

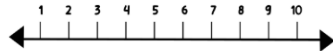
Representacions vectorials de text en problemes de classificació

Treball de Final de Grau

Toni Esteve Gené

Universitat Autònoma de Barcelona
Departament de Matemàtiques

Motivació



Continguts



- 1 Introducció a les representacions vectorials i TF-IDF
- 2 Latent Semantic Analysis
- 3 Latent Dirichlet Allocation
- 4 Metodologia
- 5 Annex

Introducció a les representacions vectorials



Objectiu: Convertir documents de text en vectors numèrics per fer-los tractables computacionalment.

Corpus de documents

Format per un conjunt de documents

$$D = \{d_1, d_2, \dots, d_n\}$$

Vocabulari

Conjunt de termes únics $\mathcal{V} = \bigcup_{d \in D} T_d$

Espai vectorial semàntic

Es una parella $(\mathbb{R}^{|\mathcal{V}|}, \phi)$ on ϕ es una funció que associa a cada $d_i \in D$ amb in vector $v_i \in \mathbb{R}^{|\mathcal{V}|}$.

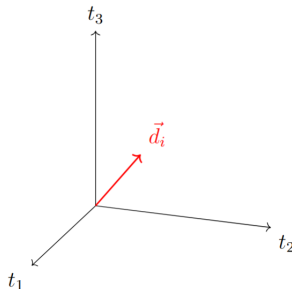


Figura 1: Representació vectorial d'un document en un vocabulari format per 3 termes.

TF-IDF



Objectiu: Mesurar la importància d'un terme dins un document tenint en compte la seva freqüència local i la seva raresa global.

Term Frequency (TF)

$$tf(t, d) = \frac{f(t, d)}{\max\{f(t', d) \mid t' \in d\}}$$

Inverse Document Frequency (IDF)

$$idf(t, D) = \log \left(\frac{|D|}{|\{d \in D \mid t \in d\}|} \right)$$

Term Frequency - Inverse Document Frequency (TF-IDF)

$$tfidf(t, d) = tf(t, d) \cdot idf(t, D)$$

A partir dels pesos *TF-IDF* s'obté la següent representació:

$$\phi(d_i) = (tfidf(t_1, d_i), \dots, tfidf(t_{|V|}, d_i))$$

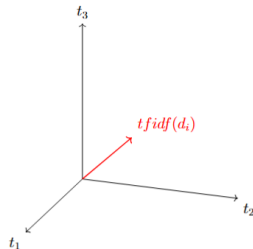
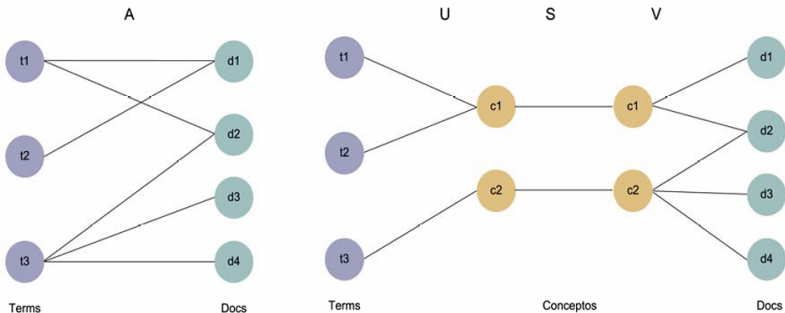


Figura 2: Representació vectorial d'un document en un vocabulari format per 3 termes.

Introducció per a Latent Semantic Analysis



Figura 3: Relacions conceptuais captades per LSA





Matriu d'entrada A

Donat un corpus de documents D i un vocabulari \mathcal{V} considerem

$$\begin{matrix}
 & d_1 & d_2 & \dots & d_j & \dots & d_n \\
 \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_i \\ \vdots \\ t_m \end{matrix} & \left[\begin{array}{cccccc}
 a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\
 a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2n} \\
 \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
 a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{in} \\
 \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
 a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn}
 \end{array} \right]
 \end{matrix}$$

Figura 4: Representació esquemàtica de la matriu d'entrada A , on les files corresponen als termes t_i i les columnes als documents d_j .



Descomposició en Valors Singlars

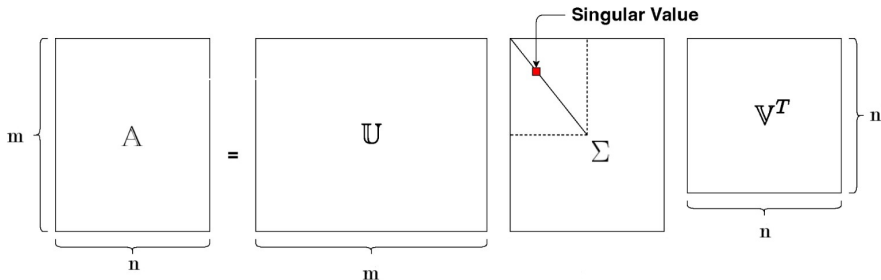
Teorema (Descomposició en Valors Singlars)

Sigui $A \in \mathbb{R}^{m \times n}$ Aleshores existeixen matrius:

$$A = U\Sigma V^T$$

amb:

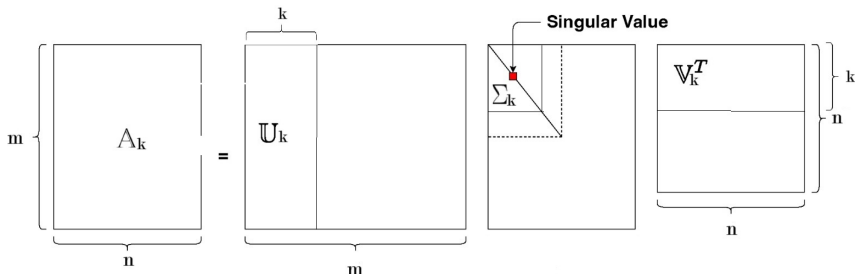
- $U \in \mathbb{R}^{m \times m}$ complint $UU^T = I_m$.
- $\Sigma \in \mathbb{R}^{m \times n}$ amb valors singulars $\sigma_1 \geq \dots \geq \sigma_r > 0$
- $V \in \mathbb{R}^{n \times n}$ complint $VV^T = I_n$.



Truncació de la Descomposició en Valors Singulars



Conservem només les k dimensions latents més informatives, eliminant soroll i redundàncies



Teorema (Eckart–Young–Mirsky):

$$\|A - A_k\|_F = \min_{\text{rank}(B) \leq k} \|A - B\|_F$$

Teoria de Perron-Frobenius

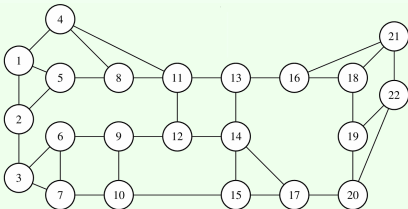


Matriu de coocurrència

Donada $A \in \mathbb{R}^{m \times n}$

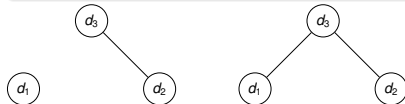
$$A^T A = \begin{bmatrix} \langle d_1, d_1 \rangle & \cdots & \langle d_1, d_j \rangle & \cdots & \langle d_1, d_n \rangle \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \langle d_i, d_1 \rangle & \cdots & \langle d_i, d_j \rangle & \cdots & \langle d_i, d_n \rangle \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \langle d_n, d_1 \rangle & \cdots & \langle d_n, d_j \rangle & \cdots & \langle d_n, d_n \rangle \end{bmatrix}$$

Graf dirigit associat a una matriu $\mathcal{G}(D)$



Matriu Reduïble i Irreduïble

Donada $D \in \mathbb{R}^{n \times n}$ diem que es irreduïble si $\mathcal{G}(D)$ és fortament connex (dreta). En cas contrari, diem que és reduïble (esquerra).



Teoria de Perron-Frobenius II

Teorema de Perron-Frobenius

Sigui $D \in \mathbb{R}^{n \times n}$ una matriu simètrica, no negativa i irreduïble \iff existeix un valor propi real i positiu $r = \rho(D)$ (radi espectral) tal que

$$D\mathbf{v} = r\mathbf{v} \quad \text{amb} \quad \mathbf{v} > 0$$

Contrarecíproc del Teorema de Perron-Frobenius

Si D és una matriu simètrica i reduïble \iff el vector propi dominant \mathbf{v}_1 conté coordenades nul·les o negatives.

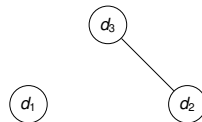
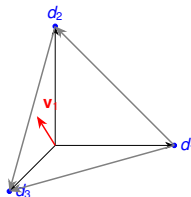
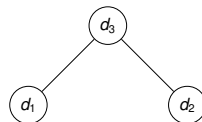
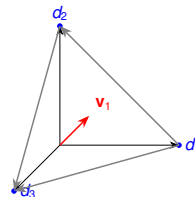


Figura 6: Representació del Teorema de Perron-Frobenius i el seu contrarecíproc.

Introducció a Latent Dirichlet Allocation

Latent Dirichlet allocation

(Blei, Ng, & Jordan, 2001; 2003)

Main difference: one
topic per word

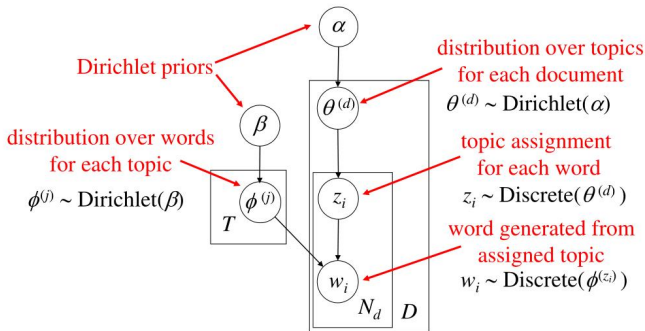


Figura 7: Flux del procés generatiu per a LDA

Procés generatriu probabilístic



Objectiu: LDA modela cada document com una barreja de temes latents, assumint que cada paraula prové d'un tema ocult.

Procés generatriu per a LDA

Fixem un nombre de temes K i un vocabulari \mathcal{V} . Per a cada tema k , es genera $\phi^{(k)} \sim \text{Dirichlet}(\beta)$: distribució de paraules dins del tema. Per a cada document $\mathbf{w} \in D$:

- ① Es tria la longitud $N \sim \text{Poisson}(\zeta)$.
- ② Es tria la distribució de temes $\theta \sim \text{Dirichlet}(\alpha)$.
- ③ Per a cada paraula w_n :
 - Es tria un tema latent $z_n \sim \text{Multinomial}(\theta)$.
 - Es tria una paraula $w_n \sim \text{Multinomial}(\phi^{(z_n)})$.



Problema inferencial

Objectiu inferencial: En LDA volem calcular $p(\theta, \mathbf{z}, \phi \mid \mathbf{w}, \alpha, \beta)$ a partir del text observat.

Fórmula de Bayes aplicada al model LDA

$$p(\theta, \mathbf{z}, \phi \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}, \phi \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

La expressió del denominador desenvolupada és intractable computacionalment:

td

El model LDA és generatiu, però el problema inferencial inverteix el procés: volem descobrir quins temes latents han generat el document observat.



Mostreig de Gibbs

Objectiu: Obtenir mostres d'un vector aleatori $x = (x_1, \dots, x_m)$ amb distribució conjunta $p(x_1, \dots, x_m)$, difícil de simular directament.

Algorisme del mostreig de Gibbs

Premisa: Coneixem les distribucions condicionals completes:

- $p(x_1 \mid x_2, \dots, x_m)$
- $p(x_2 \mid x_1, x_3, \dots, x_m)$, etc.

Passos de l'algorisme

- Inicialitzar cada x_i aleatòriament.
- Iterar per $t = 1, \dots, T$:
 - Mostrejar $x_1^{(t+1)} \sim p(x_1 \mid x_2^{(t)}, x_3^{(t)}, \dots, x_m^{(t)})$
 - Mostrejar $x_2^{(t+1)} \sim p(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \dots, x_m^{(t)})$
 - Continuar fins a x_m .



Estimació puntual de ϕ_k i θ_d

Un cop s'acaba el mostreig de Gibbs, s'estima puntualment:

- θ_d : Distribució de temes per document
- ϕ_k : Distribució de paraules per tema

Estimació de θ_d

Per a l'**estimació de θ_d** , es veu que

$\theta_d \mid \mathbf{z}_d, \alpha \sim \text{Dirichlet}(n_{d,1} + \alpha_1, n_{d,2} + \alpha_2, \dots, n_{d,K} + \alpha_K)$ i s'infereix

$$\hat{\theta}_{d,k} = \frac{n_{d,k} + \alpha_k}{\sum_j (n_{d,j} + \alpha_j)}$$

Estimació de la probabilitat que el document d tracti del tema k

Representació vectorial de longitud k donada per θ_d

$$\hat{\theta}_d = (\hat{\theta}_{d,1}, \dots, \hat{\theta}_{d,k})$$

Metodologia



Objectiu Analitzar l'impacte de les representacions vectorials TF-IDF, LSA i LDA de text en tasques de classificació binària.

S'usen diferents datasets:

- Dataset Spam (petit, vocabulari reduït).
- Twitter (gran volum, alta diversitat lèxica).
- 20 Newsgroups (temàtic, semànticament estructurat).

Resum de resultats

Spam dataset			
	auc-roc	t	k
TF-IDF	0.9826	53 s	10000
LSA	0.9818	10 s	100
LDA	0.9657	20 s	10

Figura 8: Gradient Boosting amb 500 arbres i profunditat 6.

20 Newsgroup dataset			
	auc-roc	t	k
TF-IDF	0.9915	650 s	10000
LSA	0.9902	120 s	100
LDA	0.9743	520 s	20

Figura 9: Gradient Boosting amb 500 arbres i profunditat 6

Twitter dataset			
	auc-roc	t	k
TF-IDF	0.8239	3 s	10000
LSA	0.7631	12 s	100
LDA	0.6040	100 s	10

Figura 10: Regressió logística

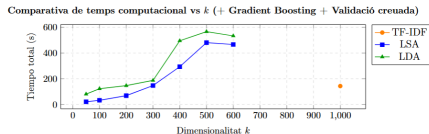









Figura 11: Comparativa del temps computacional segons la dimensionalitat k

Referències

-  Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *JMLR*, 3, 993–1022.
-  Casella, G., Berger, R. L. (2002). *Statistical Inference*. Duxbury.
-  Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
-  Blei, D. M., Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics*, 1(1), 17–35.
-  Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
-  Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240.
-  Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1), 188–230.

Gràcies per la vostra atenció!
Preguntes?

Model de perturbacions



- **Ideal**: matriu bloc-diagonal $D \rightarrow$ temes independents
- **Real**: correlacions espúries per termes ocasionals

Formulació espectral

$$A^T A = M = D + \varepsilon B$$

- D : estructura temàtica ideal (bloc-diagonal)
- εB : correlacions espúries

Exemple: dues temàtiques connectades



$$D = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad B = \begin{pmatrix} d & z \\ z & f \end{pmatrix}$$

$$M = D + \varepsilon B = \begin{pmatrix} a + \varepsilon d & \varepsilon z \\ \varepsilon z & b + \varepsilon f \end{pmatrix}$$

$$\lambda_{1,2} = \frac{\alpha + \beta}{2} \pm \frac{1}{2} \sqrt{(\alpha - \beta)^2 + 4\varepsilon^2 z^2}$$

Quan $\varepsilon > 0$:

- $\mathbf{v}_1 > 0$: correlació semàntica global
- \mathbf{v}_2 : separa documents en eixos semàntics latents

$$\mathbf{v}_1 = \frac{1}{N} \begin{pmatrix} 1 \\ \frac{\varepsilon z}{G} \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{N} \begin{pmatrix} -\frac{\varepsilon z}{G} \\ 1 \end{pmatrix}$$

Els eixos propis defineixen estructures latents separades.

Mostreig de Gibbs col·lapsat



Objectiu: Inferir les assignacions temàtiques z_i donat $\mathbf{w}, \alpha, \beta$.

Assignació de temes via mostreig de Gibbs

$$p(z_i = k \mid z_{-i}, \mathbf{w}, \alpha, \beta) \propto \underbrace{\frac{n_{d,k}^{(-i)} + \alpha_k}{\sum_j (n_{d,j}^{(-i)} + \alpha_j)}}_{\text{Proporció de tema al document}} \cdot \underbrace{\frac{n_{k,w}^{(-i)} + \beta_w}{\sum_{w'} (n_{k,w'}^{(-i)} + \beta_{w'})}}_{\text{Probabilitat de paraula al tema}}$$

- $n_{d,k}^{(-i)}$: Nombre de paraules del document d assignades al tema k (sense comptar i).
- $n_{k,w}^{(-i)}$: Nombre de vegades que la paraula w apareix assignada al tema k (sense comptar i).
- Es mostreja un nou valor per z_i segons aquesta distribució.



Ajust dels hiperparàmetres en LDA

Després del mostreig de Gibbs col·lapsat, s'ajusten els hiperparàmetres:

- α : distribució de temes per document
- β : distribució de paraules per tema

Log-versemblança marginal:

$$p(\mathbf{z} \mid \alpha) = \prod_{d=1}^D \log \frac{B(n_{d,\cdot} + \alpha)}{B(\alpha)}$$

Derivant la ecuació anterior respecte α_k i igualant a 0 s'obté el següent resultat

$$\begin{aligned} \frac{\partial}{\partial \alpha_k} \log p(\mathbf{z} \mid \alpha) &= \sum_d \left[\psi(n_{d,k} + \alpha_k) - \psi \left(\sum_j (n_{d,j} + \alpha_j) \right) \right] \\ &\quad - D \left[\psi(\alpha_k) - \psi \left(\sum_j \alpha_j \right) \right] = 0 \end{aligned}$$



Estimació puntual de ϕ_k i θ_d

Un cop s'acaba el mostreig de Gibbs, s'estima puntualment:

- θ_d : Distribució de temes per document
- ϕ_k : Distribució de paraules per tema

Estimació de θ_d

Per a l'**estimació de** θ_d , es veu que

$\theta_d \mid \mathbf{z}_d, \alpha \sim \text{Dirichlet}(n_{d,1} + \alpha_1, n_{d,2} + \alpha_2, \dots, n_{d,K} + \alpha_K)$ i s'infereix

$$\hat{\theta}_{d,k} = \frac{n_{d,k} + \alpha_k}{\sum_j (n_{d,j} + \alpha_j)}$$

Estimació de la probabilitat que el document d tracti del tema k

Estimació de ϕ_k

Per a l'**estimació de** ϕ_k , es veu que $\phi_k \sim \text{Dirichlet}(\gamma_1, \gamma_2, \dots, \gamma_V)$ i s'infereix

$$\hat{\phi}_{k,w} = \frac{n_{k,w} + \beta_w}{\sum_{w'} (n_{k,w'} + \beta_{w'})}$$

Estimació la probabilitat que el tema k generi la paraula w