LSAren oinarri matematikoa

A. Zelaia, I. Baragaña eta Y. Yurramendi

Konputazio Zientziak eta Adimen Artifiziala saila
Informatika Fakultatea
Euskal Herriko Unibertsitatea
DONOSTIA
ccpzejaa@si.ehu.es

Laburpena: Testu idatzien semantika adierazteko gaitasuna duen tresna bat da Latent Semantic Analysis (LSA). Modu matematikoan adierazten ditu testuko paragrafo eta hitzak. Ondoren, adierazpen matematiko horren gainean zenbait eraldaketa burutzen ditu, eta horrela, testuen eta bertan dauden hitzen arteko erlazio semantikoak neurtzeko gai da. Artikulu honetan adierazpen matematikoa eta egiten zaizkion eraldaketak aztertzen dira, tresnaren funtzionamendua ulertzeko; teorikoki hasteko eta adibide baten bidez ondoren, kontzeptu teorikoak ulertzen lagunduko duelakoan gaude.

1. TEORIA

LSAren oinarria den teoria pausoz pauso garatuko dugu atal honetan, LSAri emandako testuak nola antolatzen dituen ezagutzetik hasi eta erabiltzaileak egindako antzekotasun semantikoen eskaerei erantzuna ematen dien arte. Bertan azaldutakoa ulertzeko nahiko ezagutza matematikoa ez duen irakurleari [7] eta [9] lanetara jotzea gomendatzen diogu. Bestalde, artikulu honetan azaldutakoak modu sakonagoan aztertzeko bibliografiako [1], [2] eta [4] aztertzea gomendatzen dugu.

1.1. Corpusetik M matrizera. Aurretiko zenbait definizio

LSA tresnak testuak irakurtzen ditu. Testu horiek **corpusa** osatzen dute, eta normulean oso testu luzeak izaten dira. Ezer baino lehen, corpusari dagokion adierazpen matematikoa egingo du matrize baten bidez.

Corpuseko paragrafo bakoitza, hau da, bi lerro zuriren artean dagoen testu zatia, **dokumentua** dela esaten da. LSAk dokumentu bakoitza matrizeko zutabe batekin egokituko du.

Bestalde, corpusean zurigunez banatutako karaktere-segida oro hitza da, eta hitz batzuk **termino** izango dira (kasurik errazenean, dokumentu bat baino gehiagotan agertzen direnak, beste aukera batzuk badaude ere). LSAk termino bakoitzari matrizeko errenkada bat egokituko dio.

Termino bakoitza dokumentu bakoitzean zenbat aldiz agertzen den kalkulatu, eta m_{ij} maiztasun horiek izango dira matrizean kokatuko dituen balioak. m termino eta n dokumentu izanik, matrizeko errenkada bakoitza termino bat adierazten duen bektorea izango da $\mathbb{R}^n$ espazioan, eta zutabe bakoitza dokumentu bat adierazten duen bektorea $\mathbb{R}^m$ espazioan. Horrela lortuko du ondoren aipatzen den $\mathbf{M} \in \mathbb{R}^{m \times n}$ balio errealeko m errenkadako eta n zutabeko matrizea.

1.2. M Matrizearen aurreprosezaketa

LSAk sortu duen **M** matrize hori bere horretan erabil daiteke, edo erabiltzaileak hala eskatuta, eraldatu egin daiteke. Eraldaketa horri matrizearen **aurreprosezaketa** esaten zaio, eta helburna dokumentuetatik aukeratutako terminoei garrantzi maila desberdinak ematea da. Izan ere, **M** matrizean kokatutako m_{ij} balioak maiztasunak dira, eta balio horiek oso desberdinak izaten dira maiz. Termino bakoitzaren agerpenaren mailak, bai dokumentu bakoitzaren barruan eta bai corpus mailan, termino hori zenbateraino esanguratsua den iradoki dezake, eta matrizearen aurreprosezaketak informazio hori erabili nahi du esanguratsuagoa izango den beste matrize bat lortzeko.

Aurreprosezaketa pisuen bidez egiten da; $L(i, j)$ **pisu lokalak** i terminoak j dokumentuan duen garrantzia neurtzen du eta $G(i)$ **pisu globalak** i terminoaren garrantzia neurtzen du corpus mailan. Ez da pisu globalik definitzen dokumentuetarako. Horrek esan nahi du dokumentu guztiei garrantzi edo pisu berbera ematen zaiela corpusean.

Erabiltzaileak LSAri esan beharko dio zein pisu lokal eta zein pisu global erabili nahi dituen. Horren arabera, **M** matrizeko m_{ij} balioak beste balio hauen bidez ordezkatuko dira:

$$m'_{ij} = L(i, j) \cdot G(i).$$

Pisu lokal eta globaletarako LSA tresnak eskaintzen dituen aukerak honakoak dira:

— $L(i, j)$ pisu lokalak. i terminoak j dokumentuan duen garrantzia neurtzeko LSA-k hiru maila bereizten ditu:

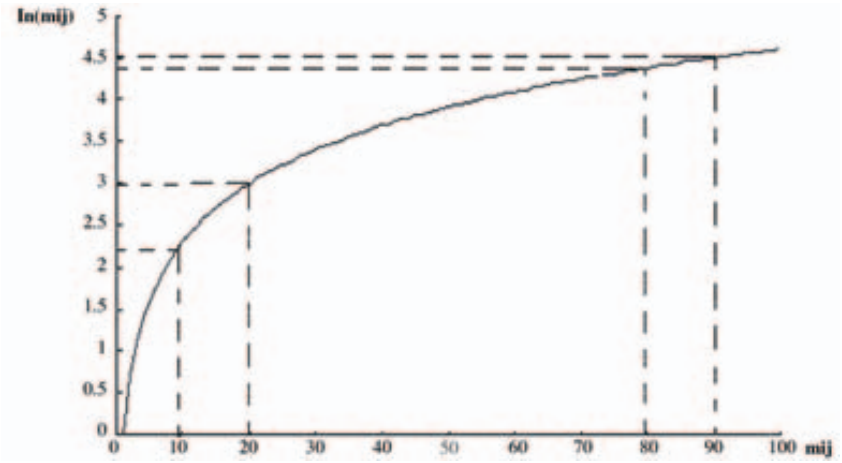
- **tf** edo «Term frequency». $L(i, j) = \mathbf{tf}(\mathrm{i}, \mathrm{j}) = m_{ij}$
  i terminoa j dokumentuan zenbat aldiz agertzen den adierazten du, terminoaren maiztasuna dokumentu horretan, alegia. Hori da hain

zuzen ere, jatorrizko **M** matrizeak jasotzen duen informazioa, eta hortaz, pisar lokal hau erabiltzea lokalki eraldaketarik ez egitea bezalakoa da. Zenbat eta maiztasun handiagoa izan termino batek dokumentu babean, orduan eta pisar handiagoa izango du terminoak lokalki.
Ikus dezagun adibide bat. Demagun 7 dokumentuz (d_1, ..., d_7) osatutako corpus babean bi terminaren (t_1, t_2) agerpenen maiztasunak honakoak direla:

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
t_1	2	100	8	0	300	10	50
t_2	4	0	4	100	40	10	500

Taulan ikus daitekeenez, terminoen agerpenen maiztasunak nahiko desberdinak izan daitezke dokumentu desberdinetarako.

- **log:** $L(i, j) = \boldsymbol{log}(i, j) = log(m_{ij} + 1)$.
  Logaritmo funtzioak maiztasunen arteko aldea murriztea eragiten du. Ikus dezagun grafiko baten bidez logaritmo funtzioaren eragina zein den.



Grafikan terminoen m_{ij} maiztasunak, hau da, jatorrizko **M** matrizearen balioak, eta halen logaritmoak daude irudikatuta. Logaritmoak terminoen maiztasun handienen arteko aldea txikitzera jotzen dar. Horrela, maiztasunen ordez beren logaritmoak idatziz, matrizeko balioak pixka bat uniformeagoak egiten ditugu. Kontuan izan behar da oso maiztasun desberdinak ager daitezkeela **M** jatorrizko matrizean.
Azter dezagun adibide baten bidez zein den pisu lokal honen eragina. Har dezagun **tf**(*i, j*) pisar lokala aztertzeko adibide moduan erabili dugun taula berbera. Taula horretako maiztasun bakoitza-

ren gainean **log**(*i, j*) funtzioa kalkulatuz gero, honako baste taula hau lortuko dugu:

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
t_1	1.1	4.6	2.1	0	5.7	2.4	3.9
t_2	1.6	0	1.6	4.6	3.7	2.4	6.2

Ikus daitekeen bezala, taulako balioek berdintze aldera jo dute. Gainera, zenbat eta handiagoak izan taulako maiztasunak, orduan eta gehiago txikitu dira balioak logaritmoaren eraginez.

- **bin:** $L(i, j) = \mathbf{bin}(i, j) = \min\{m_{ij}, 1\}$
  Eraldaketa lokal honek matrizeko m_{ij} maiztasunak baztertzea dakar, eta maiztasun horien ordez balio bitarrak kokatuko dira, hau da,

$$\mathbf{bin}(i, j) = \begin{cases} 1 & i \text{ terminoa } j \text{ dokumentuan gutxienez behin agertzen bada} \\ 0 & \text{bestelako kasuetan.} \end{cases}$$

  Honela, agerpena bai ala ez besterik ez du jasoko eraldatu ondorengo matrizeak.
  Aurreko adibidearekin jarraituz, honela geratuko dira taulako maiztasunak **bin**(*i, j*) funtzioa ezarri eta gero:

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
t_1	1	1	1	0	1	1	1
t_2	1	0	1	1	1	1	1

Laburbilduz, lokalki ditugun hiru aukera horien arabera, maiztasunak bere horretan utz ditzakegu, desberdintasunak berdintze aldera jo, edo bestela, matrizeko balioak bitar bihurtu.

— $G(i)$ pisar globalak. *i* terminoak corpus mailan duen garrantzia neurtzen du. Horrela, termino bakoitzari bere pisu globala egokituko zaio. LSAk lau maila bereizten ditu:

- **none:** $G(i) = \mathbf{none}(i) = 1$. Pisu global hau konstantea da, hau da, berbera da termino guztietarako. Horrek esan nahi du termino guztiei garrantzi berbera eman nahi zaiela corpus mailan.
  Pisa lokal moduan **tf**(*i, j*) eta global moduan **none**(*i*) aukeratzeak jatorrizko **M** matrizea bere horretan uztea dakar.

$$m'_{ij} = \mathbf{tf}(i, j) \cdot \mathbf{none}(i) = m_{ij} \cdot 1 = m_{ij}$$

- **normal:** $G(i) = \mathbf{normal}(i) = \dfrac{1}{\|\mathbf{t}_i\|} = \dfrac{1}{\sqrt{\sum_{j=1}^{n} m_{ij}^2}}$

non $\mathbf{t}_i$ den t_i terminoa adierazten duen bektorea. Pisa global hau t_1, ..., t_m terminoei ezarriz terminoak adierazten dituzten bektore guztiak normalizatzea lortzen da, hau da, bektoreak luzera berekoak izatea.
Adibidea. Har dezagun d_1, ..., d_6 dokumentuez osatutako corpus bat (ez da aurreko adibidekoa) eta azter ditzagun corpuseko t_1 eta t_2 terminoen agerpenak:

	d_1	d_2	d_3	d_4	d_5	d_6	$\|\mathbf{t}_i\|$	$\frac{1}{\|\mathbf{t}_i\|}$
t_1	10	4	2	6	10	4	$\sqrt{100+16+4+36+100+16} = 16,5$	$\frac{1}{16,5} = 0,06$
t_2	2	0	0	2	0	0	$\sqrt{4+0+0+4+0+0} = 2,83$	$\frac{1}{2,83} = 0,35$

Taulan egindako kalkuluen arabera, t_1 terminoari **normal**(1) = = 0.06 pisu globala dagokio eta t_2 terminoari **normal**(2) = 0.35.
normal(i) funtzioak **M** matrizearen eraldaketan duen eragina ulertzeko, har dezagun maiztasunen taula, eta eralda ditzagun maiztasunak modu honetan:

$$m'_{ij} = \mathbf{tf}(i, j) \cdot \mathbf{normal}(i).$$

Taula honela geratuko da:

	d_1	d_2	d_3	d_4	d_5	d_6
t_1	0.6	0.24	0.12	0.36	0.6	0.24
t_2	0.7	0	0	0.7	0	0

Eraldatu ondoren lortutako taulan t_1 eta t_2 terminoei dagozkien bektoreak normalizatuta daude (norma 1 da). Egiazta dezagun hori:

$$\|t_1\| = \sqrt{0,6^2 + 0,24^2 + 0,12^2 + 0,36^2 + 0,6^2 + 0,24^2} = 0,99$$

$$\|t_2\| = \sqrt{0,7^2 + 0,7^2} = \sqrt{0,5 + 0,5} = 1$$

Ikus daitekeenez, bi terminoak adierazten dituzten bektoreak normalizatuta egotea lortu dugu (biribilketa-erroreak barne).

- **idf** edo «Inverse Document Frequency». $G(i) = \mathbf{idf}(i) = \frac{gf(i)}{df(i)}$, non

  * $gf(i)$: «global frequency». Terminoaren maiztasun globala neurtzen du, hau da, corpus osoan zenbat aldiz agertzen den:

$$gf(i) = \sum_{j=1}^{n} m_{ij}\,.$$

* *df*(*i*)*:* «document frequency». Terminoa zenbat dokumentutan azaltzen den adierazten du:

$$df(i) = \sum_{j=1}^{n} \min\left\{m_{ij}, 1\right\}.$$

Azter dezagun adibide bat. Corpusa 7 dokumentuz osatuta dago, d_1, d_2, ..., d_7 eta t_1, t_2 eta t_3 errenkadek **M** matrizeko hiru errenkada adierazten dituzte, hiru terminori dagozkien maiztasunak alegia.

Terminoak	d_1	d_2	d_3	d_4	d_5	d_6	d_7	$gf(i)$	$df(i)$	$idf = \frac{gf(i)}{df(i)}$
t_1	2	0	5	0	0	8	0	15	3	5
t_2	5	5	5	5	5	5	5	35	7	5
t_3	100	50	10	0	500	65	0	725	5	145

t_1 terminoa 3 dokumentutan besterik ez da agertzen ($df(1) = 3$), guztira 15 aldiz ($gf(1) = 15$), hau da, batez bastean 5 aldiz (**idf**(1) = 5). t_2 terminoa aldiz, dokumentu guztietan agertzen da (7 dokumentu), eta batez beste bakoitzean 5 aldiz agertzen da (**idf**(2) = 5). t_3 terminoa 5 dokumentutan agertzen da, batez beste 145 aldiz dokumentu bakoitzean (**idf**(3) = 145). Hortaz, pisu global honen arabera t_1 eta t_2 terminoek pisu berbera dute (**idf**(1) = **idf**(2) = 5); t_3 terminoari pisu handiagoa egokitu zaio (**idf**(3) = 145).
Informazioaren erauzketaren alorrean **tf**(*i, j*) pisar lokala **idf**(*i*) pisu globalarekin konbinatuta erabili ohi da (ikus 1.6. atala):

$$m'_{ij} = \mathbf{tf}(i, j) \cdot \mathbf{idf}(i).$$

Adibidearekin jarraituz, taula honela eraldatuko litzateke.

Terminoak	d_1	d_2	d_3	d_4	d_5	d_6	d_7
t_1	10	0	25	0	0	40	0
t_2	25	25	25	25	25	25	25
t_3	14.500	7.250	1.450	0	72.500	9.425	0

Ikus daitekeen bezala, t_3 terminoari dagokion pisu globala oraingoan oso altua izan da, eta hortaz, matrizearen eraldaketak matrizeko balioak beren artean are gehiago bereiztea eragin du.

- **entropy:**

$$G(i) = \mathbf{entropy}(i) = -\sum_{j=1}^{n} p_{ij} \log_2\left(p_{ij}\right), \quad \text{non } p_{ij} = \frac{m_{ij}}{gf(i)}.$$

Entropiak terminoek dokumentuetan duten maiztasunen arteko desoreka neurtzen du. Ikus dezagun adibide baten bidez horrek zer esan nahi duen.

Ondoko taulan 6 dokumentuz osatutako corpusa eta dokumentu horietan zazpi terminok dituzten m_{ij} maiztasunak ageri dira. Azken zutabean termino bakoitzari dagokion entropia kalkulatuta dago:

	d_1	d_2	d_3	d_4	d_5	d_6	**entropy**(i)
t_1	20	20	20	20	20	20	$-6\left(\frac{20}{120}\log_2\frac{20}{120}\right) = \log_2 6 = 2{,}5$
t_2	6	6	6	6	6	6	$-6\left(\frac{6}{36}\log_2\frac{6}{36}\right) = \log_2 6 = 2{,}5$
t_3	10	4	2	6	10	4	$-\left(\frac{10}{36}\log_2\frac{10}{36} + ... + \frac{4}{36}log_2\frac{4}{36}\right) = 2{,}33$
t_4	2	0	0	2	0	0	$-\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}\right) = 1$
t_5	10	0	0	0	10	0	$-\left(\frac{10}{20}\log_2\frac{10}{20} + \frac{10}{20}log_2\frac{10}{20}\right) = 1$
t_6	1	0	0	0	3	0	$-\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}log_2\frac{3}{4}\right) = 0{,}8$
t_7	0	0	0	0	0	500	$-\left(\frac{500}{500}\log_2\frac{500}{500}\right) = 0$

Taulan ikus dezakegunez, t_7 terminoa corpuseko dokumentu bakar batean ageri da, eta ondorioz, entropia zero da. t_4, t_5 eta t_6 terminoak bi dokumentutan ageri dira, baina beren entropiak ez dira berdinak, terminoen agerpenak ez direlako parekoak. t_4 eta t_5 terminoek entropia berbera dute, bi terminoak bi dokumentutan besterik ez direlako ageri, eta gainera, bi dokumentu horietan modu orekatuan ageri direlako (2 aldiz t_4 eta 10 aldiz t_5). t_6 terminoa, aldiz, nahiz eta bi dokumentutan besterik ez agertu, batean bastean baino maiztasun handiagoaz ageri da, eta desoreka horren adierazgarri, entropia txikiagoa lortzen da. t_3 terminoa dokumentu guztietan ageri da, kopuru berberean agertzen ez bada ere, eta dagokion entropia altuagoa da. t_1 eta t_2 terminoek agertzen dute entropiarik handiena; bietarako berbera, eta 6 dokumentuz osatutako corpus babean egon daitekeen entropiarik handiena.

Hortaz, ikus daiteke entropiak maiztasunen arteko oreka neurtzen duela. Maiztasunak zenbat eta orekatuagoak agertu, orduan eta entropia handiagoa izango dute.

LSAn pisar lokal moduan **log**(i, j) eta global moduan entropy(i) erabili ohi da.

$$m'_{ij} = \mathbf{log}(i, j) \cdot \mathbf{entropy}(i)$$

Eralda ditzagun taulako balioak pisu horiek erabiliz:

	d_1	d_2	d_3	d_4	d_5	d_6
t_1	7.5	7.5	7 5	7.5	7.5	7.5
t_2	4.75	4.75	4.75	4.75	4.75	4.75
t_3	5.6	3.7	2.6	4.4	5.6	3.7
t_4	1.1	0	0	1.1	0	0
t_5	2.4	0	0	0	2.4	0
t_6	0.56	0	0	0	1.12	0
t_7	0	0	0	0	0	0

t_7 terminoaren entropia zero izateak terminoaren agerpen guztiak zero bihurtzea eragin du.

1.3. **Matrizea belio singularretan deskonposatzea**

Matrizea sortu eta erabiltzaileak esandako pisu lokal eta globalak erabiliz eraldatu ondoren, **M** matrize horren deskonposaketetako bat kalkulatuko da. Ondoren aipatzen den teoremak ziurtatzen du $\mathbf{M} \in \mathbb{R}^{m\times n}$ matrize oro deskonposa daitekeela **balio singularretan deskonposatzea** esaten zaion beste hiru matrizeren biderkaketa moduan. Teoremak honela dio:

1 teorema Izan bedi $\mathbf{M} \in \mathbb{R}^{m\times n}$ matrizea. Beti da posible balio singularretan honela deskonposatzea:

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T \qquad k = \min\{m, n\},$$

non

- $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_m] \in \mathbb{R}^{m\times m}$: $\mathbf{u}_i$ zutabe bektoreak ortonormalak dira elkarren artean. **M**ren ezker bektore singularrak direla esaten da.
- $\Sigma \in \mathbb{R}^{m\times n}$ matrize diagonala da. Diagonaleko σ_i balioak **M** matrizearen **balio singularrak** dira non $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k \geq 0$ betetzen den $k = \min\{m, n\}$.
- $\mathbf{V} \in \mathbb{R}^{m\times n}$ $\mathbf{v}_i$ zutabe bektoreak ortonormalak dira. **M** matrizearen eskuin bektore singularrak direla esaten da.

Zeroren desberdinak diren balio singularren kopurua **M** matrizearen r heina da, $r \leq k$ delarik. **M** matrizearen balio singularrak bakarrak dira.

Badira matrize baten balio singularrak eta dagozkien bektore singularrak kalkulatzeko teknikak, eta horrela lortzen ditu LSAk **U**, Σ eta **V** matrizeak.

1.4. **Bektore-espazioaren dimentsioaren murrizketa: espazio murriztua**

Behin matrizea balio singularretan deskonposatzea lortu denean, posible izango da jatorrizko **M** matrize horren hurbilketa izango den beste $\mathbf{M}_p$ matrize bat kalkulatzea. Horretarako egin behar dena da balio singularrez osatutako Σ matrizean zenbait balio singular, balio txikieneko batzuk, baztertu. Σ matrizean $p \leq r$ balio singular $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$ eta dagozkien $\mathbf{u}_i \in \mathbb{R}^m$ eta $\mathbf{v}_i \in \mathbb{R}^n$ bektore singularrak mantenduz lortuko da $\mathbf{M}_p$ matrizea. Eragiketa horri **dimentsioaren murrizketa** esaten zaio, eta aukeratu beharreko *p* balioari **dimentsioa**.

Ondoren aipatzen den beste teorema hori $\mathbf{M}_p$ matrize honi buruzkoa da, eta ziurtatzen du $\mathbf{M}_p$ dela heina *p* edo txikiagoa duten matrizeen artetik jatorrizko **M** matrizera gehien hurbilduko den matrizeetako bat.

2 teorema (Eckart, Young) Izan bedi aurreko teoreman (1) ekuazioak emandako **M** matrizearen deskonposizioa balio singularretan, non **M** matrizearen heina *r* den, $r = rang(\mathbf{M}) \leq k = \min(m, n)$. Ondokoa definitzen da:

$$\mathbf{M}_p = \mathrm{U}_p \, \Sigma_p \, \mathrm{V}_p^T = \sum_{i=1}^{p} \sigma_i u_i v_i^T \quad p \leq r,$$

non $\mathbf{U}_p$ eta $\mathbf{V}_p$ matrizeak diren **U** eta **V** matrizeen lehenengo *p* zutabeez osatutako matrizeak. Ondokoa betetzen da:

$$\min_{rang(A) \leq p} \|\mathbf{M} - \mathbf{A}\| = \|\mathbf{M} - \mathbf{M}_p\| = \sigma_{p+1}$$

non $\|\mathbf{A}\|$ notazioaren bidez matrizeen norma espektrala adierazten den.

Jatorrizko **M** matrizearen eta haren hurbilketa den $\mathbf{M}_p$-ren arteko distantzia σ_{p+1} dela ziurtatzen zaigu. Hortaz, dimentsioa murrizten denean, guk neur dezakegu zenbateraino gerturatu nahi dugun $\mathbf{M}_p$ matrizea jatorrizko **M** matrizera, *p* dimentsioa guk aukeratuko dugulako.

Dimentsioa murriztearen ondorioz, jatorrizko **M** matrizearen zutabeek sortutako bektore-espaziotik $\mathbf{U}_p$ matrizearen zutabeek sortutakora pasa gara, $Im\ \mathbf{U}_p = Im\ \mathbf{M}_p$, eta espazio berri honi *p* dimentsioko **bektore-espazio murriztua** edo **espazio semantikoa** esaten zaio.

Erabiltzaileak erabaki eta LSAri esan behar dio zein *p* dimentsiotara murriztu nahi duen jatorrizko bektore-espazioa. *p*-ren ankeraketa oso garrantzitsua da, LSA-k emandako emaitzak *p* dimentsio horren arabera aldatuko baitira. Hala ere, ez da ezagutzen dimentsio egokia aukeratu ahal izateko metodorik. Hortaz, *p* dimentsioa enpirikoki aukeratu behar da, hau da, dimentsio desberdinetarako probak egin, eta emaitza onenak ematen dituena ankeratu bekarko da.

Zergatik murriztu espazioaren dimentsioa? Zergatik lan egin dimentsio murriztuko espazio horretan? Azken finean, dimentsioa murriztea jatorrizko bektoreekin lan egin ordez bektore horiek *p* dimentsioko espazioan duten proiekzioarekin lan egitea da, eta bidean informazioa galtzen dela pentsa dezakegu. Hala ere, harrigarria badirudi ere, antzekotasun semantikoen neurketek emaitza hobeak ematen dituzte espazio murriztuan jatorrizkoan baino, *p*-ren ankeraketa egokia izan bada behintzat; bektoreen arteko erlazio semantikoak hobeto adierazten dira espazio murriztuan.

1.5. **Bektoreen arteko antzekotasunaren neurketa espazio murriztuan**

M matrizea sortu, aurreprosezaketa egin, balio singularretan deskonposatu eta dimentsioa murriztu ondoren lortzen den espazio murriztu horretan egingo ditugu bektoreen arteko antzekotasunen neurketak. Espazio semantiko horrek adierazten du LSAk corpuseko dokumentuen eta terminoen semantikari buruz duen ezagutza. Orain prest dago antzekotasun semantikoak neurtzen hasteko.

Espazio murriztu horretan bai terminoak adierazten dituzten bektoreen proiekzioak eta bai dokumentuak adierazten dituzten bektoreenak ditugu. Horrela, bi terminoren arteko antzekotasuna, bi dokumenturen arteko antzekotasuna edota termino baten eta dokumentu baten arteko antzekotasuna kalkula ditzakegu. Izan bitez **s** eta **t** termino edota dokumentu horiek adierazten dituzten bi bektore. Bi bektore horien proiekzioa espazio murriztuan honela kalkulatzen da:

$$\mathrm{s}_p = \mathrm{s}^T \mathrm{U}_p \, \Sigma_p^{-1}$$

$$\mathrm{t}_p = \mathrm{t}^T \mathrm{U}_p \, \Sigma_p^{-1} .$$

$\mathbf{s}_p$ eta $\mathbf{t}_p$ bektoreen osagaiak **s** eta **t** bektoreen proiekzioen koordenatuak dira espazio murriztuan. $\mathbf{s}_p$ eta $\mathbf{t}_p$ bektoreen arteko antzekotasuna neurtzeko bi neurri proposatzen dira:

—**Biderkadura eskalarra.** $\mathbf{s}_p^T \cdot \mathbf{t}_p$.

$$\mathrm{s}_p^T \cdot \mathrm{t}_p = \left(s_1, s_2, ..., s_p\right) \cdot \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{pmatrix} = \sum_{i=1}^{p} s_i t_i = s_1 t_1 + s_2 t_2 + ... + s_p t_p$$

Biderkadura eskalarraren emaitza 0tik gertu badago, $\mathbf{s}_p$ eta $\mathbf{t}_p$ bektoreek adierazten dituzten termino edota dokumentuak semantikoki antzekoak ez direla interpretatuko dugu.

Kontuan izan behar da s_i eta t_i balioak oso altuak izan daitezkeela, balio horiek m_{ij} maiztasunetan oinarrituta daudelako, eta hortaz, biderkadura eskalarrean kointzidentzia bakar batek emaitza altua emango luke. Hala ere, horrek ez luke antzekotasun semantiko handia adierazi beharko. Arazo hori ekidingo duen neurketa bi bektoreek osatzen duten angeluaren kosinua da.

—**Kosinua**. Espazio murriztuko $\mathbf{s}_p$ eta $\mathbf{t}_p$ bi bektoreren arteko θ angeluaren kosinua honela kalkulatzen da:

$$\cos\,\theta = \frac{\mathbf{s}_p^T\,\mathbf{t}_p}{\|\mathbf{s}_p\|\,\|\mathbf{t}_p\|}$$

Ikus daitekeenez, kosinuaren kalkuluan ere biderkadura eskalarra daukagu, baina normalizatuta dago; bektoreen normen biderkaduraz zatituta. Honek esan nahi du bi bektoreek norabide berbera badute, θ angelua zero izango dela, eta ondorioz, kosinua 1. Semantikoki guztiz berdinak direla esango dugu. Bestalde, bektoreak perpendikularrak badira, θ angeluaren kosinua zero izango da, eta horrek antzekotasun semantiko eza iradokiko du.

Erabiltzaileak LSAri jakinaraziko dio espazio murriztuko zein termino edota dokumenturen arteko antzekotasun semantikoa neurtu nahi duen, hala nola biderkadura eskalarra edo kosinuaren emaitza nahi duen. Neurketaren emaitza itzuliko dio LSAk erabiltzaileari.

1.6. **LSA informazioaren erauzketarako: LSI**

Informazioaren erauzketa edo «Information Retrieval» delakoa erabiltzen da erabiltzaileak gai baten inguruko testuak aurkitu nahi dituenean. Horretarako, erabiltzaileak zenbait hitz idatziko ditu bilatzaile batean, eta honek horien agerpenak dituzten dokumentuak aurkitu eta erabiltzaileari itzuliko dizkio.

Edonola, bilaketa modu honetan egiteak arazoak sortzen ditu. Kontuan izan behar da zenbaitetan hitz batek esanahi bat baino gehiago duela, eta balitekeela modu honetan aurkitutako zenbait dokumentu erabiltzailearen interesekoak ez izatea. Baliteke baita sinonimoak direla eta, dokumenturen batean bere interesekoa den gai bat jorratzea, baina beste hitz sinonimo bat erabiltzeagatik bilatzaileak ez aurkitzea.

Hori dela eta, LSA oso tresna erabilgarria izan daiteke, testuen semantikan oinarritzen delako bilaketa eta ez kointzidentzia hutsean. LSA tresna informazioaren erauzketarako erabiltzen denean **LSI** esaten zaio: **Latent Semantic Indexing.**

Informazioaren erauzketa egiteko modua honakoa da: aztertuko diren dokumentu guztiekin corpusa osatzen da, eta aurreko ataletan azaldu den moduan corpus honi dagokion matrizea eraiki eta aurreprosezaketa, deskonposaketa eta dimentsioaren murrizketa eginez, espazio murriztua lortuko da. Corpusa handia bada, prosezu honek denbora luzea eska dezake eta espazio okupazio handia. Ondoren, erabiltzaileak hitz-zerrenda bat idatziko du corpus horretan bere interesekoa izango den dokumentu bat bilatu nahi duen bakoitzcan: «kontsulta». Kontsulta hori espazio semantikoko dokumentuekin konparatu ahal izateko, haren adierazpen bektoriala kalkulatu beharko da. Eraikiko dugun **q** bektorea balio bitarrez osatutako $\mathbf{q}^T = (q_1, q_2, ..., q_m)$ bektorea izango da, non

$$q_i = \begin{cases} 1 & \text{espazioko i terminoa kontsultan agertzen bada} \\ 0 & \text{bestelako kasuetan} \end{cases}$$

Bektore horrek *m* osagai ditu, espazio semantikoko termino bakoitzeko bat. Horrela, corpuseko dokumentuek jatorrizko **M** matrizean duten adierazpenaren parekoa den bektorea osatzen da, matrizeak zutabe bat gehiago balu bezala. Horregatik, bektore honi **sasidokumentu** esaten zaio.

Ondoren, $\mathbf{q}^T$ bektorearen adierazpena lortu behar da espazio murriztuan:

$$\mathrm{q}_p = \mathrm{q}^T \mathrm{U}_p \sum\nolimits_p^{-1}.$$

$\mathbf{q}_p$ bektorearen osagaiak **q** bektorearen proiekzioaren koordenatuak dira espazio murriztuan. Behin hori izanda, $\mathbf{q}_p$ bektoreak espazio murriztuko $\mathbf{s}_p$ dokumentuekin duen antzekotasuna neur daiteke. Kosinua erabiliz,

$$cos\ \theta = \frac{\mathbf{q}_p^T \mathbf{s}_p}{\|\mathbf{q}_p\| \, \|\mathbf{s}_p\|},$$

non $\mathbf{s}_p$ bektorea **s** dokumentuari *p* dimentsioko espazioan dagokion proiekzioaren koordenatuak diren.

Erabiltzaileari kosinu handieneko dokumentuak itzuliko zaizkio bilaketaren emaitza moduan, horiek izango baitira kontsultarekin antzekotasun semantiko handienekoak.

2. **ADIBIDEA**

Azaldutako teoria hobeki ulertzeko, adibide txiki bat hartu eta hasieratik bukaeraraino garatuko da. Corpusa 9 dokumentu motzek osatuko dute, **M** matrizea tamaina txikikoa izan dadin. Adibide hau bibliografian aurkitu dugu (ikus [5], [6] eta [3]).

2.1. **Corpusa eta M matrizea**

Corpusa osatuko duten dokumentuak testu teknikoen 9 izenburu dira. Dokumentu horiek bi motatakoak dira: «gizaklaren eta konputagailuaren arteko interakzioa»ri buruzko 5 dokumentu (c1-c5) eta «grafoen teoria»ri buruzko 4 dokumentu (ml-m4). Bi arlo hauek elkarren artean nahiko desberdinak dira.

Dokumentu batean baino gehiagotan agertzen diren hitzak termino gisa ankeratuko dira. Guztira 12 dira, eta azpimarratuta datoz. Hona hemen 9 dokumentuak.

cl: Human machine interface for Lab ABC computer applications.

c2: A survey of user opinion of computer system response time.

c3: The EPS user interface management system.

c4: System and human system engineering testing of EPS.

c5: Relation of user-perceived response time to error measurement.

ml: The generation of random, binary, unordered trees.

m2: The intersection graph of paths in trees.

m3: Graph minors IV: Widths of trees and well-quasi-ordering.

m4: Graph minors: A survey.

9 dokumentu eta 12 termino horiekin ondoren zehaztuko den $\mathbf{M} \in \mathbb{R}^{12\times 9}$ matrizea eraikiko du. Matrizeko osagaiak maiztasunak dira, hau da, terminoa dokumentuan zenbat aldiz agertzen den. Hona hemen **M** matrizea (bi posizio laukitxotan sartuta ageri dira aurrerago egingo zaien ereferentzia errazteko asmoz):

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & \boxed{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ c1 & c2 & c3 & c4 & c5 & m1 & m2 & m3 & m4 \end{pmatrix} \begin{matrix} \\ \rightarrow \text{human} \\ \rightarrow \text{interface} \\ \rightarrow \text{computer} \\ \rightarrow \text{user} \\ \rightarrow \text{system} \\ \rightarrow \text{responde} \\ \rightarrow \text{time} \\ \rightarrow \text{EPS} \\ \rightarrow \text{survey} \\ \rightarrow \text{trees} \\ \rightarrow \text{graph} \\ \rightarrow \text{minors} \\ \\ \end{matrix}$$

Nahi izanez gero, matrize horren **aurreprosezaketa** egin daiteke, eta pisuen bidez matrizeko osagaiak eraldatu. Adibide hau ulergarriagoa izan

dadin, pisu lokal moduan **tf**(*i, j*) eta global moduan **none**(*i*) aukeratu dira (ikus 1.2), **M** matrizea bere horretan uzteko.

2.2. **Balio singularretan deskonposatzea**

M matrizearen balio singularrak eta dagozkien bektore singularrak kalkulatuz, honela deskonposatzen da **M** matrizea balio singularretan:

$$\mathbf{U} = \begin{pmatrix} 0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & -0.41 \\ 0.20 & -0.07 & 0.14 & -0.55 & 0.28 & 0.50 & -0.07 & -0.01 & -0.11 \\ 0.24 & 0.04 & -0.16 & -0.59 & -0.11 & -0.25 & -0.30 & 0.06 & 0.49 \\ 0.40 & 0.06 & -0.34 & 0.10 & 0.33 & 0.38 & 0.00 & 0.00 & 0.01 \\ 0.64 & -0.17 & 0.36 & 0.33 & -0.16 & -0.21 & -0.17 & 0.03 & 0.27 \\ 0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\ 0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\ 0.30 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & -0.17 \\ 0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & -0.58 \\ 0.01 & 0.49 & 0.23 & 0.03 & 0.59 & -0.39 & -0.29 & 0.25 & -0.23 \\ 0.04 & 0.62 & 0.22 & 0.00 & -0.07 & 0.11 & 0.16 & -0.68 & 0.23 \\ 0.03 & 0.45 & 0.14 & -0.01 & -0.30 & 0.28 & 0.34 & 0.68 & 0.18 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 3.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.54 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.35 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.64 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.50 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.31 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.85 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.56 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.36 \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} 0.20 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & -0.06 \\ 0.61 & 0.17 & -0.50 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & 0.24 \\ 0.46 & -0.13 & 0.21 & 0.04 & 0.38 & 0.72 & -0.24 & 0.01 & 0.02 \\ 0.54 & -0.23 & 0.57 & 0.27 & -0.21 & -0.37 & 0.26 & -0.02 & -0.08 \\ 0.28 & 0.11 & -0.51 & 0.15 & 0.33 & 0.03 & 0.67 & -0.06 & -0.26 \\ 0.00 & 0.19 & 0.10 & 0.02 & 0.39 & -0.30 & -0.34 & 0.45 & -0.62 \\ 0.01 & 0.44 & 0.19 & 0.02 & 0.35 & -0.21 & -0.15 & -0.76 & 0.02 \\ 0.02 & 0.62 & 0.25 & 0.01 & 0.15 & 0.00 & 0.25 & 0.45 & 0.52 \\ 0.08 & 0.53 & 0.08 & -0.03 & -0.60 & 0.36 & 0.04 & -0.07 & -0.45 \end{pmatrix}$$

Egiazta daiteke hiru matrize horien biderkaketa jatorrizko **M** matrizea dela (biribilketa-errore txikiak baztertuta). Σ matrizearen diagonalean dauden σ_i balio singularrak $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_9 \geq 0$ handienetik txikienera ordenatuta daude.

2.3. **Bektore-espazioaren dimentsioaren murrizketa: espazio murriztua**

Orain bektore-espazioaren dimentsioaren murrizketa egingo dugu. Espazioa bi dimentsiotara murriztuko dugu, horrela posible izango baita planean dokumentuak irudikatzea eta beren arteko antzekotasuna modu grafikoan ikustea. Bi dimentsiotara murrizteko, Σ matrizeko bi balio singularrik handienak eta dagozkien $\mathbf{U}_p$ eta $\mathbf{V}_p$ matrizeetako zutabe-bektoreak besterik ez ditugu kontuan izango:

$$\mathrm{M}_2 = \mathrm{U}_2 \, \Sigma_2 \, \mathrm{V}_2^T = \sum_{i=1}^{2} \sigma_i \mathrm{u}_i \mathrm{v}_i^T = \sigma_1 \mathrm{u}_1 \mathrm{v}_1^T + \sigma_2 \mathrm{u}_2 \mathrm{v}_2^T .$$

$$\mathbf{M}_2 = \begin{pmatrix} 0.22 & -0.11 \\ 0.20 & -0.07 \\ 0.24 & 0.04 \\ 0.40 & 0.06 \\ 0.64 & -0.17 \\ 0.27 & 0.11 \\ 0.27 & 0.11 \\ 0.30 & -0.14 \\ 0.21 & 0.27 \\ 0.01 & 0.49 \\ 0.04 & 0.62 \\ 0.03 & 0.45 \end{pmatrix} \begin{pmatrix} 3.34 & 0 \\ 0 & 2.54 \end{pmatrix} \begin{pmatrix} 0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.02 & 0.02 & 0.08 \\ 0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \end{pmatrix}$$

Hiru matrize horien biderkaketa burutuz gero, jatorrizko **M** matrizearen hurbilketa den honako $\mathbf{M}_2$ matrizea lortzen da:

$$\mathbf{M}_2 = \begin{pmatrix} 0.16 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\ 0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\ 0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\ 0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\ 0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\ 0.10 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & \boxed{0.42} \\ -0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & \boxed{0.66} \\ -0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\ -0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62 \end{pmatrix}$$

$\mathbf{M}_2$ matrizeari buruz bi gauza esan behar dira:

1. Dimentsioa zenbat eta handiagoa aukeratu, hau da, zenbat eta balio singular gehiago mantendu, orduan eta antzekoagoa izango da $\mathbf{M}_p$ matrizea jatorrizko **M** matrizearekin.
2. Ez da **M** erabili nahi dugun matrizea, haren hurbilketa den $\mathbf{M}_2$ baizik, bigarren horretan hobeto adierazten baitira terminoen eta dokumentuen arteko erlazioak.

2.4. **M-tik M_2-rako eraldaketaren interpretazioa. Korrelazioak**

Jatorrizko **M** matrizetik haren hurbilketa den $\mathbf{M}_2$ matrizera pasa gara. Azter ditzagun bi matrize horiek. Itxuraz, **M** eta $\mathbf{M}_2$ matrizeak nahiko desberdinak dira; osoko zenbakiak ditu lehenak, eta bigarrengoan denak erreal izatera aldatu dira. Gainera, azken honetan balio negatiboak ageri dira, jatorrizko matrizean ez bezala. Hala ere, $\mathbf{M}_2$ matrizeak terminoen eta dokumentuen arteko erlazioak hobeto adierazten ditu, eta hori da atal honetan aztertuko duguna. **M**-tik $\mathbf{M}_2$-rako eraldaketa hau interpretatzeko korrelazioak kalkulatuko ditugu [8]. Korrelazioa antzekotasun-neurri estatistikoa da, eta kasu honetan Pearson-en korrelazioa da erabili duguna. Hiru ikuspuntutatik aztertuko ditugu bi matrizeak.

(a) Matrizeetako m_{ij} balioen eraldaketa. Azter ditzagun **M** eta $\mathbf{M}_2$ matrizeetan survey eta trees terminoei eta m4 dokumentuari dagozkion bi posizioak (laukitxoan sartuta daude).

—trees terminoa m4 dokumentuan ez da ageri eta horregatik posizio horretan **M** matrizean 0 balioa ageri da. Balio hori $\mathbf{M}_2$ matrizean 0.66 izatera pasa da. Eraldaketa horretan graph eta minors terminoek eragina izan dute. Izan ere, bi termino horiek m4 dokumentuan ageri dira, eta trees, graph eta minors terminoak adierazten dituzten bektoreak oso antzekotzat jo dira. (Ikus **M** matrizean dagozkien errenkadak).

—survey terminoa m4 dokumentuan ageri da, eta jatorrizko **M** matrizean posizio horretan lekoa dago. leko hori 0.42 bihurtu da $\mathbf{M}_2$ matrizean. Honekin adierazi nahi dena da m4 dokumentuaren testuinguru horretan ez duela termino horrek pisu handiegirik, eta m4 dokumentuaren karakterizazioa den $\mathbf{M}_2$ matrizeko zutabean garrantzia kendu zaio.

(b) Terminoen arteko korrelazioak. Azter ditzagun human, user eta minors terminoak adierazten dituzten bektoreak, bai **M** matrizean eta bai $\mathbf{M}_2$ matrizean ere.

—human-user terminoen arteko korrelazioa. Bi termino horiek antzeko esanahia dutela ulertzen dugu (gizaki, erabiltzaile). **M** matrizean human eta user terminoak adierazten dituzten bektoreak hartuz (errenkadak), eta haien arteko korrelazioak kalkulatuz gero, r(human,user) = –0.38 lortzen da. $\mathbf{M}_2$ matrizeko bektoreekin kalkulatuz aldiz, r(human,user) = 0.94 lortzen da. Korrelazioak gora egin du nabarmen. Horrek esan nahi du $\mathbf{M}_2$ matrizean human eta user terminoak adierazten dituzten bektoreak beren artean gertuago daudela **M** matrizean baino. Zergatik gertatu da hau? human eta user terminoak *antzeko* esanahia daten dokumentuetan ageri direlako, dokumentu berberean

biak batera inoiz agertu ez badira ere. Hortaz, testuinguruak lagundu du.

—human-minors terminoen arteko korrelazioa. Bi termino horiek semantikoki ez dira antzekoak, minors grafoen teoriako kontzeptu bat izanik, «gizakia»ren kontzeptuarekin zerikusi handirik ez duelako. **M** matrizean bi termino horiek adierazten dituzten bektoreen arteko korrelazioa r(human,minors) = –0.29 da, eta $\mathbf{M}_2$ matrizeko bektoreena r(human,minors) = –0.83. Bi korrelazioak negatiboak dira, eta hortaz, bai batak eta bai besteak antzekotasun semantiko eza adierazten digute. Hala ere, $\mathbf{M}_2$ matrizeko bektoreen korrelazioak nabarmenago uzten du antzekotasun semantiko eza, –1 baliotik are gertuago dagoelako. Hau zergatik gertatu den ulertzeko, orain ere human terminoa eta minors terminoa zein testuingurutan agertzen diren aztertu behar da; bi testuinguruen arteko kointzidentzia eza izan da hori eragin duena.

(c) Dokumentuen arteko korrelazioak. Corpusak 9 dokumentu izanik, dokumentu pare guztietarako korrelazioak kalkulatu dira, bai **M** matrizeko zutabeak eta bai $\mathbf{M}_2$ matrizeko zutabeak kontuan izanik. Jatorrizko **M** matrizearekin kalkulatutako korrelazioak honakoak dira:

M	c1	c2	c3	c4	c5	m1	m2	m3
c2	–0.19							
c3	0.00	0.00						
c4	0.00	0.00	0.47					
c5	–0.33	0.58	0.00	–0.31				
m1	–0.17	–0.30	–0.21	–0.16	–0.17			
m2	–0.26	–0.45	–0.32	–0.24	–0.26	0.67		
m3	–0.33	–0.58	–0.41	–0.31	–0.33	0.52	0.77	
m4	–0.33	–0.19	–0.41	–0.31	–0.33	–0.17	0.26	0.56

c1-c5 dokumentuak jakintza-arlo baten ingurukoak eta ml-m4 baste batekoak direla dakigunez, beren arteko korrelazioen batezbestekoa kalkulatu eta honako taula lortu dugu:

	clc2c3c4c5	mlm2m3
c2c3c4c5	0.02	
mlm2m3m4	–0.30	0.44

«Gizaki-konputagailu interakzioa»ri buruzkoak diren cl-c5 izenburuen arteko korrelazioa nahiko baxua da, batez bestekoan 0.02. «Grafoen teoria»ri buruzko diren ml-m4 dokumentuen arteko korrelazioen batezbestekoa 0.44koa da. Bi arlo desberdinak elkarren

artean konparatuz, aldiz, –0.30ekoa atara da. Korrelazioak oso adierazgarriak ez badira ere, ikus daiteke korrelaziorik baxuena arlo desberdinen konparazioak eman duela.
Hala ere, emaitzak adierazgarriagoak dira $\mathbf{M}_2$ matrizearekin lan eginez gero. Honakoak dira dokumentu pare guztien arteko korrelazioak $\mathbf{M}_2$ matrizerako:

$\mathbf{M}_2$	cl	c2	c3	c4	c5	ml	m2	m3
c2	0.91							
c3	1.00	0.91						
c4	1.00	0.88	1.00					
c5	0.85	0.99	0.85	0.81				
m1	–0.85	–0.56	–0.85	–0.88	–0.45			
m2	–0.85	–0.56	–0.85	–0.88	–0.44	1.00		
m3	–0.85	–0.56	–0.85	–0.88	–0.44	1.00	1.00	
m4	–0.81	–0.50	–0.81	–0.84	–0.37	1.00	1.00	1.00

Lehen egin bezala, kalkula ditzagun batezbestekoak, eta bil ditzagun taula babean

	clc2c3c4c5	mlm2m3
c2c3c4c5	0.92	
mlm2m3m4	–0.72	1.00

Bi dimentsioko espaziorako topikoen arteko banaketa askoz ere bereiziago dago. «Gizaki- konputagailu interakzioa»ri buruzko izenburuen korrelazioen batezbestekoa 0.02tik 0.92ra igo da. Horrek ez du esan nahi izenburuak elkarren artean oso antzeko direnik. Izan ere, ez dira antzekoak. Baina, beste arlo batekoekin konparatu direnean, bereizi egin ditu.
«Grato teoria»ri buruzko izenburuen korrelazioen batezbestekoa 1.00 izatera pasa da. Bi arloak elkarren artean konparatuta, aldiz, korrelazioa oso txikia dela ikus daiteke: –0.72
Honetatik guztitik ondoriozta daiteke **M** matrizea balio singularretan deskonposatzeak eta espazioaren dimentsioa murrizteak inferentzia egoki asko eragin dituela.

2.5. **Dokumentuen adierazpen grafikoa espazio murriztuan**

M matrizea bi dimentsioko espaziora murriztu dugunez, planora alegia, posible da 9 dokumentuek planean zein leku hartu duten irudikatzea. Horrela, grafikoki ikusi ahal izango dugu zein den dokumentuek elkarren arte-

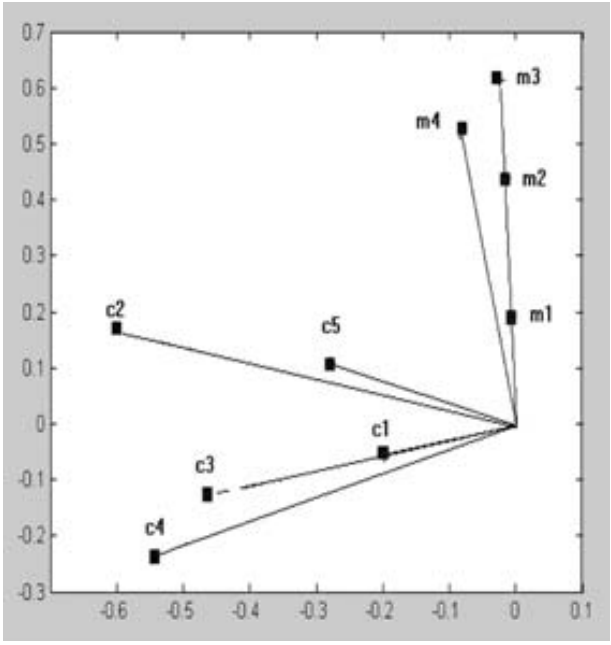
an duten antzekotasuna. **d** dokumentu bakoitzaren proiekzioa planean kalkulatzeko, hauxe egingo dugu:

$$\mathbf{d}_2 = \mathbf{d}^T\mathbf{U}_2\Sigma_2^{-1},$$

non **d** bektorea **M** matrizeko zutabe-bektoreetakoa den. $\mathbf{d}_2$ bektoreak bi osagai ditu: **d** bektorearen proiekzioaren koordenatuak dira planean. Modu honetan kalkulatuko ditugu 9 dokumentnen koordenatuak, eta beraiekin osatutako matrizea honakoa da:

$$\mathbf{D}_2 = \begin{pmatrix} -0.20 & -0.61 & -0.46 & -0.54 & -0.28 & -0.01 & -0.01 & -0.02 & -0.08 \\ -0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \end{pmatrix}$$

$\mathbf{D}_2$ matrizeko zutabe bakoitzak dokumentu baten koordenatuak adierazten ditu planean: lehenengo bost zutabeek c1-c5 dokumentuen koordenatuak eta azken lau zutabeek ml-m4 dokumentuena. Ondorengo irudian dokumentuen proiekzioak ageri dira.



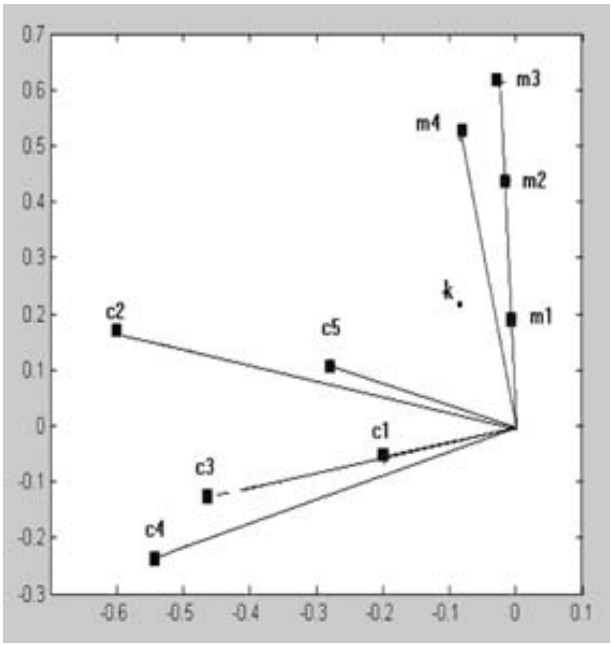
(0, 0) puntua jatorria izanik, dokumentuak adierazten dituzten bektoreak marraztu ditugu. Garbi ikus daiteke c1-c5 dokumentuak adierazten dituzten bektoreen arteko θ angelua nahiko txikia dela, eta hortaz, kosinua letik gertu egongo dela. Gauza bera gertatzen da ml-m4 dokumentuen artean ere. Grafikan c1-c5 dokumentuak alde batetik eta ml-m4 dokumentuak beste alde batetik multzokatuta ageri direla ikus daiteke.

Demagun, orain, erabiltzaileak honako beste dokumentu hau duela: «Graph theory with applications to engineering and computer science» eta

demagun jakin nahi duela dokumentu hau corpuseko zein dokumenturen antzekoa den. Dokumentu honi kontsulta esango diogu, eta hasteko, bektore bidezko adierazpena kalkulatuko dugu. Kontsultan espazio semantikoko «graph» eta «computer» terminoak ageri dira. Bi termino horiek matrizeko hirugarren eta hamaikagarren errenkadei dagozkie. Hori dela eta, kontsultari dagokion bektorea da $\mathbf{k}^{\mathrm{T}} = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0)$. Bektore honen proiekzioa planoan honela kalkulatuko dugu.

$$\mathbf{k}_2 = \mathbf{k}^T \mathbf{U}_2 \Sigma_2^{-1},$$

$\mathbf{k}_2$ bektoreak bi osagai ditu: $\mathbf{k}_2^T = (-0.0828, 0.2620)$, kontsultaren koordenatuak planoan. Marraz dezagun kontsulta grafikoan beste dokumentuekin batera. *k* etiketaz adierazita dator kontsultaren proiekzioa adierazten duen puntua.



Grafikan ikus daitekeenez, kontsulta gertuago dago «Grafo teoria»ri buruzko ml-m4 dokumentuetatik «Gizaki-konputagailu interakzioa»ri buruzko cl-c5 dokumentuetatik baino.

Hala ere, kontsultan ageri diren espazio semantikoko terminoak bi besterik ez dira: «graph» terminoa «grafo teoria»ri buruzko dokumentuekin erlazionatuta dagoena, eta «computer» terminoa, «gizaki-konputagailu interakzioa»ri buruzkoekin erlazionatutakoa. Hortaz, kontsultako terminoak aztertuta bakarrik, jatorrizko espazioan alegia, ez da nabarmen ikusten zein arlorekin dagoen erlazionatuta kontsulta. Aldiz planoan, espazio murriztuan, bai.

3. **ONDORIOAK**

Artikulu honetan ikusi dugunez, posible da testuen semantika modu matematikoan adieraztea, eta teknika matematiko batzuen bidez adierazpen hori eraldatzearen ondorioz, testaren eta bai bertako hitzen arteko erlazio semantikoak hobe neurtuko dira. Honek aplikazio interesgarri asko dira. Adibidez, hezkuntzaren arloan ikasleek egindako laburpenen ebaluazioan erabili da beste zenbait hizkuntzatan (laburpenaren eta jatorrizko testuaren arteko antzekotasuna neurtzeko). Informazioaren erauzketaren arloan ere tresna baliogarria da, eta lagun dezake erabiltzailearen interesekoak diren testuak aurkitzen testaren semantikan oinarrituta. Honek bilatzaile sintaktikoek dituzten zenbait arazo ekiditen ditu (sinonimia, polisemia).

ERREFERENTZIAK

[1] M.W. BERRY, S.T. DUMAIS, G.W. O'BRIEN (1994) «Using Linear Algebra for Intelligent Information Retrieval». *SIAM Review,* 37(4): 573-595.

[2] M.W. BERRY, Z. DRMAC, E.R. JESSUP (1999) «Matrices, vector spaces and Information Retrieval». *SIAM Review*, 41(2): 335-362.

[3] S. DEERWESTER, S.T. DUMAIS, G.W. FURNAS, T.K. LANDAUER, R. HARSHMAN (1990) «Indexing by Latent Semantic Analysis». *Journal of the Ameritan Society for Information Science* 41:391-407.

[4] J.M. GRACIA (2002) *Álgebra Lineal tras los Buscadores de Internet.* http://www.vc.ehu.es/campas/centros/formacia/deptosf/depme/gracia2.htm

[5] T.K. LANDAUER, D. LAHAM, P.W. FOLTZ (1998) «Introduction to Latent Semantic Analysis». *Discourse Processes*, 25: 259-284.

[6] T.K. LANDAUER, S.T. DUMAIS (1997) «A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction and Representation of Knowledge». *Psychological Review, 104: 211-240*

[7] C.D. **M**EYER (2000) ***M**atrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics.

[8] S. RÍOS (1994) *Iniciación Estadística*. Editorial Paraninfo.

[9] G. STRANG (1988) *Linear Algebra and its Applications*. Harcourt Brace Jovanovich.