Energie textuelle de documents

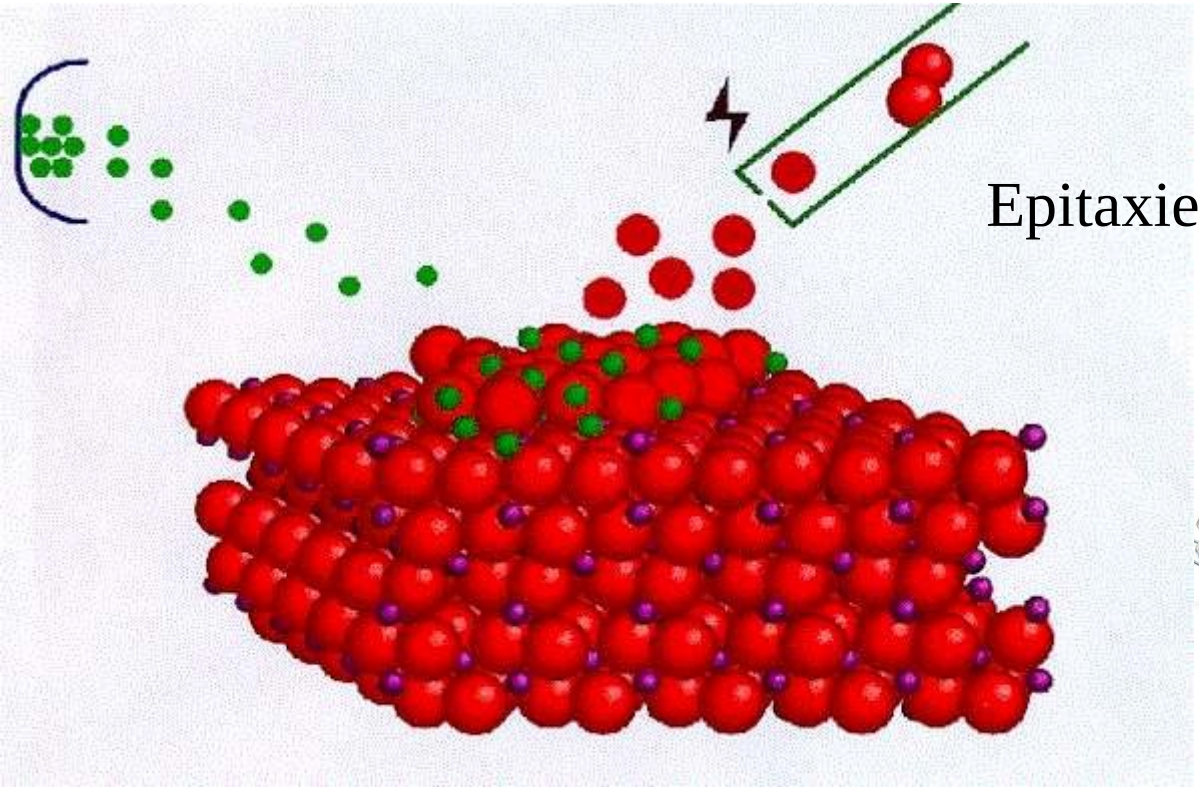
Juan-Manuel TORRES

juan-manuel.torres@univ-avignon.fr

AVIGNON UNIVERSITÉ

LABORATOIRE INFORMATIQUE D'AVIGNON
Université d'Avignon et des Pays de Vaucluse

Micro introduction à la Physique Statistique...



Epitaxie par jet moléculaire

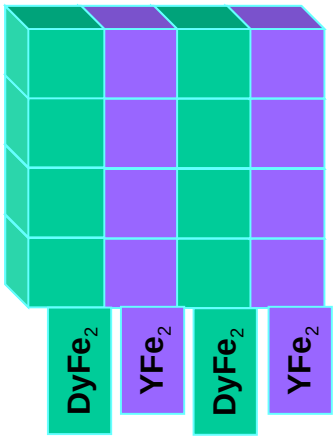
La croissance de couches atomiques par jets moléculaires, c'est la technique de l'épitaxie. Crédit : phocea.CEA

https://www.futura-sciences.com/sciences/actualites/physique-supraconducteurs-nanometriques-nouvelle-electronique-16981/

Micro introduction à la Physique Statistique...

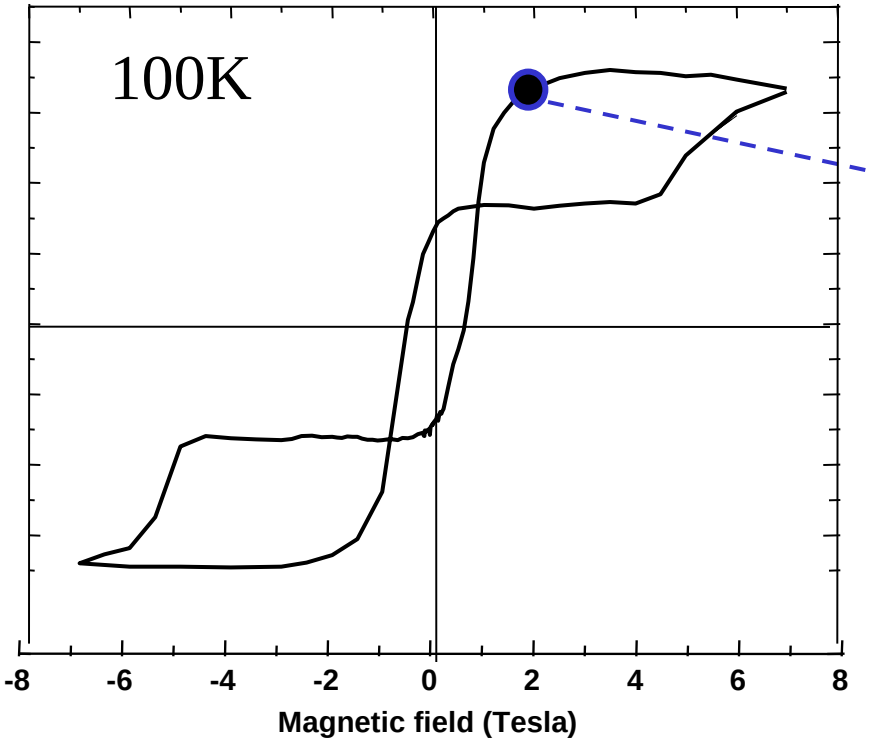
Nouveaux matériaux magnétiques

Epitaxie par jet moléculaire

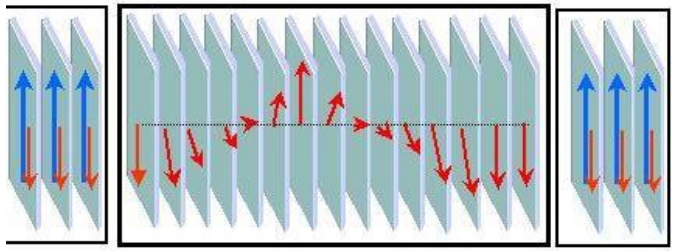


Mesures magnétiques

Mesures magnétiques et configuration de spins



Spin : représentation de chaque atome comme un petit aimant



Modèles théoriques de la Physique Statistique:

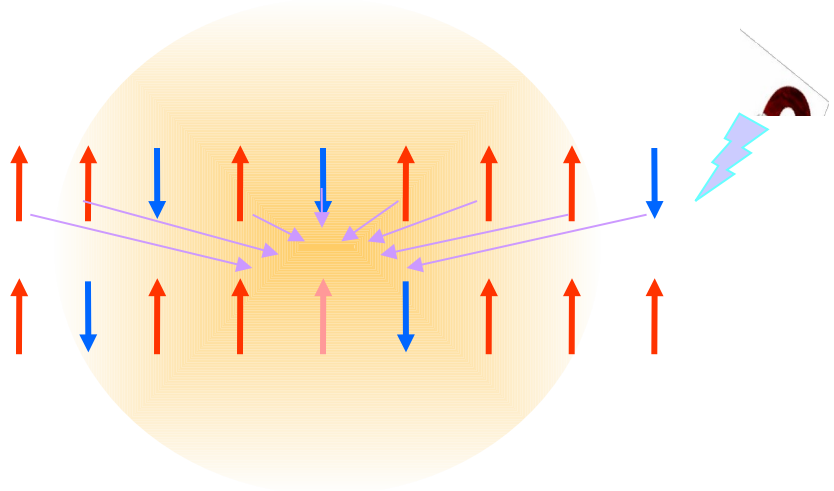
Modèle d'Ising: deux orientations possibles ↓↑

Energie du système

$$E = E\,(interactions) + E(champ)$$

$$E_{ij} = J_{ij}\, s_i\, s_j \qquad + \qquad E_i = H\, s_i$$

$$J_{ij} = J_{ji}$$



Configuration de spin final : minimisation de E

p(état du système) = $f\,(E, T, Z)$; Z=fonction de partition ;
T = température

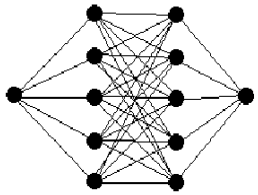
Mais... où entre le TALN dans toute cette histoire ?

Applications *exotiques* de la physique statistique

Physique Statistique

Modèle de spins d'Ising

Modèle de Hopfield



Réseaux neuronaux

TALN

Résumé automatique

Segmentation thématique

...

Mémoire associative

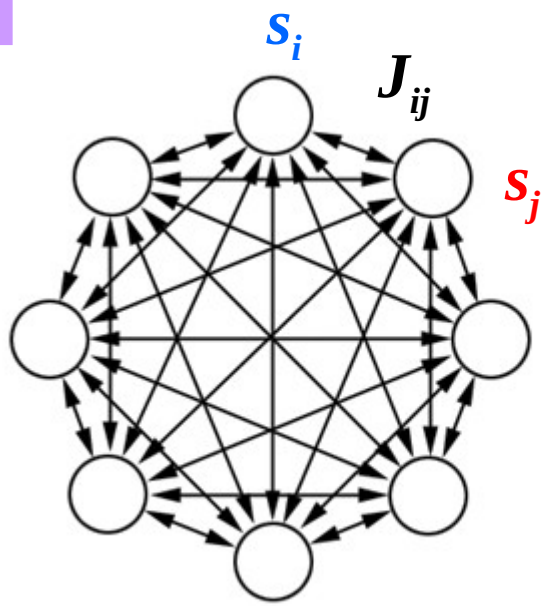
(Hopfield, 1982)

Modèle de spins d'Ising

neurone = spin ↓↑

$J_{ij} = J_{ji}$

$E_{ij} = J_{ij}\, s_i\, s_j$



Réseaux neuronaux

Règle d'Hebb

$J_{ij} = s_i\, s_j$

Apprentissage

Récupération: minimisation de E

Limitations :

- Patrons corrélés → erreur de récupération
- Capacité ≈ 0, 14 N

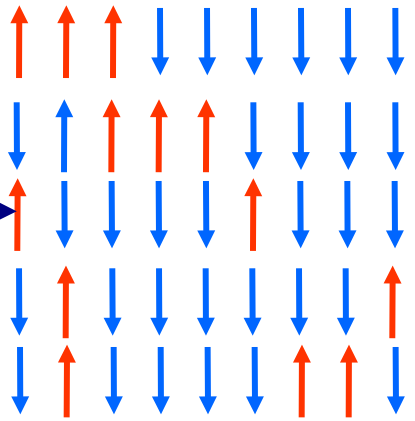
Codage des documents comme un système de spins

- Modèle vectoriel (sac de mots)
- Mots filtrés, normalisés et lemmatisés
  (Porter, 1980; Manning & Schutze, 2000)

Les maisons bleues de ma tante.
Un de mes tantes s'appelle Lulu.
J'adore tellement sa maison.
Le bleu est ma couleur préférée !
J'ai des chaussures blues toutes neuves.

maison	bleu	tante	appeler	lulu	adorer	neuf	chaussure	couleur
TF	TF	TF	0	0	0	0	0	0
0	0	TF	TF	TF	0	0	0	0
TF	0	0	0	0	TF	0	0	0
0	TF	0	0	0	0	0	0	TF
0	TF	0	0	0	0	TF	TF	0

Corrélés



Interaction entre spins

mot ~ neurone ~ spin s_i

$\begin{bmatrix} \text{TF} & \text{TF} & 0 \ldots & 0 \end{bmatrix} = \begin{bmatrix} s_0 & s_1 & s_2 \ldots . & s_N \end{bmatrix}$ Phrase ~ chaîne de spins

Doc → $S = \begin{pmatrix} s_1^1 & s_2^1 & \cdots & s_N^1 \\ s_1^2 & s_2^2 & \cdots & s_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^P & s_2^P & \cdots & s_N^P \end{pmatrix}$ Phrases x mots

$$J^\mu = \begin{pmatrix} s_1^\mu \\ \vdots \\ s_i^\mu \\ \vdots \\ s_N^\mu \end{pmatrix} \times \left(s_1^\mu \cdots s_i^\mu \cdots s_N^\mu \right) = \begin{pmatrix} j_{1,1}^\mu & j_{1,j}^\mu & \cdots & j_{1,N}^\mu \\ \vdots & \vdots & \ddots & \vdots \\ j_{i,1}^\mu & j_{i,j}^\mu & \cdots & j_{i,N}^\mu \\ \vdots & \vdots & \ddots & \vdots \\ j_{N,1}^\mu & j_{N,j}^\mu & \cdots & j_{N,N}^\mu \end{pmatrix}$$

$J = \Sigma\, J^\mu = (S^T \times S)$: c'est la mémoire d'Hopfield

L'énergie n'est pas utilisée

Energie textuelle

$$E = - \begin{pmatrix} s_1^1 & s_2^1 & \cdots & s_N^1 \\ s_1^2 & s_2^2 & \cdots & s_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^P & s_2^P & \cdots & s_N^P \end{pmatrix} \times J \times \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^P \\ s_2^1 & s_2^2 & \cdots & s_2^P \\ \vdots & \vdots & \ddots & \vdots \\ s_N^1 & s_N^2 & \cdots & s_N^P \end{pmatrix} = - \; S \times (S^T \times S) \times S^T$$

$$= -(S \times S^T) \times (S \times S^T)$$

$$¿ -(S \times S^T)^2$$

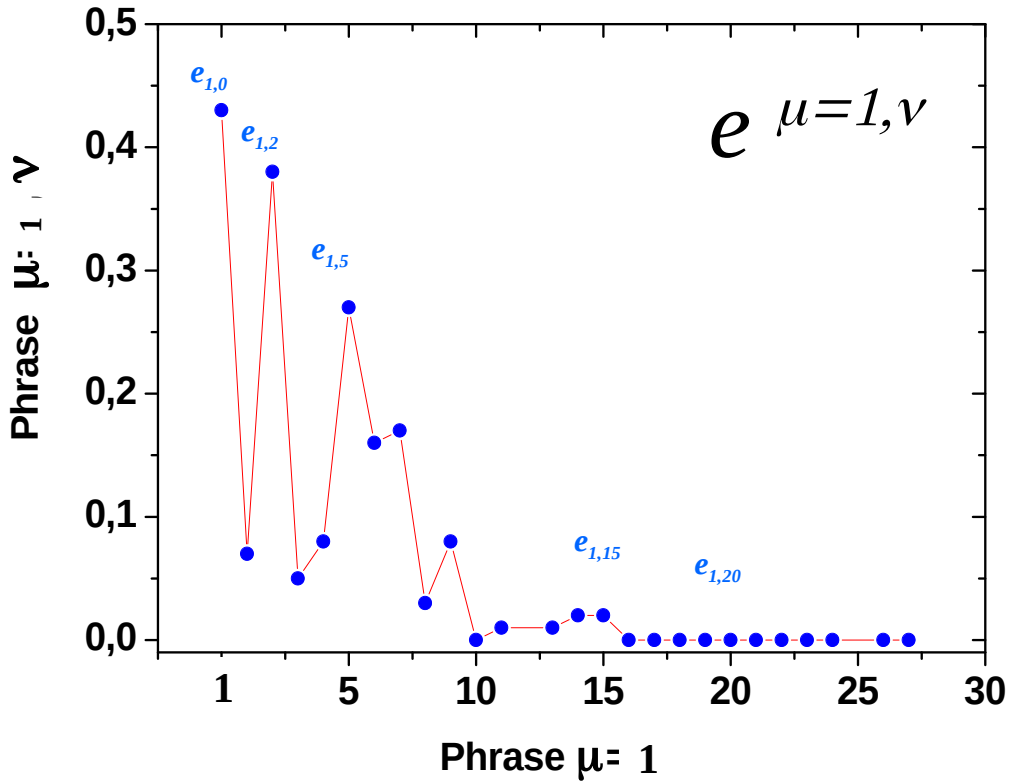
$$E = \begin{pmatrix} e^{1,1} \ldots & e^{1,P} \\ \vdots & \vdots \\ e^{\mu,1} \ldots & e^{\mu,P} \\ \vdots & \vdots \\ e^{P,1} \ldots & e^{P,P} \end{pmatrix}$$

$e^{\mu,\nu}$ = énergie entre la phrase μ et la phrase ν

Energie textuelle (phrase)

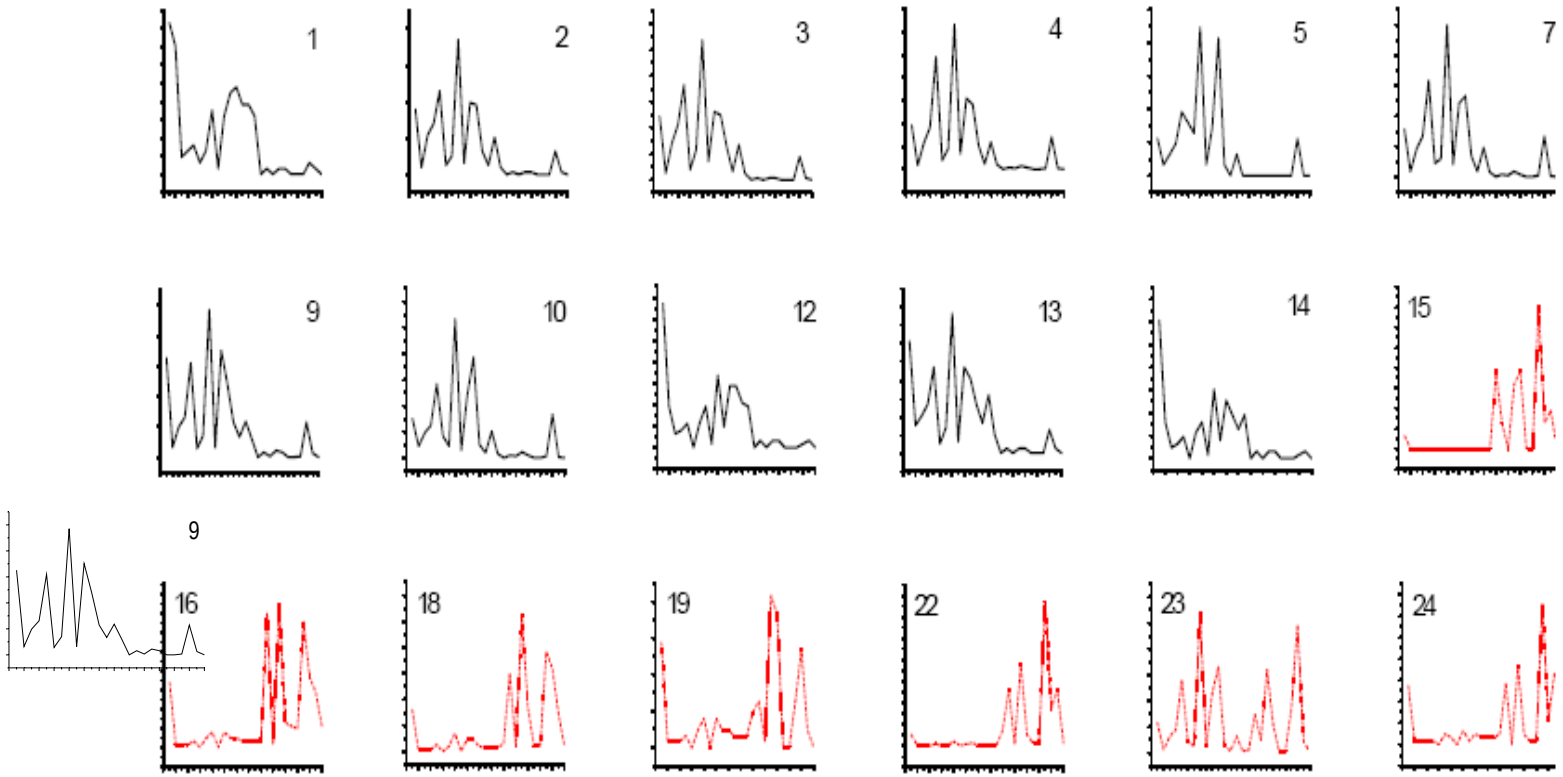
$$E = \begin{pmatrix} e^{1,1} & \cdots & e^{1,P} \\ \vdots & & \vdots \\ e^{\mu,1} & \cdots & e^{\mu,P} \\ \vdots & & \vdots \\ e^{P,1} & \cdots & e^{P,P} \end{pmatrix}$$

$$E^1 = -\sum_{\nu}^{p} e^{\nu,1}$$



Energie totale de la phrase $\mu=1$

Energie textuelle (doc)



| E^{μ} |des phrases :
Résumé automatique

Concordance entre courbes :
Segmentation thématique

Interprétation (théorie de graphes)

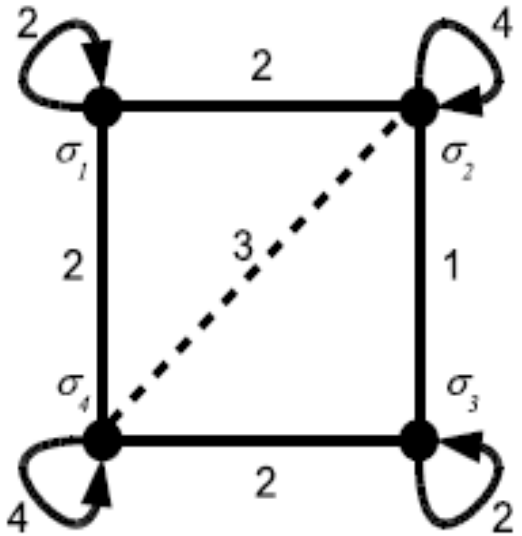
Energie textuelle : $E = -(S \times S^T) \times (S \times S^T)$ **O(P^2)**

$\iota -(S \times S^T)^2$

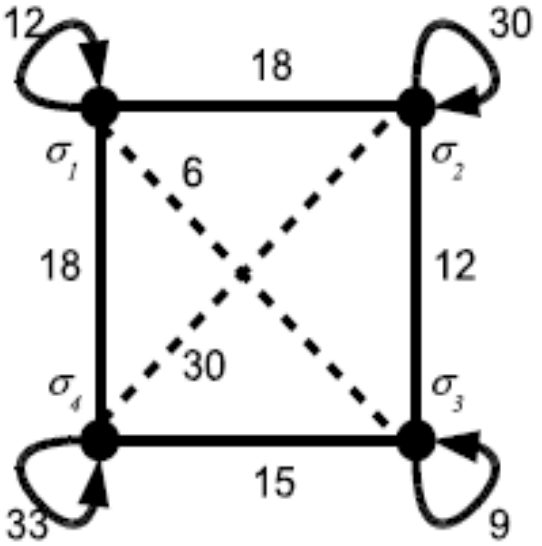
Exemple

	σ_1	σ_2	σ_3	σ_4
σ_1	2	2		2
σ_2	2	4	1	3
σ_3		1	2	2
σ_4	2	3	2	4

$S \times S^T$



$I(S)$



$G(S \times S^T)^2$

$\sigma_1 \cap \sigma_3 = \varnothing$ mais $\sigma_1 \cap \sigma_4 \neq \varnothing$ et $\sigma_4 \cap \sigma_3 \neq \varnothing$

O(P log P)

$\Rightarrow$ l'énergie entre σ_1 et σ_3 n'est pas nulle

Somme de trajets de **longueur 2** dans le graphe

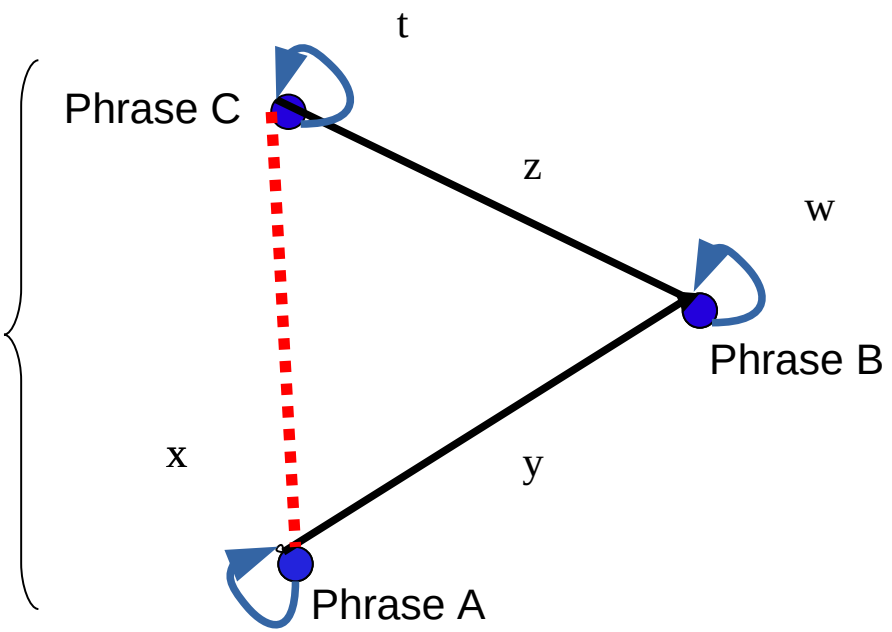
$S \times S^T$

Interactions entre phrases (A,B) et (B,C) ayant des mots en commun

$(S \times S^T)^2$

Interactions entre phrases ne partageant pas des mots (A,C) mais ayant des mots en commun avec des *phrases voisines* (B)

Coût (A, C) = $y \times z + w \times z + z \times t$



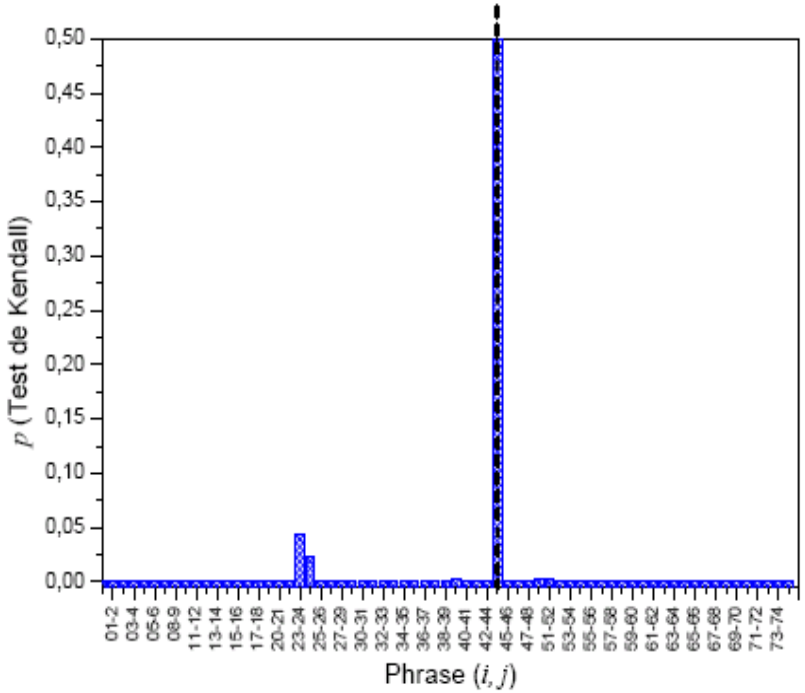
$$E = (S \times S^{T})^{2}$$

Résultats :

Frontières thématiques

Résumé automatique

Detéction de frontières : W Kendall

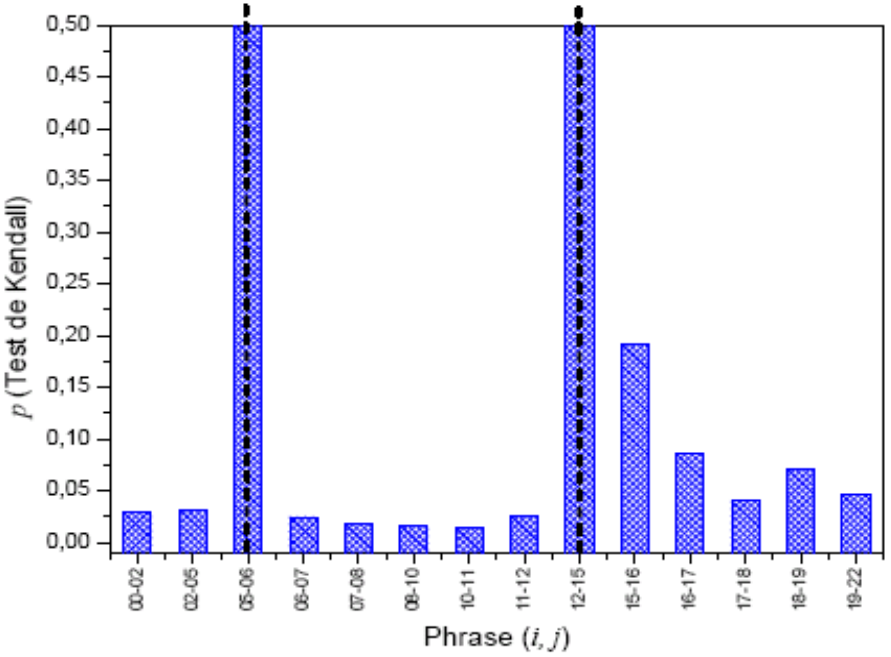


2 thématiques (en)

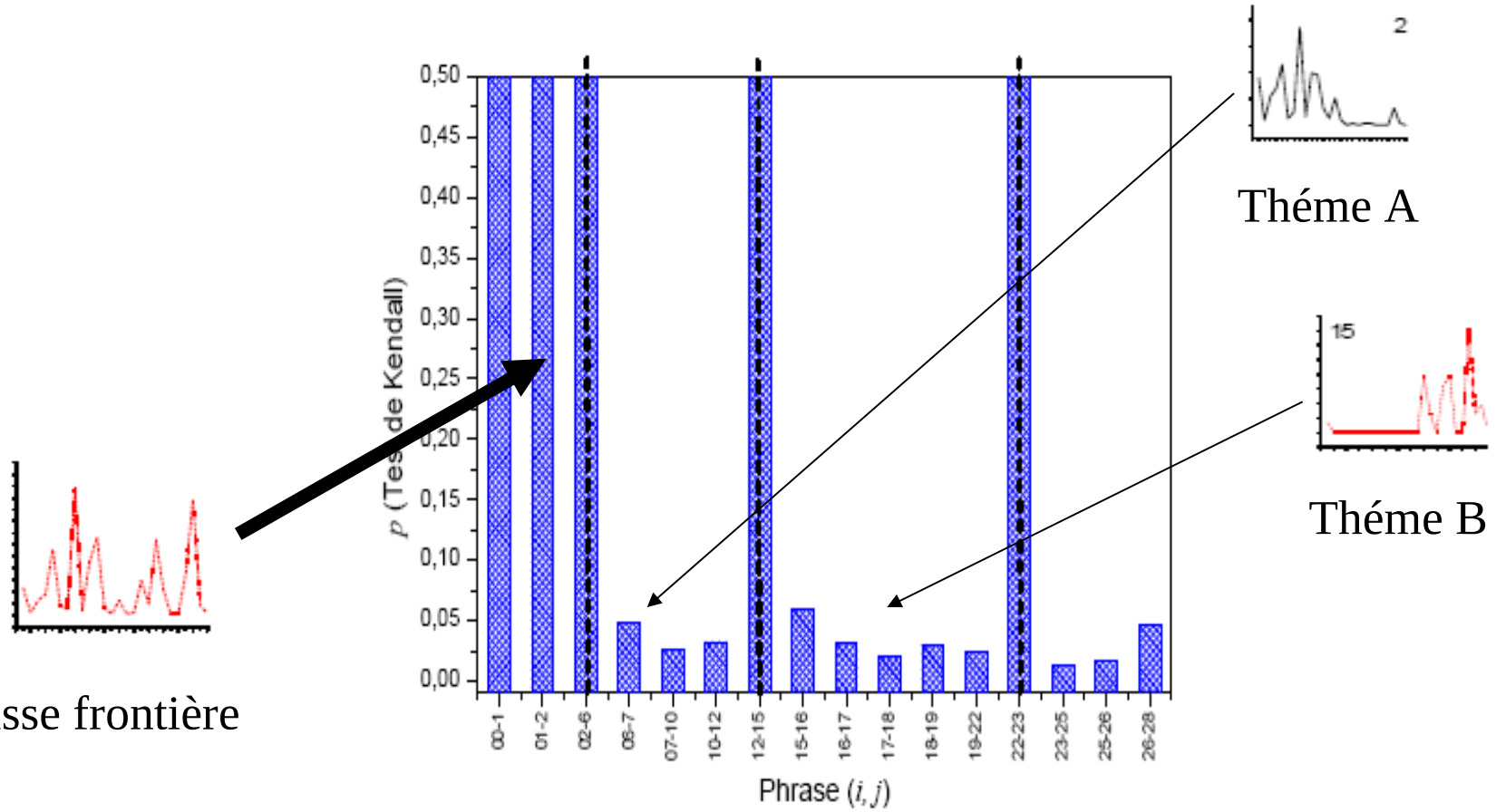
Coefficient de concordance *W* de Kendall et sa probabilité *p*

Test non-paramétrique
(Siege & Castellan 1988)

3 thématiques (fr)

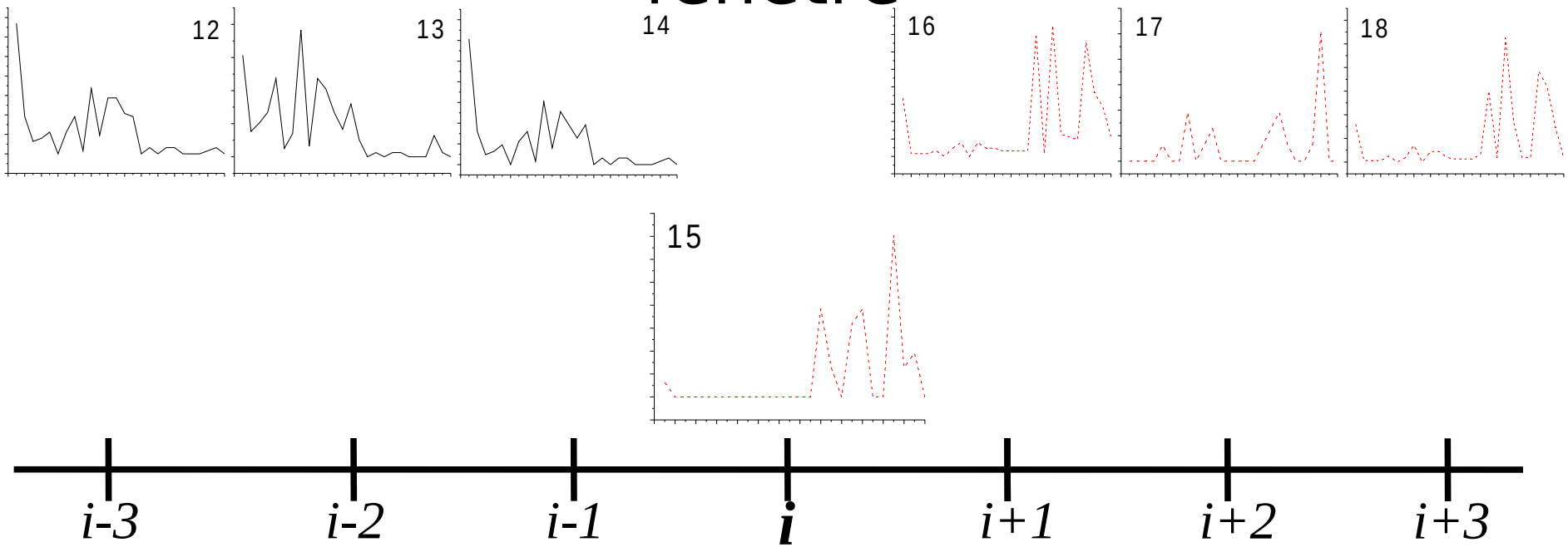


Erreurs en frontières



Texte avec 4 thématiques

Extraction de frontières : Kendall en fenêtre



$p_{i-3}=0,1$ $p_{i-2}=0,4$ $\mathbf{p_{i-1}=0,002}$ $p_{i+1}=0,001$ $p_{i+2}=0,004$ $p_{i+3}=0,003$

Seuil test de Kendall = 0.01 $\begin{cases} \text{pred++} & \text{si } p > 0.01 \\ \text{succ++} & \text{si } p < 0.01 \end{cases}$

$\text{pred}(i) = 2/3 \quad \&\& \quad \text{succ}(i) = 3/3$

$\Rightarrow$ ***i* est une frontière**

Frontières thématiques (français)

Taille du segment (en phrases)	Lcseg *	LIA_seg *	Energie	Energie en fenêtre	
					<front trouvées>
9-11	0,3272	**(0.3187-**0.4635**)**	0,4419	0,4134	7,1 / 9
3-11	0,3837	**(0,3685-**0,5105**)**	0,4403	0,4264	7,15 / 9
3-5	0,4344	(0,4204-0,5856)	**0,4167**	**0,4140**	5,08 / 9

* Le nb moyen de frontières n'est pas rapporté par (Sitbon et Bellot, 2005)

Résumé générique

F-score – Rouge-SU4 normalisé *(Lin, 2004)*

Corpus	Energie	Cortex*	*Baseline*
3-mélanges (web, Fr) 27 phr, 826 mots, 25%, 8 réf	**0,47150**	0,43068	0,32936
puces (web, Fr) 29 phr, 653 mots, 25%, 8 réf	0,53574	**0,55628**	0,32723
J'accuse (E. Zola, Fr) 206 phr, 4936 mots, 12%, 6 réf	0,58479	**0,60037**	0,26152
Lewinsky (Wikipedia, En) 30 phr, 816 mots, 20%, 7 réf	0,47757	**0,51076**	0,29248
Québec (Wikipedia, En) 44 phr, 1190 mots,25%, 8 réf	0,51179	**0,55656**	0,35244

**Torres et al. 2002*

Conclusion

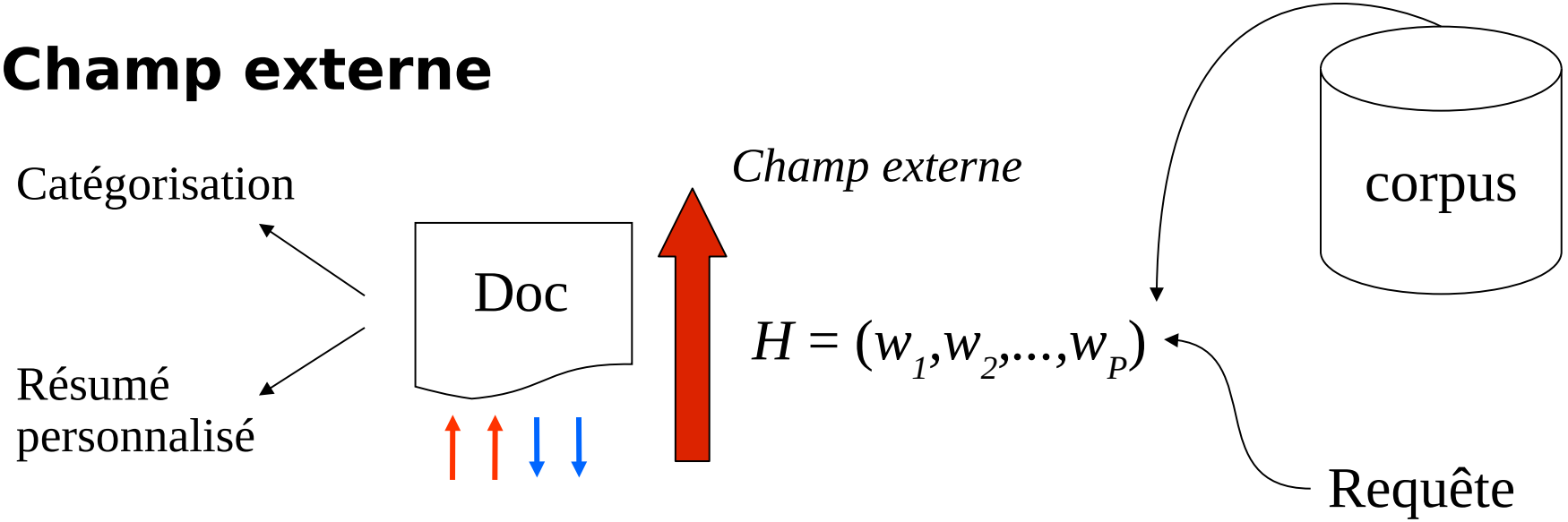
- Pont entre la Physique Statistique et le TALN
- Notion d'énergie textuelle
- Applications
  - **Résumé générique**
    - comparable au système Cortex (génerique) en termes de précision, rappel et *F*-score
  - **Frontières thématiques**
    - Combinaison avec une méthode non paramètrique (test de Kendall)
    - Extraction par fenêtre glissante

Perspectives

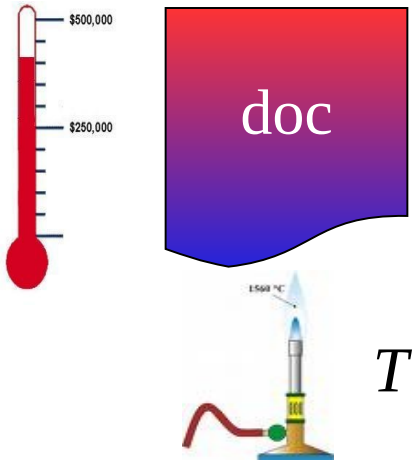
- Résumés guidés par des requêtes
- Multilinguisme
- Améliorer la détection des frontières
- Restructurer des paragraphes?
- Mesure de similitude… évaluer des systèmes produisant du langage naturel?

Perspectives *exotiques*

Champ externe



Température



$p(S_i = \pm 1) = 1/(1+e^{(2\beta h i)})$

$\beta = 1/k_B T$ = température inverse

k_B = cte de Boltzmann

⇒ Modifier le paysage d'énergie

References

- S Fernández, E SanJuan, J-M Torres-Moreno, Textual Energy of Associative Memories: Performant Applications of Enertex Algorithm in Text Summarization and Topic Segmentation, MICAI 2007: pp 861-871
- I da Cunha, S Fernández, P Velázquez, J Vivaldi, E SanJuan, J-M Torres-Moreno, A New Hybrid Summarizer Based on Vector Space Model, Statistical Physics and Linguistics, MICAI 2007: MICAI 2007: Advances in Artificial Intelligence pp 872-882
- S Fernandez, E SanJuan, JM Torres-Moreno, Énergie textuelle de mémoires associatives, Traitement Automatique des Langues Naturelles, 25-34