Topic model et Réduction de dimension

Text Mining

Ricco Rakotomalala

La dimensionnalité est un problème en fouille de textes. Dans ce support nous étudions des techniques qui permettent de représenter les documents dans un espace intermédiaire préservant la proximité entre eux.

Topic Model. L'espace correspond à un ensemble de « topics » (thèmes) définis par les termes avec des poids ± élevés (soft/fuzzy clustering), et qui permettent de décrire les documents dans un nouvel espace de représentation. Les documents peuvent être associés à des divers degrés à des topics (ex. un ouvrage de machine learning sous Python).

Cet espace permet de produire un résumé propice à une meilleure compréhension de la nature des informations disponibles, notamment à travers les outils de visualisation.

La réduction de la dimensionnalité permet aussi une mise en œuvre efficace des techniques de data mining par la suite, dans l'espace de représentation réduit.

Description initiale des données. Matrice documents termes, p est souvent très élevé (plusieurs milliers).

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	...	Term p
Doc.1										
Doc.2										
Doc.n										

Permet de comprendre la nature des topics.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	...	Term p
Topic 1										
Topic 2										
Topic 3										
...										
Topic K										

Les valeurs caractérisent l'association. Il peut s'agit de coefficients d'une combinaison linéaire, des probabilités,… Un terme peut être associé à divers degrés à différents topics.

	Topic 1	Topic 2	Topic 3	...	Topic K
Doc.1					
Doc.2					
Doc.n					

Description des documents dans l'espace des topics. Avantageux si (1) K << p ; (2) on peut associer une sémantique aux topics.

1. Latent semantic indexing (LSI)
2. Analyse factorielle des correspondances (AFC)
3. Latent dirichlet allocation (LDA)
4. Bibliographique

Latent semantic analysis (LSA)

Analyse sémantique latente – Indexation sémantique latente

LATENT SEMANTIC INDEXING (LSI)

Le LSI est une technique qui élabore un espace de représentation synthétique préservant au mieux les propriétés des données, en particulier les distances entre les termes.

En réalité il s'agit simplement d'une technique factorielle équivalente à l'ACP (analyse en composantes principales) où les variables ne sont ni réduites, ni centrées. Avec les mêmes objectifs et les mêmes outils pour évaluer la qualité de représentation.

On dispose dès lors des outils d'interprétation usuels : qualité de représentation des termes et des documents sur les facteurs ; contribution des termes et des documents aux facteurs.

Corpus de 3 documents (Grossman, page 71):

D1 : "shipment of gold damaged in a fire"

D2 : "delivery of silver arrived in a silver truck"

D3 : "shipment of gold arrived in a truck"

Matrice termes-documents **M**

(pondération Term-frequency)

	D1	D2	D3
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

$p = 8$

$n = 3$

Après retrait des stop words.

Principe de la décomposition en valeurs singulières (*singular value decomposition* - ***SVD***)

$$M = U\Delta V^T \quad \text{avec} \quad \begin{cases} Mv_k = \delta_k u_k \\ M^T u_k = \delta_k v_k \end{cases}$$

Nombre max. de facteurs

$$\mathrm{H} = \min(p, n)$$

Pour exprimer la fidélité de représentation d'un facteur $\mathrm{F_k}$, nous calculons l'équivalent de la valeur propre d'une ACP

$$\lambda_k = \frac{\delta_k^2}{p-1}$$

$$\lambda_k \, (\mathrm{k} = 1,\ldots,3) = (1.175, 0.740, 0.228)$$

⇨ En pourcentage cumulé d'information $(54.81, 89.36, 100.0)$

LSI – Matrices intermédiaires (K = 2)

M

arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

=

U

-0.465	0.020	-0.369
-0.084	-0.330	0.487
-0.289	0.215	0.217
-0.084	-0.330	0.487
-0.260	-0.525	-0.098
-0.260	-0.525	-0.098
-0.578	0.429	0.433
-0.465	0.020	-0.369

Coordonnées des termes dans le nouvel espace de représentation

Δ

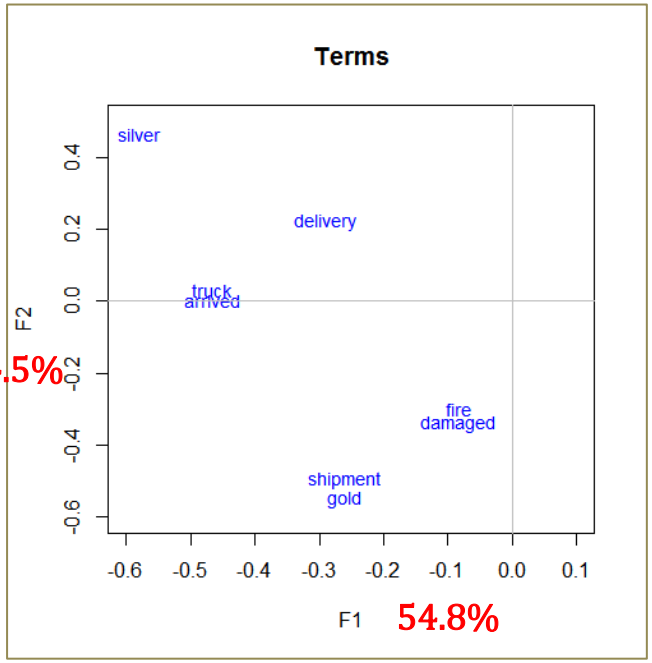
2.867347		
	2.276482	
		1.26331

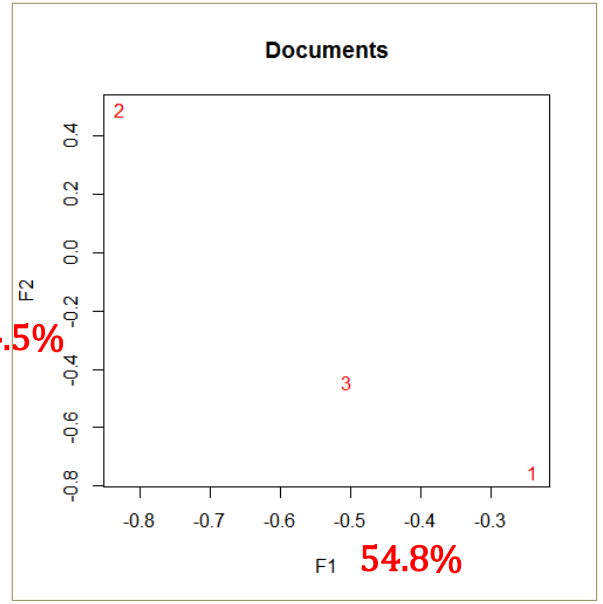
Indicateurs de qualité de représentation

V^T

-0.240	-0.828	-0.506
-0.751	0.489	-0.444
0.615	0.274	-0.739

Coordonnées des documents dans le nouvel espace de représentation





$U =$

-0.465	0.020	-0.369
-0.084	-0.330	0.487
-0.289	0.215	0.217
-0.084	-0.330	0.487
-0.260	-0.525	-0.098
-0.260	-0.525	-0.098
-0.578	0.429	0.433
-0.465	0.020	-0.369

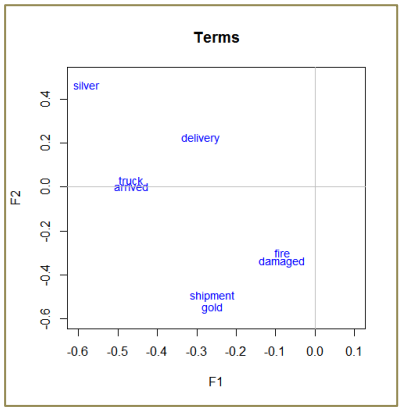
La somme des carrés des valeurs en colonne est égale à 1

$V^T =$

-0.240	-0.828	-0.506
-0.751	0.489	-0.444
0.615	0.274	-0.739

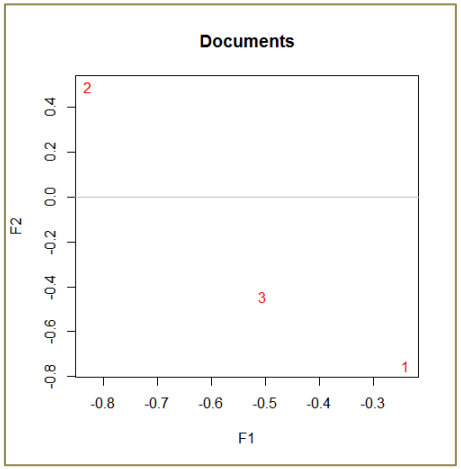
La somme des carrés des valeurs en ligne est égale à 1

Il est dès lors facile d'approfondir l'interprétation sous l'angle de la qualité de représentation (COS^2) et des contributions (CTR) des termes et des documents.



	F1	F2
arrived	0.22	0.00
damaged	0.01	0.11
delivery	0.08	0.05
fire	0.01	0.11
gold	0.07	0.28
shipment	0.07	0.28
silver	0.33	0.18
truck	0.22	0.00

Même commentaire pour les documents.



	D1	D2	D3
F1	0.06	0.69	0.26
F2	0.56	0.24	0.20

Plus le terme est loin de l'origine, plus il contribue.

	D1	D2	D3
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

Comme en ACP, il est possible de positionner un document supplémentaire (ex. un document requête en recherche d'information).

	q
arrived	0
damaged	0
delivery	0
fire	0
gold	1
shipment	0
silver	1
truck	1

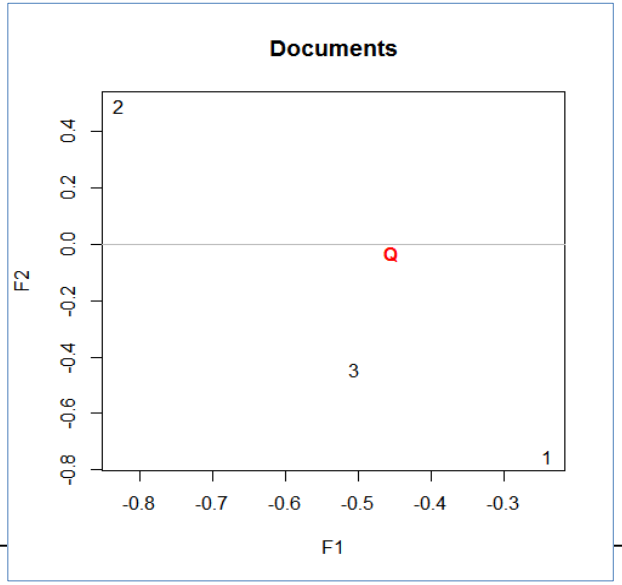
Q : « gold silver truck »

Pour une projection du document supplémentaire dans le plan (K = 2):

$$F^* = q^T U_K (\Delta_K)^{-1}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} -0.465 & 0.020 \\ -0.084 & -0.330 \\ -0.289 & 0.215 \\ -0.084 & -0.330 \\ -0.260 & -0.525 \\ -0.260 & -0.525 \\ -0.578 & 0.429 \\ -0.465 & 0.020 \end{bmatrix} \begin{bmatrix} 2.867347 & 0 \\ 0 & 2.276482 \end{bmatrix}^{-1} = \begin{bmatrix} -0.455 & -0.033 \end{bmatrix}$$

On positionne un document à partir des termes qui le compose



Il est possible d'approximer le tableau de données initial dans l'espace de représentation réduit (*la reconstitution est exacte si on prend les H facteurs*). Cela peut donner une indication sur la qualité des K facteurs sélectionnés.

$$M_K = U_K \Delta_K {V_K}^T$$

-0.465	0.020		2.867347	0		-0.240	-0.828	-0.506
-0.084	-0.330		0	2.276482		-0.751	0.489	-0.444
-0.289	0.215							
-0.084	-0.330							
-0.260	-0.525							
-0.260	-0.525							
-0.578	0.429							
-0.465	0.020							

	D1	D2	D3
arrived	0.3	1.1	0.7
damaged	0.6	-0.2	0.5
delivery	-0.2	0.9	0.2
fire	0.6	-0.2	0.5
gold	1.1	0.0	0.9
shipment	1.1	0.0	0.9
silver	-0.3	1.9	0.4
truck	0.3	1.1	0.7

	D1	D2	D3
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

La SVD peut être vue comme un système de compression des données avec pertes (la reconstitution est approximative, mais la qualité peut être modulée).

Le gain en espace de stockage n'est intéressant que si

$$K \ll \min(p, n)$$

LSI repose sur la décomposition en valeurs singulières de la matrice termes documents (valable quelle que soit la pondération utilisée). « Facteur = Topic » ici.

Le calcul peut être long, des stratégies existe pour accélérer les calculs lorsque seuls les K premiers facteurs sont demandés.

Elle est équivalente à une ACP (analyse en composantes principales) où les variables ne sont ni centrées, ni réduites (cf. *prcomp() vs. princomp() sous R*). La lecture des résultats et les aides à l'interprétation sont transposables dans ce contexte. Il est également possible de positionner les documents supplémentaires dans le repère factoriel.

LSI = Topic Modeling. Les topics (facteurs) sont définis par les termes à divers degrés (ce que l'on peut voir aussi comme un clustering flou), il est possible de positionner les documents dans le nouvel espace de représentation.

Analyse à partir du tableau de contingence « termes – documents »

ANALYSE FACTORIELLE DES CORRESPONDANCES

L'AFC peut s'appliquer à tout tableau croisé de valeurs positives ou nulles dès lors que les notions de marge et de profils ont un sens. Individu statistique = occurrence d'un terme dans un document (Lebart et Salem, Chapitre 3).

C'est le cas pour la matrice termes documents, en particulier pour les pondérations booléennes et fréquences.

	D1	D2	D3	Somme
arrived	0	1	1	2
damaged	1	0	0	1
delivery	0	1	0	1
fire	1	0	0	1
gold	1	0	1	2
shipment	1	0	1	2
silver	0	2	0	2
truck	0	1	1	2
Somme	4	5	4	13

Nombre d'apparition du terme dans l'ensemble des documents

Nombre de termes composant un document

Profils lignes

	D1	D2	D3	Somme
arrived	0.00	0.50	0.50	1.00
damaged	1.00	0.00	0.00	1.00
delivery	0.00	1.00	0.00	1.00
fire	1.00	0.00	0.00	1.00
gold	0.50	0.00	0.50	1.00
shipment	0.50	0.00	0.50	1.00
silver	0.00	1.00	0.00	1.00
truck	0.00	0.50	0.50	1.00
Somme	0.31	0.38	0.31	1.00

P(terme / document)

Profils colonnes

	D1	D2	D3	Somme
arrived	0.00	0.20	0.25	0.15
damaged	0.25	0.00	0.00	0.08
delivery	0.00	0.20	0.00	0.08
fire	0.25	0.00	0.00	0.08
gold	0.25	0.00	0.25	0.15
shipment	0.25	0.00	0.25	0.15
silver	0.00	0.40	0.00	0.15
truck	0.00	0.20	0.25	0.15
Somme	1.00	1.00	1.00	1.00

P(document / terme)

Y / X	x_1	x_l	x_L	Σ
y_1				
		$\vdots$		
y_k		n_{kl}		$n_{k.}$
		$\vdots$		
y_K				
Σ		$n_{.l}$		n

K termes, L documents

n_{kl} nombre d'apparition du terme k dans le doc. l

$n_{k.}$: # du terme k dans l'ensemble des documents

$n_{.l}$: # de termes dans le document l

n : nombre total de couples « termes – documents »

Distance du KHI² (exacerbe le rôle des modalités rares)

Distance entre profils lignes

(entre les termes)

$$d^2(k,k') = \sum_{l=1}^{L} \frac{n}{n_{.l}} \left(\frac{n_{kl}}{n_{k.}} - \frac{n_{k'l}}{n_{k'.}} \right)^2$$

$$d^2(shipment, gold) = \frac{1}{0.31}(0.5-0.5)^2 + \frac{1}{0.38}(0.0-0.0)^2 + \frac{1}{0.31}(0.5-0.5)^2 = 0.0$$

$$d^2(shipment, silver) = \frac{1}{0.31}(0.5-0.0)^2 + \frac{1}{0.38}(0.0-1.0)^2 + \frac{1}{0.31}(0.5-0.0)^2 = 4.2$$

Distance entre profils colonnes

(entre les documents)

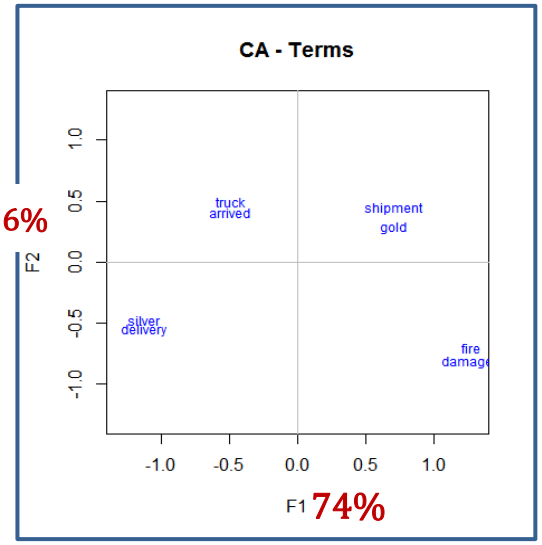
$$d^2(l,l') = \sum_{k=1}^{K} \frac{n}{n_{k.}} \left(\frac{n_{kl}}{n_{.l}} - \frac{n_{kl'}}{n_{.l'}} \right)^2$$

$$d^2(D1, D2) = \frac{1}{0.15}(0.0-0.2)^2 + \cdots + \frac{1}{0.15}(0.0-0.2)^2 = 4.5$$

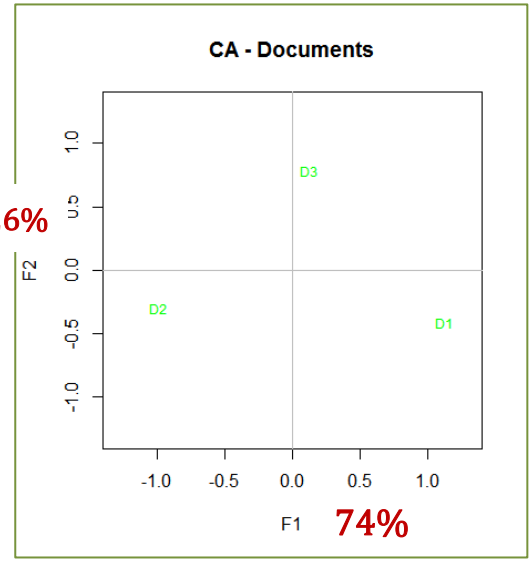
$$d^2(D1, D3) = \frac{1}{0.15}(0.0-0.25)^2 + \cdots + \frac{1}{0.15}(0.0-0.25)^2 = 2.4$$

Un des objectifs de l'AFC est de positionner les modalités lignes et colonnes dans un repère factoriel en se basant sur les profils.

Profils lignes : positionnement relatif des termes



Profils colonnes : positionnement relatif des documents



Nous disposons des outils usuels d'interprétation et d'évaluation de l'AFC.

Détection du nombre adéquat de facteurs

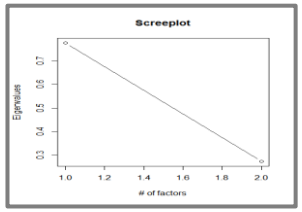
Pourcentage d'inertie expliquée

F	eigenvalue	Percentage of variance	Cumulative perc. of var.
F1	0.776	73.928	73.928
F2	0.274	26.072	100

Nombre max de facteurs

$$H = \min(K - 1, L - 1)$$

Graphique screeplot



Pas très parlant sur cet exemple précis.

Qualité de représentation (COS^2) et contribution aux facteurs (CTR)

Profils lignes

Characterization				Coord.		Contributions		COS			
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos 1		cos 2	
silver	0.154	1.600	0.246	-1.130	-0.568	25.32	18.12	0.80	(0.80)	0.20	(1.00)
fire	0.077	2.250	0.173	1.279	-0.784	16.20	17.28	0.73	(0.73)	0.27	(1.00)
damaged	0.077	2.250	0.173	1.279	-0.784	16.20	17.28	0.73	(0.73)	0.27	(1.00)
delivery	0.077	1.600	0.123	-1.130	-0.568	12.66	9.06	0.80	(0.80)	0.20	(1.00)
arrived	0.154	0.463	0.071	-0.498	0.463	4.92	12.05	0.54	(0.54)	0.46	(1.00)
truck	0.154	0.463	0.071	-0.498	0.463	4.92	12.05	0.54	(0.54)	0.46	(1.00)
gold	0.154	0.625	0.096	0.706	0.355	9.89	7.08	0.80	(0.80)	0.20	(1.00)
shipment	0.154	0.625	0.096	0.706	0.355	9.89	7.08	0.80	(0.80)	0.20	(1.00)

Profils colonnes

Characterization				Coord.		Contributions		COS			
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos 1		cos 2	
D1	0.308	1.438	0.442	1.127	-0.410	50.31	18.92	0.88	(0.88)	0.12	(1.00)
D2	0.385	1.080	0.415	-0.996	-0.297	49.14	12.40	0.92	(0.92)	0.08	(1.00)
D3	0.308	0.625	0.192	0.118	0.782	0.55	68.68	0.02	(0.02)	0.98	(1.00)

Permet d'analyser finement le rôle des modalités lignes (termes) et colonnes (documents).

Terms	D1	D2	D3	Somme
arrived	0	1	1	2
damaged	1	0	0	1
delivery	0	1	0	1
fire	1	0	0	1
gold	1	0	1	2
shipment	1	0	1	2
silver	0	2	0	2
truck	0	1	1	2
Somme	4	5	4	13

On peut mesurer l'association via la statistique du KHI-2 d'écart à l'indépendance.

$$\chi^2 = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(n_{kl}-e_{kl})^2}{e_{kl}} = 13.65 \quad \Rightarrow \quad \phi^2 = \frac{\chi^2}{n} = \frac{13.65}{13} = 1.05$$

Inertie totale

La matrice R des résidus standardisés permet de situer les attractions et répulsions entre les termes et les documents.

Résidus standardisés

Terms	D1	D2	D3
arrived	-0.784	0.263	0.490
damaged	1.248	-0.620	-0.555
delivery	-0.555	0.992	-0.555
fire	1.248	-0.620	-0.555
gold	0.490	-0.877	0.490
shipment	0.490	-0.877	0.490
silver	-0.784	1.403	-0.784
truck	-0.784	0.263	0.490

$$r_{kl} = \frac{n_{kl}-e_{kl}}{\sqrt{e_{kl}}}$$

La contribution au KHI-2 permet de mesurer l'impact des associations dans la quantité d'information globale.

Contributions au KHI-2

Terms	D1	D2	D3
arrived	4.508	0.507	1.761
damaged	11.412	2.818	2.254
delivery	2.254	7.213	2.254
fire	11.412	2.818	2.254
gold	1.761	5.635	1.761
shipment	1.761	5.635	1.761
silver	4.508	14.427	4.508
truck	4.508	0.507	1.761

$$c_{kl} = 100 \times \frac{r_{kl}^2}{\chi^2}$$

Une grande partie de l'information vient des attractions (D2, silver) et (D1, [damaged, fire]).

La représentation simultanée est possible grâce aux relations de transition (relations quasi-barycentriques) c.-à-d. il est possible d'obtenir les coordonnées d'une modalité colonne à partir des coordonnées de l'ensemble des modalités lignes, et inversement.

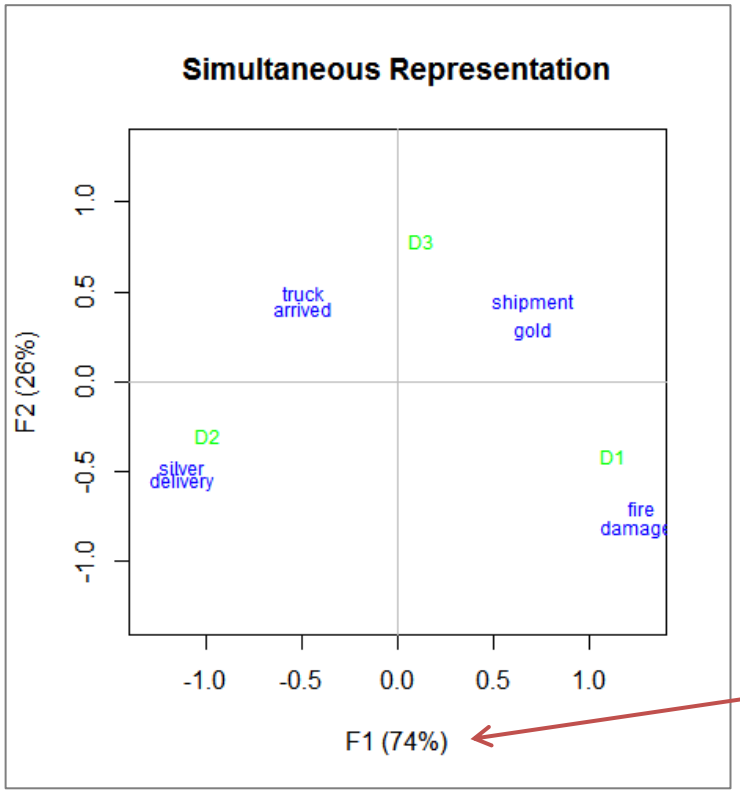
Coordonnée de la modalité ligne k sur le facteur 1

Valeur du profil $P(\text{Col. } l \,/\, \text{Ligne } k)$

$$F_{k1} = \frac{1}{\sqrt{\lambda_1}} \sum_{l=1}^{L} \frac{n_{kl}}{n_{k.}} \times G_{l1}$$

Coordonnée de la modalité colonne l sur le facteur 1

! La coordonnée de la modalité ligne k est une moyenne pondérée de l'ensemble des coordonnées des modalités colonnes.



$$G_{l1} = \frac{1}{\sqrt{\lambda_1}} \sum_{k=1}^{K} \frac{n_{kl}}{n_{.l}} \times F_{k1}$$

Le facteur 1 détermine en grande partie la lecture.

Attention, il faut situer une modalité ligne par rapport à l'ensemble des colonnes (et inversement). Ex. D1 est associé à (damaged, fire), mais on ne peut pas conclure que D3 est lié à (truck)... Cf. tableau des contributions au KHI-2.

Soit $M = \frac{1}{\sqrt{n}} R$ où R est la matrice des résidus standardisés

L'AFC consiste à calculer la décomposition en valeurs singulières de M

$$\mathbf{M} = \mathbf{U}\,\Delta\,\mathbf{V}^{\mathrm{T}}$$

U (K x K) contient les K vecteurs singuliers à gauche (modalités lignes). U est orthonormée.
Δ (K x L) est une matrice dont les éléments situés sur la diagonale correspondent aux valeurs singulières. Montés au carré, nous avons les valeurs propres.
V (L x L) contient les L vecteurs singuliers à droite (modalités colonnes). V est orthonormée.

A quoi correspond cette opération ?

(On perçoit bien l'analyse croisée ligne/colonne ici)

Une valeur singulière δ_h est telle que

$$M\,\vec{v}_h = \delta_h\,\vec{u}_h$$
$$M^T\vec{u}_h = \delta_h\,\vec{v}_h$$

Concrètement ?

1. On cherche à produire des vecteurs de projections de manière à ce que la dispersion des modalités lignes (colonnes) soit la plus grande possible sur l'axe.
2. La dispersion doit être la même pour les modalités lignes et les modalités colonnes.
3. Les facteurs sont orthogonaux deux à deux.

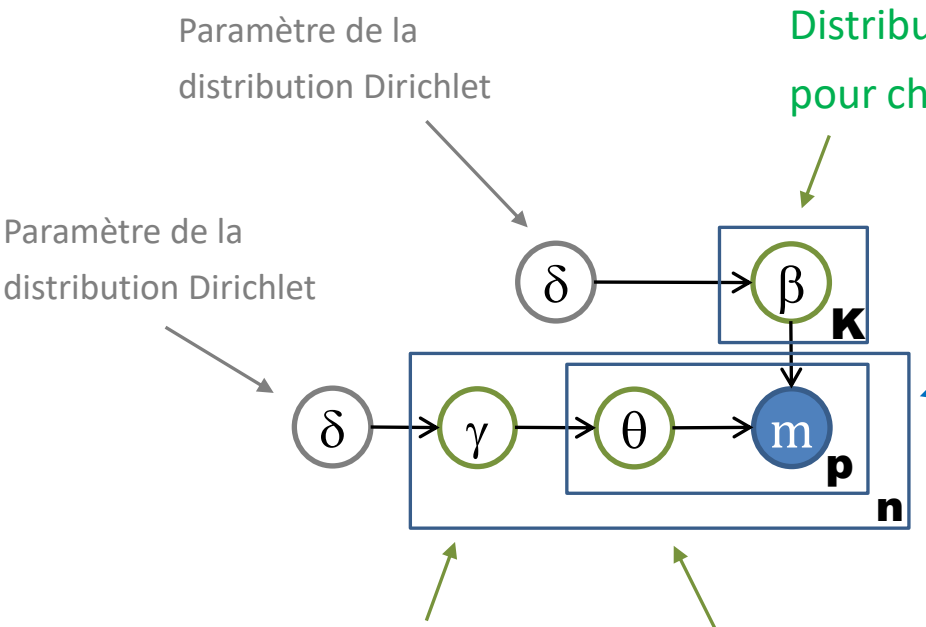
- Dans cette partie, nous avons considéré la matrice termes documents sous l'angle d'un tableau de contingence.
- La quantité d'information disponible est quantifiée par l'inertie totale $\phi^2 = \frac{\chi^2}{n}$; où χ^2 est le KHI-2 d'écart à l'indépendance de Pearson.
- L'AFC permet de positionner les termes entre eux (en fonction des documents qui les contiennent) et les documents entres eux (en fonction des termes qu'ils contiennent). Ici, à l'instar de la LSI, « Facteur = Topic ».
- Elle permet aussi de situer les associations termes – documents.
- Du point de vue des calculs, l'AFC repose sur la décomposition en valeurs singulières de la matrice des résidus standardisés du KHI-2.
- A partir de son profil colonne [P(terme/document)] et de la relation de transition (coordonnées des termes sur les facteurs), il est possible de positionner un document supplémentaire dans le repère factoriel.

Allocation de Dirichlet Latente

LATENT DIRICHLET ALLOCATION (LDA)

Modèle probabiliste génératif : modéliser le processus de génération des données c.-à-d. des paires documents-termes à l'aide de facteurs latents (sous-jacents). Modèle de mélange.

K : nombre de topics
n : # de documents
p : # de termes



Le même terme dans deux documents différents peut être associé à des topics différents.

!

La modélisation va nous fournir les éléments en vert. On a un changement de représentation, intéressant si K nombre de topics << p nombre de termes.

Hypothèse de travail simplificatrice

Les topics ne sont pas censés être corrélés entre eux.

Sélectionner le topic (k) pour le terme (j)

$$\varphi_{kj}$$

Distribution des termes (j) pour chaque topic (k) : distribution de Dirichlet symétrique de paramètre δ

$$\beta_k = \frac{\Gamma(p\delta)}{[\Gamma(\delta)]^p} \prod_{j=1}^{p} \varphi_{kj}^{\delta-1}$$

Sélectionner le topic pour chaque couple terme document

$$\theta_{ik}$$

Distribution des topics pour chaque document (Dirichlet)

$$\gamma_i = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1}$$

L'estimation des paramètres de la LDA passe par l'estimation des distributions des variables latentes à partir des données observées (*posterior inference*). On peut le voir sous l'angle de la maximisation de la log-vraisemblance. Nous passons par des heuristiques.

Gibbs sampling est une méthode de Monte-Carlo. Elle commence par assigner aléatoirement les topics puis, sur des échantillons, calcule les distributions conditionnelles et assigne les topics aux termes selon une certaine probabilité. On recommence un grand nombre de fois pour obtenir une bonne approximation des distributions.

Algorithme EM (espérance-maximisation), un algorithme itératif comprenant deux phases : espérance (E), calcul de l'espérance de la vraisemblance à valeurs des paramètres fixés ; maximisation (M) : calcul des paramètres maximisant la vraisemblance obtenue à l'étape E. On répète jusqu'à convergence.

Problèmes de temps de calcul et/ou d'espace mémoire pour les grandes volumétries.

Voir la doc. du package « topicmodels » de R.

On travaille sur la matrice documents-termes cette fois-ci

Termes = {arrived,…, truck}

Documents = {D1, D2, D3}

K = 2

M =

	arrived	damaged	delivery	fire	gold	shipment	silver	truck
D1	0	1	0	1	1	1	0	0
D2	1	0	1	0	0	0	2	1
D3	1	0	0	0	1	1	0	1

β

	P(terme/topic)							
	arrived	damaged	delivery	fire	gold	shipment	silver	truck
Topic 1	0.161	0.064	0.109	0.137	0.080	0.129	0.194	0.127
Topic 2	0.147	0.089	0.045	0.017	0.228	0.179	0.114	0.181

Topic 1 est avant tout déterminé par les termes « arrived » et « silver », Topic 2 par les termes « gold » et « truck ».

	P(topic/document)	
	Topic 1	Topic 2
D1	0.4999	0.5001
D2	0.5038	0.4962
D3	0.4963	0.5037

γ

Pas très convaincant sur cet exemple. Mais on se rend compte surtout que les documents sont placés dans un nouvel espace de représentation, celui des topics.

On peut appliquer des algorithmes de data mining par la suite.

Assignation des termes aux topics selon les documents

θ

	arrived	damaged	delivery	fire	gold	shipment	silver	truck
D1	0	2	0	1	2	2	0	0
D2	1	0	1	0	0	0	1	2
D3	1	0	0	0	2	2	0	2

0 = pas d'assignation

Pas de « surprises » ici, les termes sont associés aux mêmes topics, quels que soient les documents.

La LDA permet de mettre en évidence un ensemble de « topics » sous-jacents qui régissent un ensemble de documents.

Les topics sont décrits dans l'espace des termes. Les documents peuvent être décrits dans l'espace des topics.

Il existe un mécanisme pour la projection des documents supplémentaires dans l'espace des topics (puisque nous disposons de la description des topics dans l'espace des termes).

Le choix du nombre de topics (K) reste un problème ouvert (ex. graphique de décroissance de la déviance en fonction du nombre de topics).

Ouvrages

Aggarwal C., Zhai C., « Mining Text Data », Springer, 2012.

Grossman D.A., Frieder O. « Information retrieval – Algorithms and heuristics », Second Edition, Springer, 2004.

Lebart L., Salem A., « Statistique textuelle », Dunod, 1994.

Autres références

Blei D., Lafferty J., « Topic Models », in *Text Mining: Classification, clustering and applications*, A. Srivastava& M. Sahami, editors, Chapman & Hall , 2009.

Grun B., Hornik K., « topicmodels: An R Package for Fitting Topic Models », in Journal of Statistical Software, 40(13): 1-30, 2011.

Chen E., « Introduction to Latent Dirichlet Allocation », 2011.