaa relevanssi-yhdistelmäsuure, jossa parametri λ (välillä [0, 1]) määrittää, missä suhteessa kumpaakin mittaria käytetään.

$$relevanssi(w, k \mid \lambda) = \lambda \, log(\phi_{kw}) + (1 - \lambda) \, log(\frac{\phi_{kw}}{p_w})$$

Eduskuntakeskusteluista laskettujen LDA-analyysien vertailussa ja puheenaiheiden sisällön tunnistamisessa pyLDAvis-ohjelman avulla käytännössä on osoittautunut toimivimmaksi tarkastella puheenaihetta ensin arvolla λ = 0,6 ja tarvittaessa lisäksi ääriarvoilla 0 ja 1. Muitakin arvoja voi käyttää, mutta se ei ole yleensä mielekästä työajan pitämiseksi kohtuullisena.

Kaksiulotteisen visualisoinnin tuottamasta kuvasta voi arvioida aiheiden läheisyyttä sekä katsoa, kuinka suuri osa puheista kuuluu aiheisiin. Tässä yhteydessä kyseistä informaatiota ei ole kuitenkaan tarkoitus analysoida järjestelmällisesti.



Esimerkki pyLDAvis-tulosteesta. Korostettuna on puheenaihe nro. 6. Tulosteesta nähdään ylhäältä vasemmalta lukien, että mallissa on 68 881 sanaa ja 30 puheenaihetta. Lisäksi nähdään LDA-laskennan parametrit decay ja chunksize, joista ihmislukijan ei tarvitse tässä yhteydessä välittää.

Oikeanpuoleisessa kuvaajassa punaisella on merkitty kunkin sanan merkitys kyseiselle puheenaiheelle ja sinisellä sanan yleisyys koko korpuksen tasolla. Valitsemalla lambda välillä [0, 1] voidaan tarkastella sanan merkitystä koko puheenaiheessa, koko korpuksessa tai painottaen näiden välillä.

Yksittäisen sanan merkitystä ja esiintymistä muissa puheenaiheissa voidaan tarkastella viemällä kursori kyseisen sanan ylle. Esimerkiksi jos jokin poliittisesti tärkeä sana on malliintunut puheenaiheeseen, johon se ei kuuluisi, tilanne voi olla varsin ongelmaton, mikäli sana esiintyy sopivissakin puheenaiheissa, mutta hankala, mikäli sana esiintyy vain yhdessä puheenaiheessa.

pyLDAvis-tuloste sisältää paljon muutakin ohjelman käyttöön liittyvää informaatiota, jonka voi tässä yhteydessä ohittaa. Vaikka kyseessä on offline-tiedostolta vaikuttava paketti, tiedoston käyttö edellyttää internetyhteyttä. pyLDAvis toimii ainakin Mozilla Firefox -selaimessa.

YHTEYSTIETOJA JA LISÄKYSYMYKSIÄ

Kaikissa ongelmatilanteissa ja lisätietoja tarvittaessa ota yhteyttä: Kimmo Makkonen, puh. 040 516 8090, ja sähköpostitse kikama@utu.fi.

Mikäli jokin puheenaihe ei sovi mihinkään luokista, sen voi jättää toistaiseksi luokittamatta, ja miettiä, olisiko luokitusta täydennettävä jollain tapaa.

LÄHTEET

Blei, David M., Andrew Y. Ng, Michael I. Jordan ja John Lafferty. 2003. Latent Dirichlet Allocation. *Journal Of Machine Learning Research* 3:4/5, 993-1022.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM*. 55:4, 77-84.

Chuang, Jason, Sonal Gupta, Christopher D. Manning ja Jeffrey Heer. 2013. *Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment*. International Conference on Machine Learning (ICML), 2013. Saatavilla: http://vis.stanford.edu/files/2013-TopicModelDiagnostics-ICML.pdf Luettu: 16.3.2017.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, ja A. McCallum. 2011. Optimizing semantic coherence in topic models. *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 262-272.

Sievert, Carson ja Kenneth. E. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.

MUUTOKSET LUOKITTELUOHJEESEEN

Versio 1.1 (22.8.2017): Lisätty luokka **yleissanoja** ja poistettu ohje luokitella yleissanat laadukkaiksi nimikkeellä ”tavallisia sanoja”.

Versio 1.2 (5.9.2017): Täsmennys luokkaan vinoutunut: ”Jos sanasto liittyy johdonmukaisesti samaan teemaan, kyseessä on laadukas puheenaihe.”