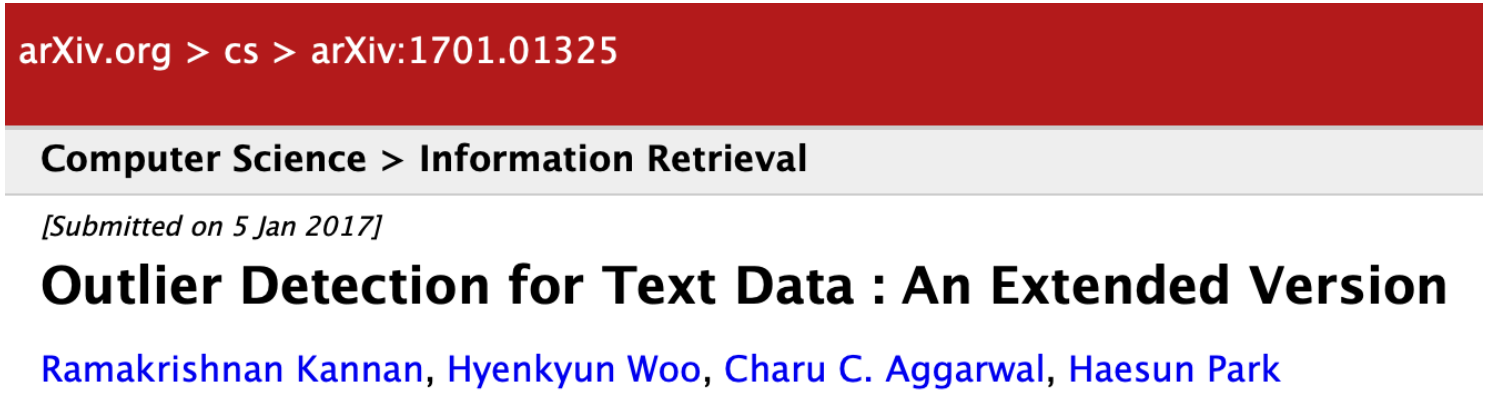
Détection d'anomalies dans les documents textuels

factorisation en matrices non-négatives et calcul de résidus

Séminaire TALia du 26 novembre 2021

Source principale



Github pour outliernmf

Un problème difficile pour les documents textuels

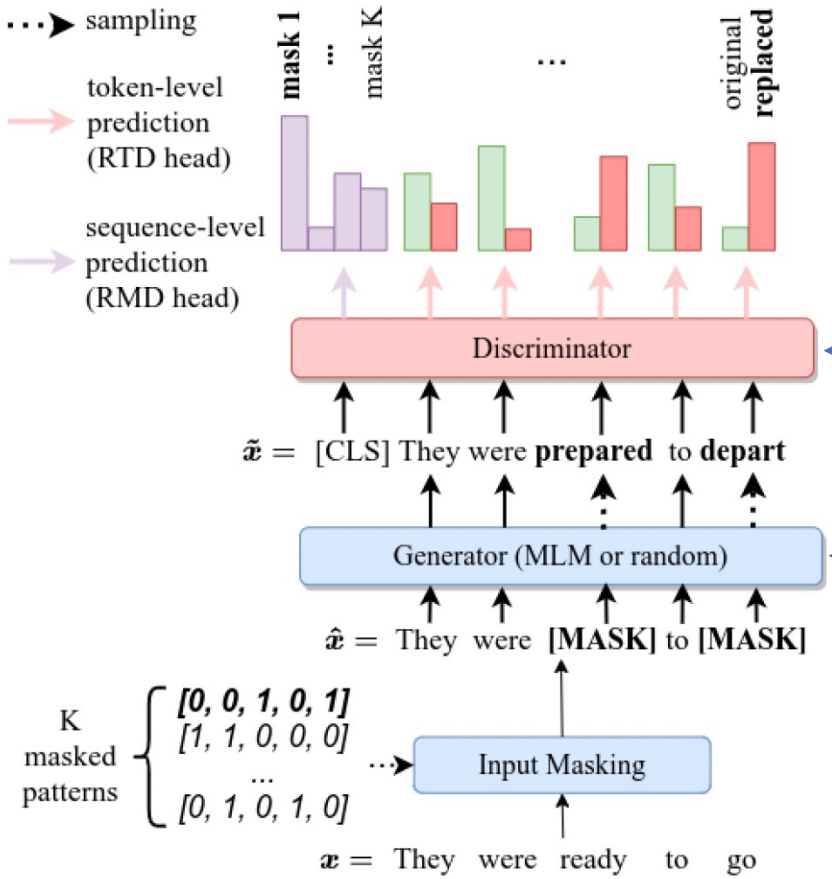
Approches pour la détection d'anomalies

1. Les méthodes basées sur une notion de **distance** : une anomalie est une observation éloignée des régions denses.
2. Les méthodes basées sur une notre de **densité** : une anomalie est une observation située dans une région de faible densité.
3. Les méthodes sur l'**erreur de reconstruction** : une anomalie est une observation qu'un AE ne parvient pas à reconstruire correctement (ou une observation que le discriminant d'un GAN classe comme anormale).
4. Les méthodes basées sur la capacité d'un modèle à **détecter des modifications des données normales**.
5. Les méthodes de type **One Class-SVM** : une anomalie est une observation qui est représentée par un point situé à l'extérieur d'une sphère qui englobe une majorité de points normaux (« quantile » multidimensionnel).

Difficulté pour les données textuelles

- Un problème en **grande dimension** : taille du vocabulaire
- Des **features** de type *fréquences de mots* ou *TF-IDF* **sont sparses** même pour les documents normaux
- Un sujet encore **peu exploré** jusqu'à récemment (cf. article publié en 2017)

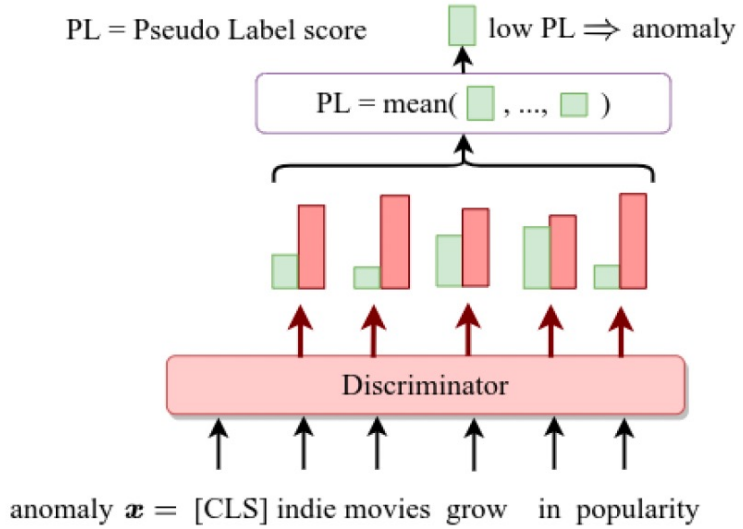
Rappel : détection d'anomalies avec le Deep Learning



$$x = [x_1, x_2, ..., x_T]$$
$$h(x) = [h_1, h_2, ..., h_T]$$

Transformer encoder *D* = « BERT »

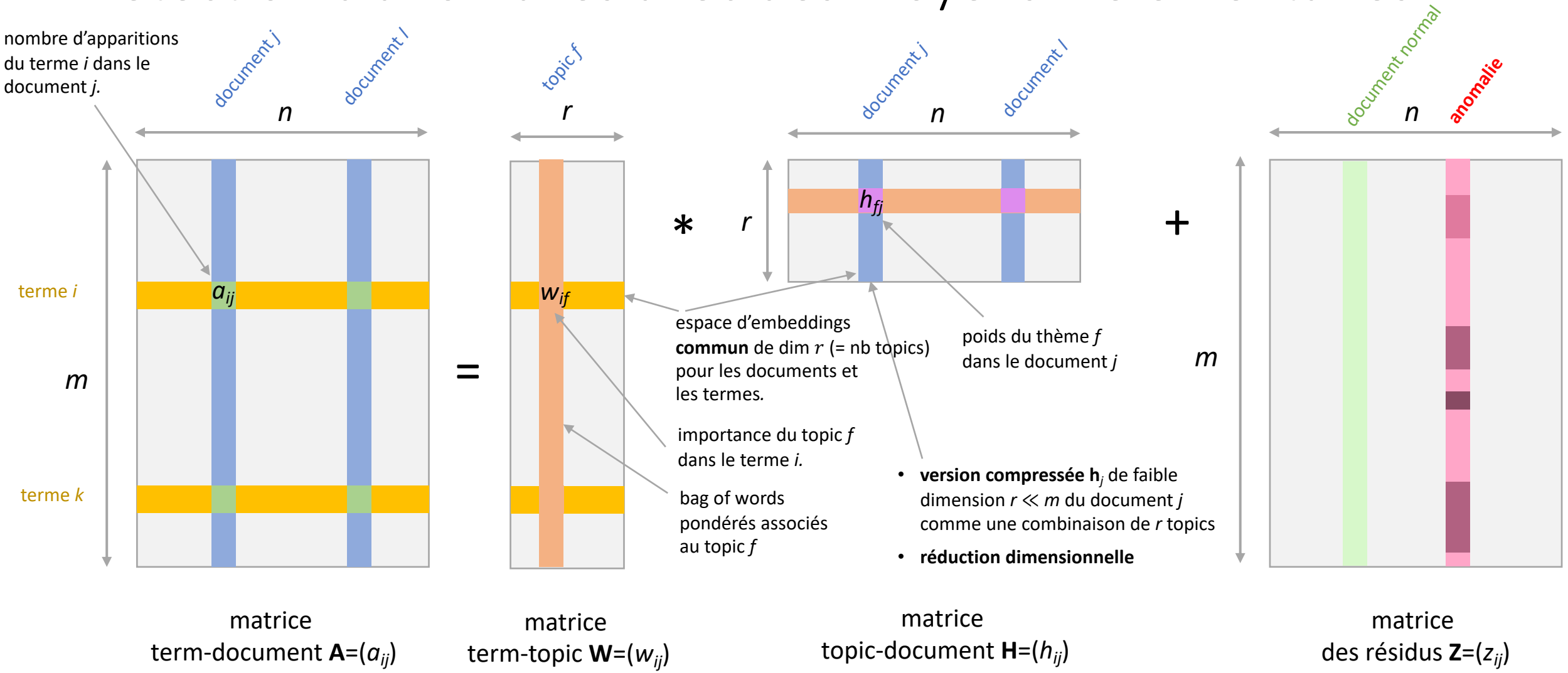
Transformer encoder *G* = « BERT »



Entraînement self-supervisé

Test

Détection d'anomalies avec des moyens « élémentaires »



matrice term-document **A**=(a_{ij})

matrice term-topic **W**=(w_{ij})

matrice topic-document **H**=(h_{ij})

matrice des résidus **Z**=(z_{ij})

Méthodes de factorisation et de réduction dimensionnelle

scikit learn

Prev Up Next

scikit-learn 1.0.1
Other versions

Please **cite us** if you use the software.

2.5. Decomposing signals in components (matrix factorization problems)
2.5.1. Principal component analysis (PCA)
2.5.2. Kernel Principal Component Analysis (kPCA)
2.5.3. Truncated singular value decomposition and latent semantic analysis
2.5.4. Dictionary Learning
2.5.5. Factor Analysis
2.5.6. Independent component analysis (ICA)
2.5.7. Non-negative matrix factorization (NMF or NNMF)
2.5.8. Latent Dirichlet Allocation (LDA)

Truncacted SVD

$$\mathbf{X} \approx \mathbf{X}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top \qquad \text{rang}(\mathbf{U}_k) = \text{rang}(\mathbf{V}_k) = \text{rang}(\mathbf{X}_k) = k$$

Connu sous le nom de LSA Latent Semantic Analysis lorsqu'appliqué à une matrice term-document calculée par un `CountVecorizer` ou un `TfidfVectorizer`.

NMF

$$\|\mathbf{X} - \mathbf{Y}\|_F = \frac{1}{2} \sum_{i,j} (X_{ij} - Y_{ij})^2$$

$$(\mathbf{W}_0, \mathbf{H}_0) = \underset{w_{ij} \geq 0, h_{ij} \geq 0}{\arg\min} \|\mathbf{A} - \mathbf{WH}\|_F^2 + \text{régularisation}$$

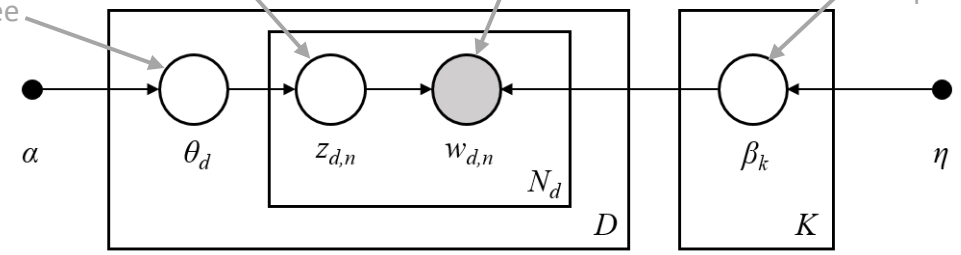
LDA

Une distribution θ_d de topics pour chacun des *D* documents d. Chaque distribution θ_d est tirée d'une loi de Dirichlet de paramètre α.

Un topic $z_{d,n}$ tiré de la distribution θ_d pour chaque mot n du document d.

Un mot tiré de la distribution β_k pour le topic $k = z_{d,n}$.

Une distribution β_k sur les V mots du vocabulaire pour chacun des *K* topics. Chaque distribution est tirée d'une loi de Dirichlet de paramètre η.



Factorisation en matrices non-négatives de rang faible

Représentation des documents comme des superpositions $\mathbf{H}_0$ de topics associés à des bag of words pondérés $\mathbf{W}_0$:

$$(\mathbf{W}_0, \mathbf{H}_0) = \underset{w_{ij} \geq 0, h_{ij} \geq 0}{\arg\min} \|\mathbf{A} - \mathbf{WH}\|_F^2 + \text{régularisation}$$

Définition d'un score d'anomalie global

$$\|\mathbf{Z}\|_{1,2} = \|[\mathbf{z}_1, \ldots, \mathbf{z}_n]\|_{1,2} := \sum_{i=1}^{n} \|\mathbf{z}_i\|_2$$

score d'anomalie pour le document n°*i*

Favoriser les représentations où peu de documents sont des anomalies

$$(\mathbf{W}_0, \mathbf{H}_0, \mathbf{Z}_0) = \underset{w_{ij} \geq 0, h_{ij} \geq 0; z_{ij}}{\arg\min} \|\mathbf{A} - \mathbf{WH} - \mathbf{Z}\|_F^2 + \alpha \|\mathbf{Z}\|_{1,2}$$

Favoriser les représentations interprétables. Une majorité de documents devraient être des superpositions de peu de topics.

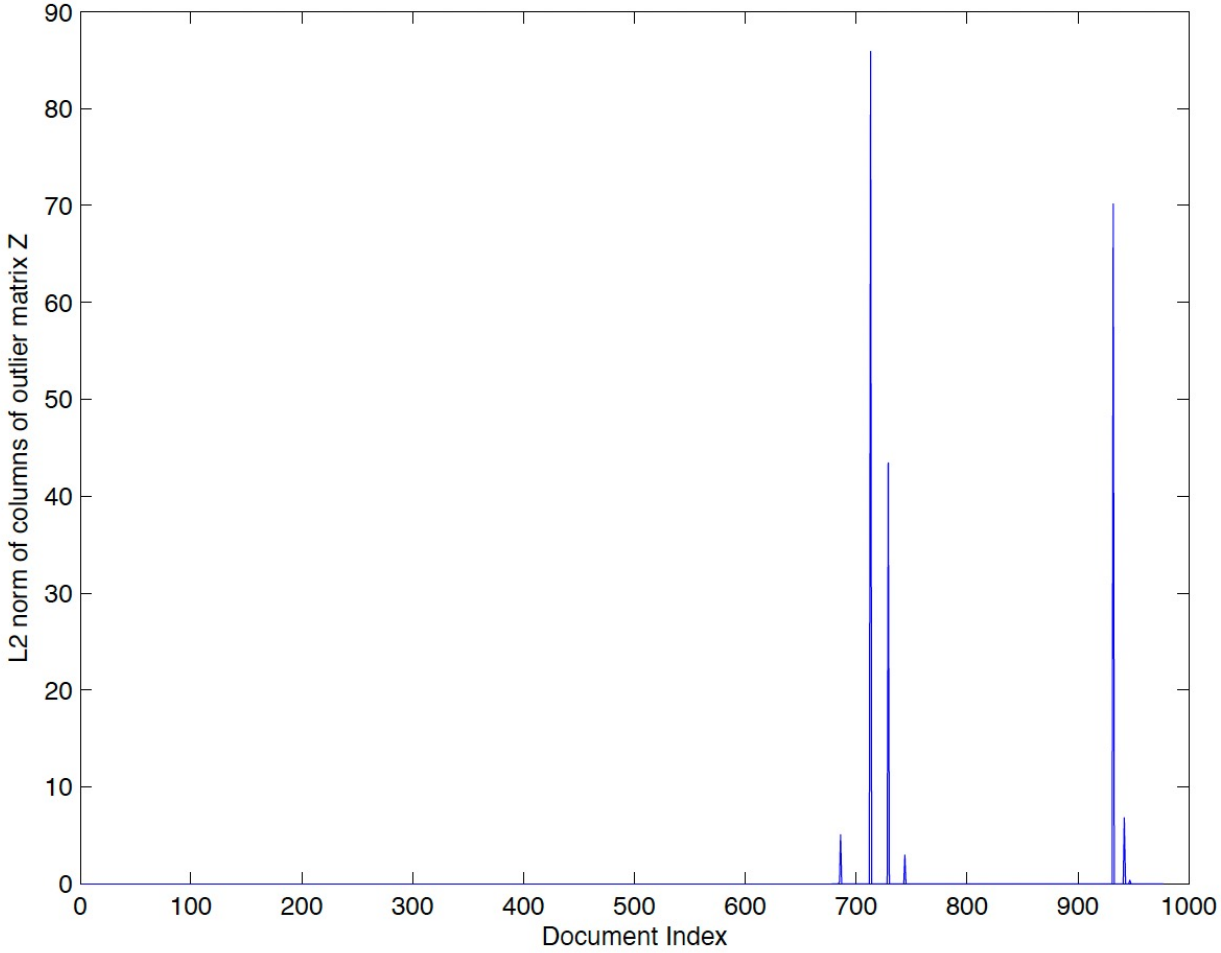
$$(\mathbf{W}_0, \mathbf{H}_0, \mathbf{Z}_0) = \underset{w_{ij} \geq 0, h_{ij} \geq 0; z_{ij}}{\arg\min} \|\mathbf{A} - \mathbf{WH} - \mathbf{Z}\|_F^2 + \alpha \|\mathbf{Z}\|_{1,2} + \beta \|\mathbf{H}\|_1$$

Calcul des scores d'anomalies dans un cas pratique

BBC dataset

http://mlg.ucd.ie/datasets/bbc.html

- 2250 documents
- 5 classes : *business*, *entertainment*, *politics*, *sport*, *tech*
- 1000 documents conservés (*business* et *politics*)
- 50 documents *tech* à identifier comme des anomalies



TONMF un algorithme d'optimisation dédié

TONMF : Text Outlier using Non-negative Matrix Factorization utilise une adaptation de la méthode d'optimisation BCD (**Block Coordinate Descent**)

$\min f(x)$ avec $x \in \mathcal{X}$ et $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$

Si $\mathbf{x}^{(i)} = (\mathbf{x}_1^{(i)}, \cdots, \mathbf{x}_m^{(i)})$

on calcule $\mathbf{x}_j^{(k+1)} \leftarrow \underset{\xi \in \mathcal{X}_j}{\operatorname{argmin}} f(\mathbf{x}_1^{(k+1)}, \cdots, \mathbf{x}_{j-1}^{(k+1)}, \xi, \mathbf{x}_{j+1}^{(k)}, \cdots, \mathbf{x}_m^{(k)})$

$$\mathbf{Z}^{(k+1)} \leftarrow \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{A} - \mathbf{Z} - \mathbf{W}^{(k)}\mathbf{H}^{(k)}\|_F^2 + \alpha\|\mathbf{Z}\|_{1,2}$$

$$\mathbf{Z}^{(k+1)} = \arg\min_{\mathbf{Z}} \sum_i \frac{1}{2}\|\bar{\mathbf{a}}_i - \mathbf{z}_i\|_2^2 + \alpha\|\mathbf{z}_i\|_2 \qquad \bar{\mathbf{a}}_i = \mathbf{a}_i - (\mathbf{W}^{(k)}\mathbf{H}^{(k)})_i$$

$$(\mathbf{W}^{(k+1)}, \mathbf{H}^{(k+1)}) \leftarrow \underset{\mathbf{W}\geq\mathbf{0}, \mathbf{H}\geq\mathbf{0}}{\arg\min} \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H} - \mathbf{Z}^{(k+1)}\| + \beta\|\mathbf{H}\|_1$$

Comparaison avec l'approche par Transformers

TONMF

- Conceptuellement simple, interprétable, similaire à d'autre techniques de factorisation de matrices ou de décomposition de signaux comme *k*-SVD ou LDA.
- Le biais inductif est très fort : on fait l'hypothèse que les documents normaux correspondent à une superposition additive d'un petit nombre de thèmes.

Méthode basée sur les Transformers en mode « GAN »

- Détecte des anomalies fines basées sur l'idée qu'une anomalie correspond à un document sur lequel le module discriminant doute sur les substitutions effectuées.
- Pas d'hypothèses de linéarité ou de superposition additive.
- Anomalies difficiles à interpréter.