

Topic Model

Modèle par thème

Enseignant : Sonia Gharsalli

- La **dimensionnalité** est un problème en fouille de textes.

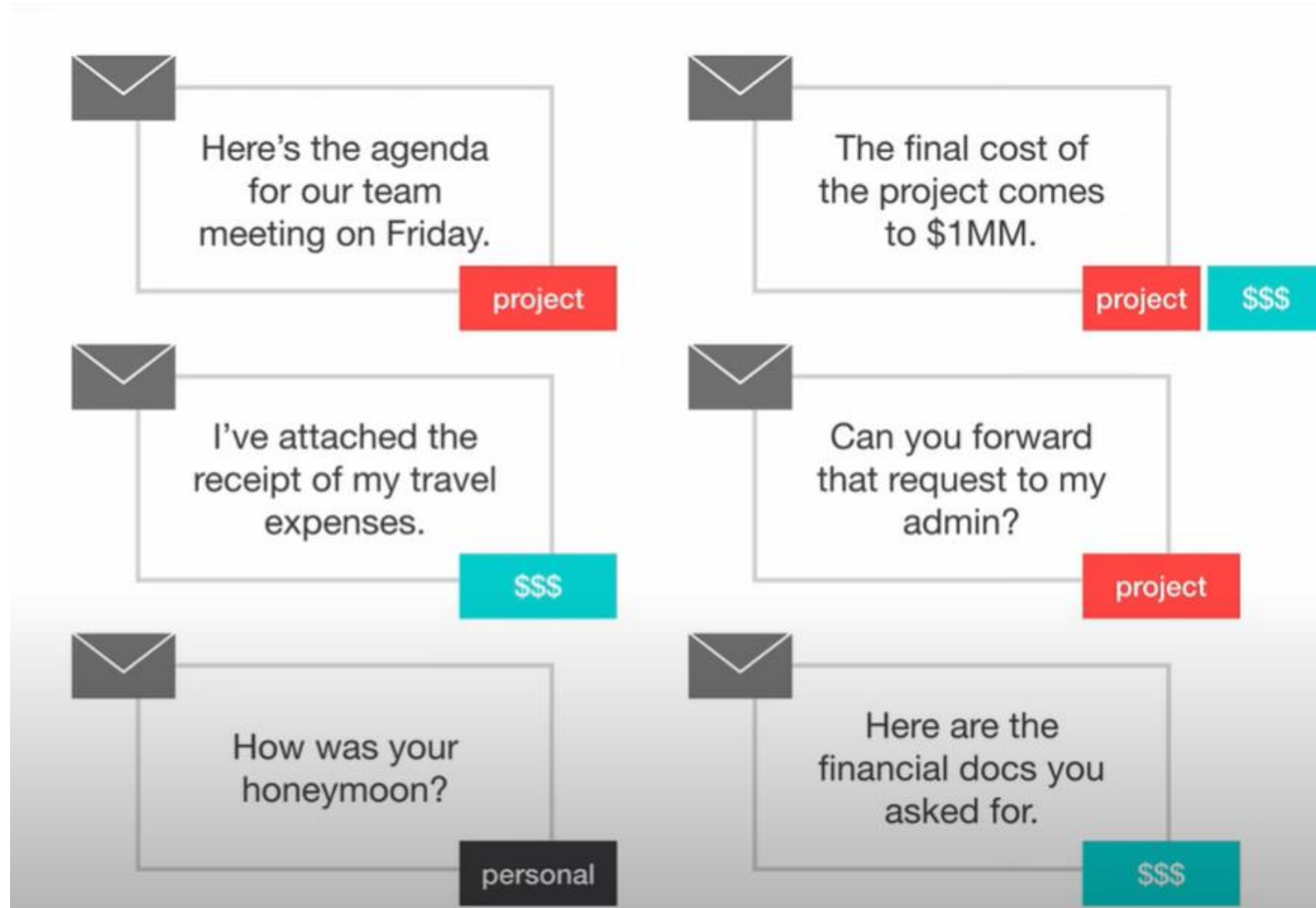
=> La représentation des documents dans un espace intermédiaire préservant la proximité entre eux peut être une des solutions permettant de réduire les dimensions.

- **Topic Model** : L'espace correspond à un ensemble de « topics » (thèmes) définis par les termes avec des poids élevés, 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)

Topic Model: exemple

- Indexation des mails



Topic Model: exemple

- Analyse des documents : Recherche des Topics de chaque document



Topic Model: example

Topic A: 40% banana, 30% kale, 10% breakfast...

Food

Topic B: 30% kitten, 20% puppy, 10% frog, 5% cute...

Animals

Topic Model



Matrice documents termes, Le nombre de termes **p** est souvent très élevé

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



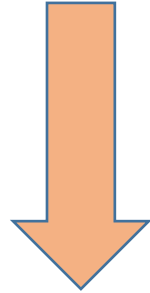
Changement vers un espace de thèmes

	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										

Un terme peut appartenir à plusieurs thèmes. Mais, avec des degrés différents.

=> Un Topic est une combinaison linéaire de plusieurs termes avec des probabilités différentes.

	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										



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

Description des documents dans l'espace des topics.

On a un réel avantage si $K \ll p$;

=> On peut associer une sémantique aux topics.

LATENT SEMANTIC INDEXING (LSI)

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.
- C'est 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.

=> Les outils d'interprétation sont : qualité de représentation des termes et des documents sur les facteurs ; contribution des termes et des documents aux facteurs

LSI: Calcul Matriciel

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

	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 en décomposition en valeurs singulières SVD

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

- Pour exprimer la fidélité de représentation d'un facteur F_k , nous calculons l'équivalent de la valeur propre d'une ACP

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


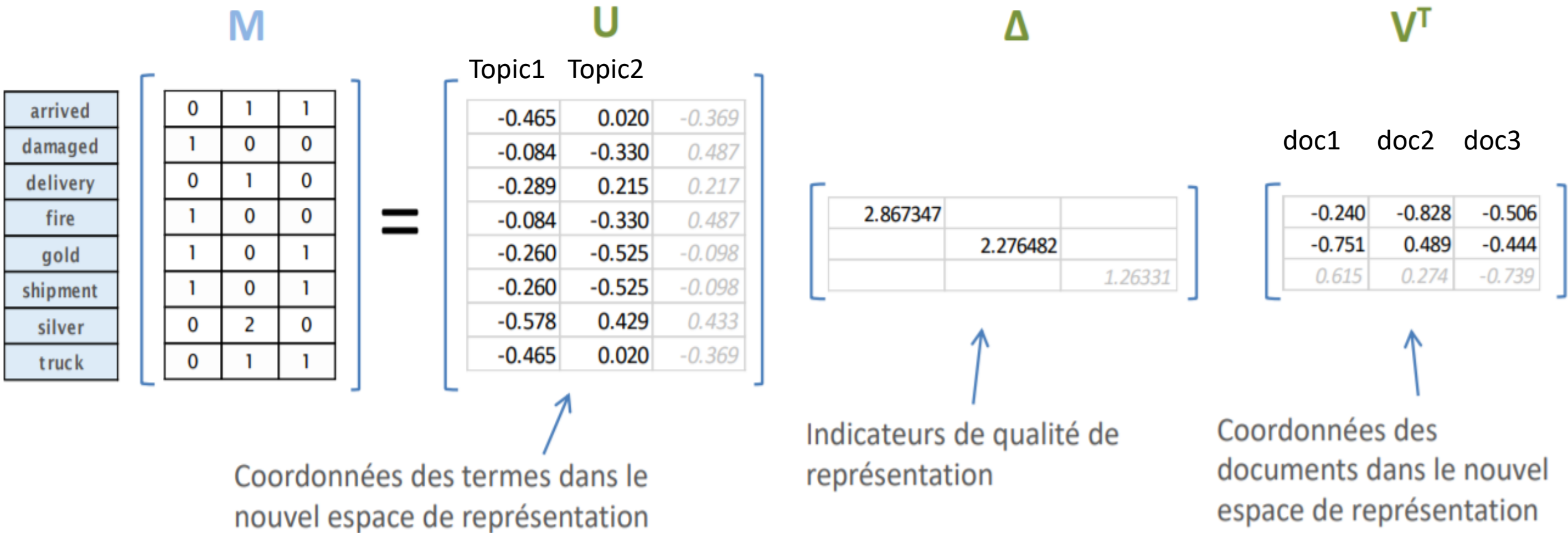
$$\lambda_k (k = 1, \dots, 3) = (1.175, 0.740, 0.228)$$



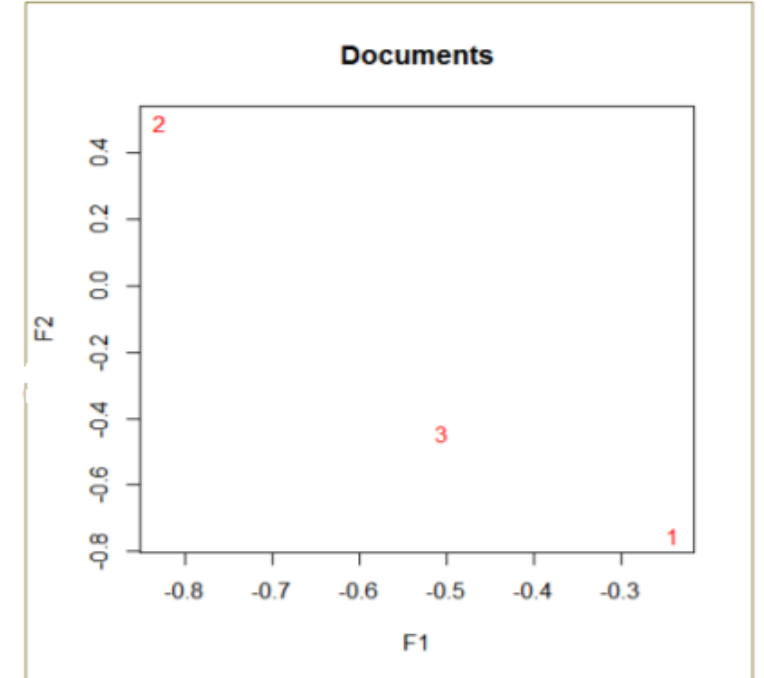
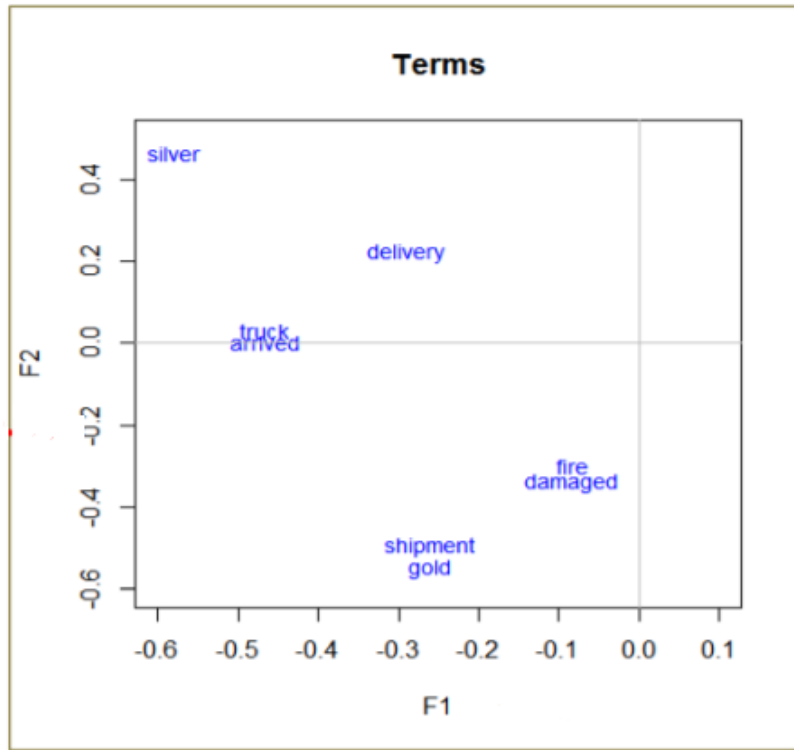
En pourcentage
cumulé d'information

(54.81, 89.36, 100.0)

Nouvelle représentation



Nouvelle représentation



=> Plus le terme (document) est loin de l'origine plus il contribue

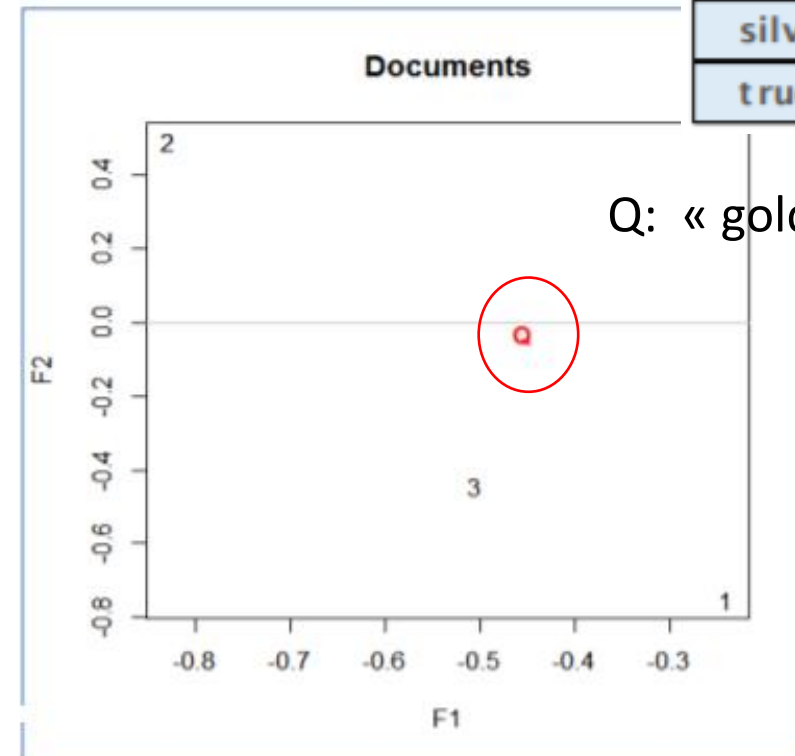
- Pour positionner un nouveau document, il faut le projeter sur le même plan (k=2)

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

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

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

Q: « gold silver truck »

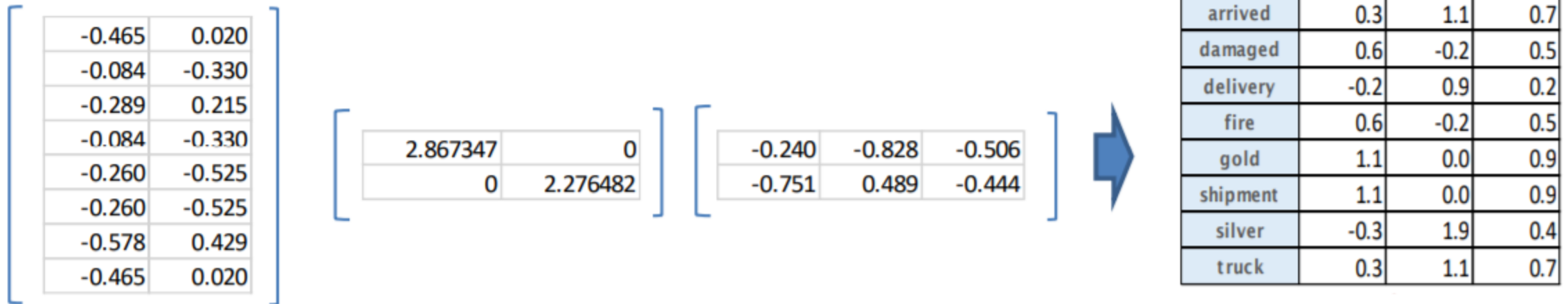


On positionne un document à partir des termes qui le compose

Reconstitution du tableau termes/documents

⇒ Il est possible d'approximer le tableau de données initial dans l'espace de représentation réduit.

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



⇒ 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 $k \ll \min(p, n)$

- Que pensez-vous de la reconstitution du tableau termes/documents pour $k = 2$?

	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

Si on applique un seuil = 0,5, alors
on peut revenir au tableau initiale

Remarques:

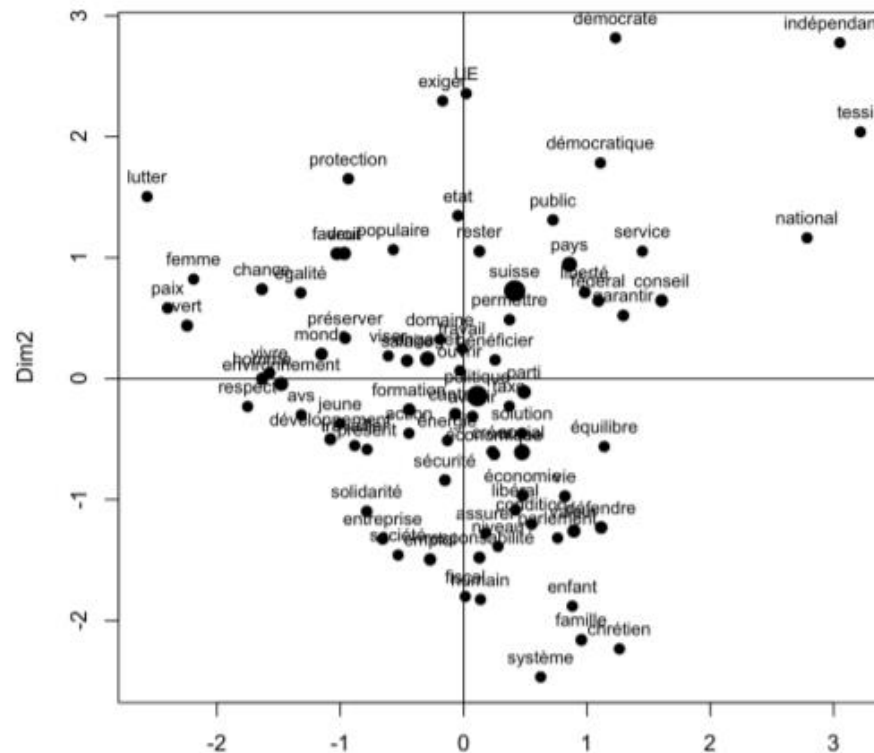
- La méthode LSI repose sur la décomposition en valeurs singulières de la matrice termes documents => *valable quelle que soit la pondération utilisée.*
- Les topics (facteurs) sont définis *par les termes à divers degrés.*
- Il est possible de positionner les documents dans le nouvel espace de représentation

ANALYSE FACTORIELLE DES CORRESPONDANCES

AFC: Analyse Factorielle des Correspondances

- L'AFC s'utilisent pour décrire et hiérarchiser les relations statistiques qui peuvent exister entre des individus placés en ligne et des variables placées en colonnes dans un tableau
- La spécificité de l'AFC est qu'elle considère *en même temps un nuage de point représentant les lignes* (individus) et *un autre représentant les colonnes (variables)*.

- L'AFC a précisément pour but de projeter le nuage des v points-termes *associée à la configuration pondérée (f, D_X)*, de haute dimensionnalité, dans un espace plus petit, typiquement à 2 dimensions, afin de pouvoir le visualiser – tout en veillant à conserver un maximum de dispersion Δ



AFC

- 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.
- C'est le cas pour la matrice termes documents, en particulier pour les pondérations binaires et fréquences d'apparition d'un terme.

Pondération fréquence
d'apparition

	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

	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

1/2

1/2

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

$P(\text{terme} / \text{document})$

4/13

5/13

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

1/5

1/4

2/13

$P(\text{document} / \text{terme})$

Distances entre profils

Y / X	x_1	x_l	x_L	Σ
y_1	$\begin{array}{ccc} & \vdots & \\ \cdots & n_{kl} & \cdots \\ & \vdots & \end{array}$			$n_{k.}$
y_k				
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 KHI 2

Distances entre termes
(distances entre profils lignes)

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

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

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

Distances entre documents
(distances entre profils colonnes)

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

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

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

Détection du nombre adéquat de facteurs

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

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

Avec K: termes et L : documents

Qualité de représentation (COS²) et contribution aux facteurs (CTR)

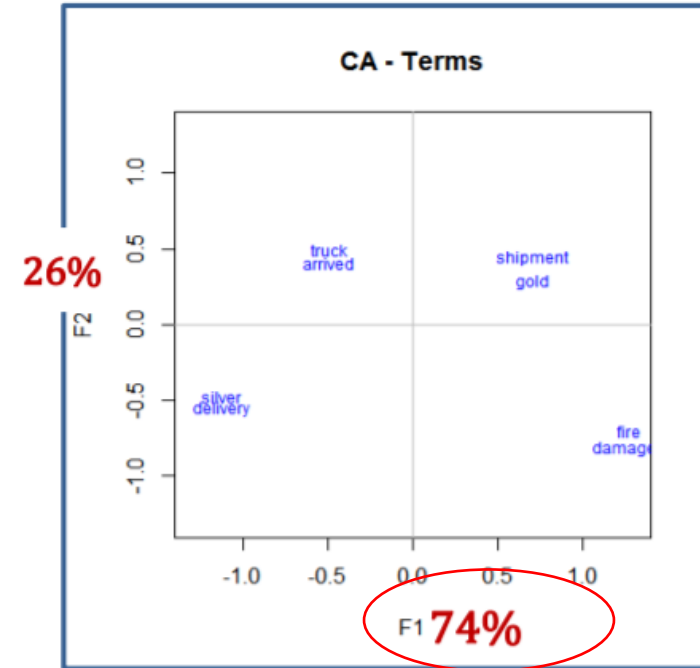
Profils sur les termes

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)

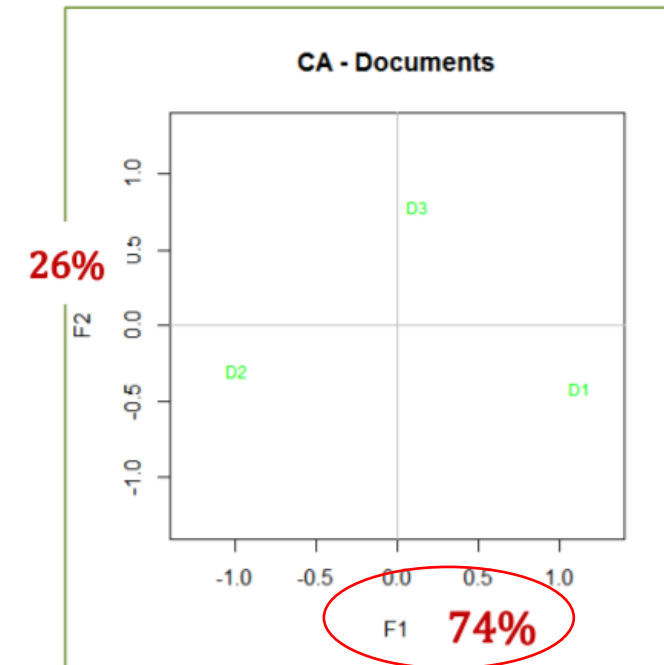
Profils sur les colonnes

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)

positionnement relatif des termes (profils des lignes)



positionnement relatif des documents (profils des colonnes)



Association termes-Documents

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

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

Inertie totale

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

- 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

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

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

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

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

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

Représentation simultanée

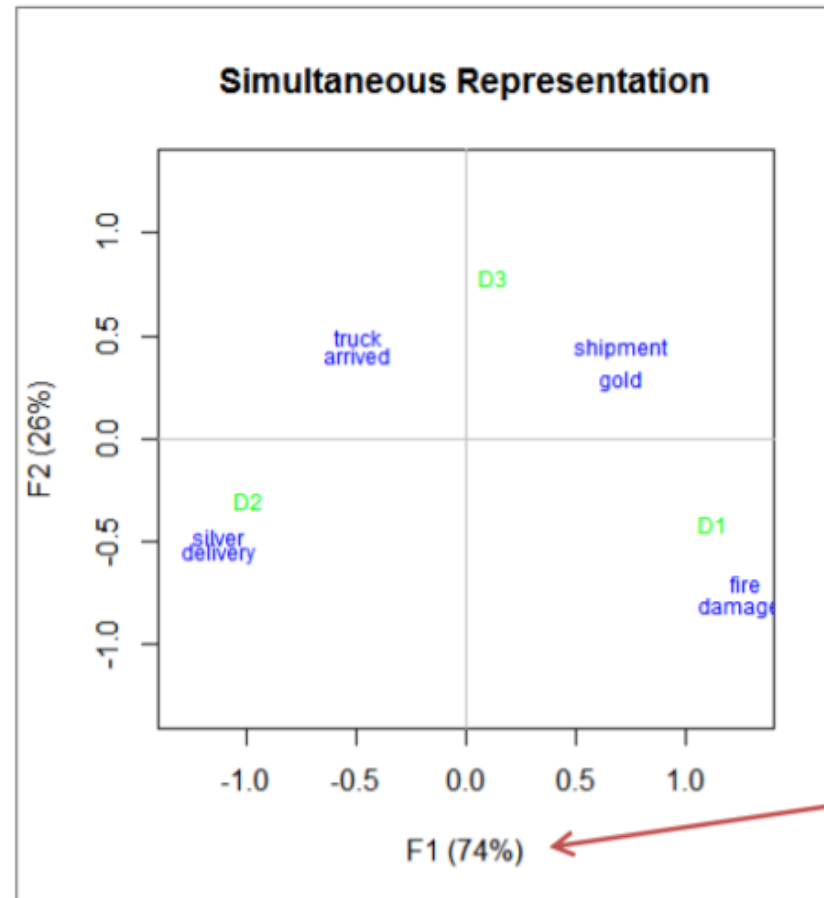
- La représentation simultanée est possible grâce aux relations de transition c.-à-d. il est possible d'obtenir les coordonnées d'une modalité colonne à partir des coordonnées de l'ensemble des modalités lignes, et inversement.

Coordonnée de la modalité ligne k sur le facteur 1

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

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

Coordonnée de la modalité colonne l sur le facteur 1



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

Le facteur 1 détermine en grande partie la lecture.

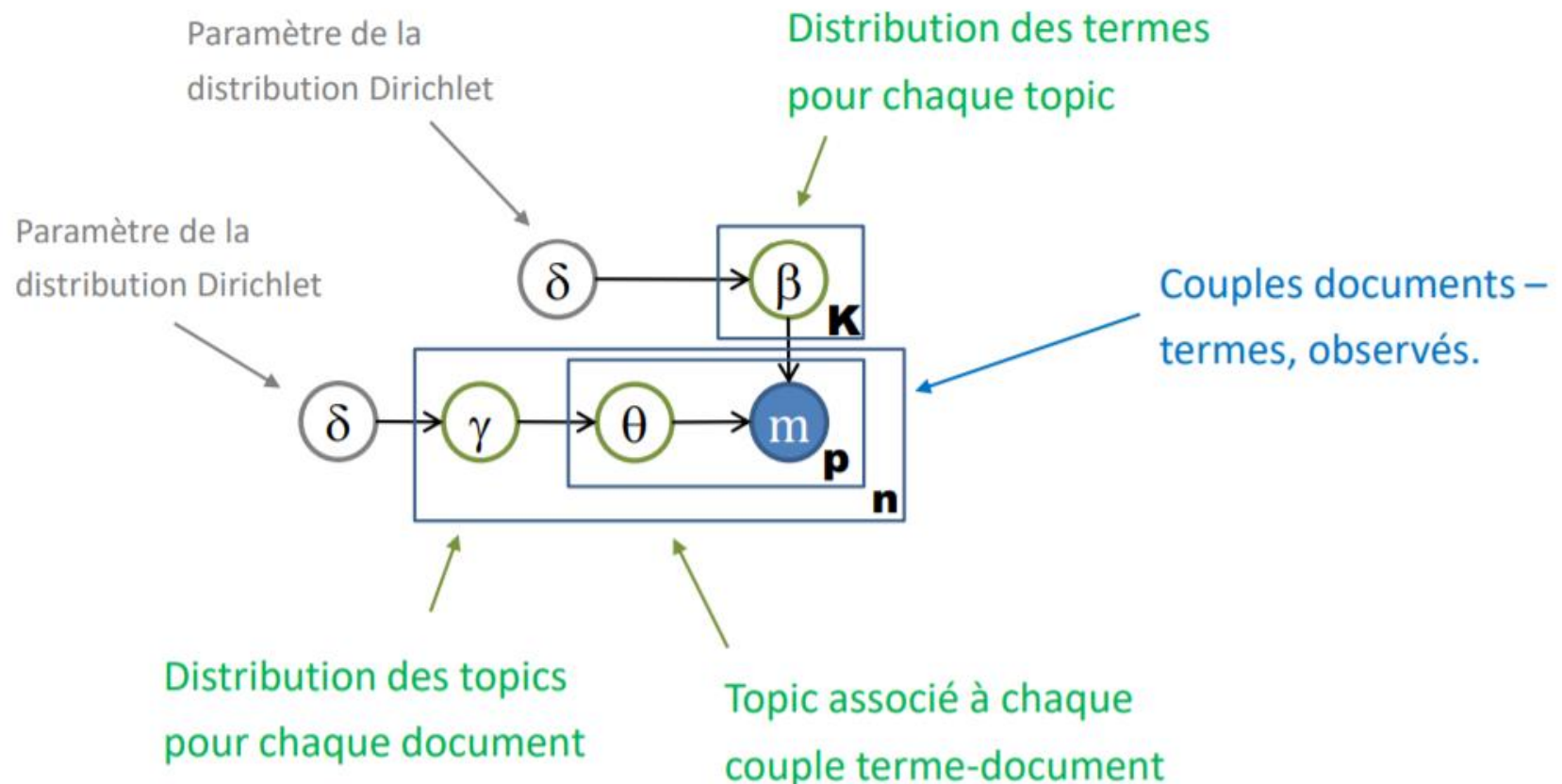
AFC: résumé

- 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.
- La dispersion doit être la même pour les modalités lignes et les modalités colonnes.
- Les facteurs sont orthogonaux deux à deux.
- L'AFC permet de positionner les termes entre eux (en fonction des documents qui les contiennent) et les documents entre eux (en fonction des termes qu'ils contiennent).
- Elle permet aussi de situer les associations termes – documents.

LATENT DIRICHLET ALLOCATION (LDA)

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



La modélisation va nous fournir les éléments en vert.

LDA

- Le changement de représentation est intéressant dans le cas où K nombre de topics \ll p nombre de termes.
- Hypothèse de travail: Les topics ne sont pas censés être corrélés entre eux.
- Distribution des termes (j) pour chaque topic (k) : distribution de Dirichlet symétrique de paramètre δ .

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

avec φ_{kj} Sélectionner le topic (k) pour le terme (j)

- Distribution des topics pour chaque document (Dirichlet)

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

θ_{ik} Sélectionner le topic pour chaque couple terme document

- 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.

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 d'autres algorithmes pour avoir plus de visibilité sur leur classification par exemple.

Assignation des termes aux topics selon les documents



	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).