

M2SOL034 Corpus, ressources et linguistique outillée

TD 2 : TXM I

Ljudmila PETKOVIĆ

Sorbonne Université

Master « Langue et Informatique » (M1 ScLan)

UFR Sociologie et Informatique pour les Sciences Humaines

Semestre 2, 2024-2025, le 7 février 2025

Table des matières

1 Exercices : commandes de base	1
2 Solutions	2

1 Exercices : commandes de base

1. Après avoir installé TXM, chargez ¹ le corpus VOEUX déjà importé dans TXM, et accédez aux propriétés du corpus.
Indiquez (a) l'année la plus ancienne et (b) l'année la plus récente de l'édition.
2. Téléchargez le corpus sous format XML-TEI `charcot.xml` depuis Moodle (*cf.* le répertoire `ressources_TD2`), et importez-le dans TXM.
En une seule requête, affichez la liste des mots-pivots dérivés du mot `hystérie`.
3. Téléchargez le corpus *Du côté de chez Swann* de Marcel Proust sous format texte brut depuis le site du projet Gutenberg <https://www.gutenberg.org/ebooks/2650>, et importez-le dans TXM. Exportez le lexique du corpus importé dans un tableur. Que pouvez-vous constater concernant la répartition des fréquences de mots?

1. charger : corpus déjà importé dans TXM auparavant ; importer : corpus brut (`txt`, `XML`, voire en provenance du presse-papier).

4. Trouver en une seule recherche les fréquences des mots-pivots dérivés du mot « patrie » à partir du corpus **VOEUX**. Quels mots avez-vous extraits ?
5. Trouver en une seule recherche le souhait de bonne année de chaque Président dans le corpus **VOEUX**.
6. Affichez la liste des cooccurents du terme **hystérie**, avec leurs fréquences, leurs indices de spécificité et leurs distances moyennes à partir du corpus **CHARCOT**.
Quel est l'indice de spécificité et la distance moyenne du cooccurrent **convulsive** ?
Comment interprétez-vous ces mesures ?

2 Solutions

1. (a) l'année la plus ancienne : 1959
(b) l'année la plus récente : 2012
2. Nous obtenons une liste des mots-pivots dérivés du mot **hystérie** à partir du corpus **CHARCOT** :

Requête	Contenu gauche	Pivot	Contenu droit
CL 000001 001 hystérie	les formes de l'hystérie et de l'	hystérie	grave. Ce n'est pas tout. M. le directeur
CL 000001 001 hystérie	dans certains cas d'angoisses, chez les	hystériques	sur exemple, une ischémie car-telle très onéreuse et très oné
CL 000001 001 hystérie	seulement que des individus sont f	hystérie	il faut dans certains cas, mais, le plus souvent,
CL 000001 001 hystérie	NEUVEME LECON De l'hystérie	hystérie	Sommeil. — Préambule. — De l'hystérie hystérie. —
CL 000001 001 hystérie	Sommeil. — Préambule. — De l'hystérie	hystérie	— Différence ou la même chose de l'hystérie. — Considérations
CL 000001 001 hystérie	l'est suspecter la stabilité de l'hystérie	hystérie	— Distinction entre l'hystérie calcaireuse et l'hystérie hystérie
CL 000001 001 hystérie	Distinction entre l'hystérie calcaireuse et l'hystérie	hystérie	Observation. — Paradoxe et contradiction hystérie. —
CL 000001 001 hystérie	ischémie hystérie. — Observation. — Paradoxe et contradiction	hystérie	— Hystérie et automatisme hystérie. —
CL 000001 001 hystérie	trismus. — Acquisition de l'hystérie	hystérie	— Précautions prises pour éviter toute cause d'erreur. —
CL 000001 001 hystérie	Suspension des accidents. Retour de l'hystérie	hystérie	— Nouvelle méthode de l'hystérie. — Grande des l'hystérie
CL 000001 001 hystérie	l'hystérie calcaireuse. — Retour de l'hystérie	hystérie	— L'hystérie de l'hystérie est en rapport avec la dose
CL 000001 001 hystérie	ischémie de l'hystérie. — Réaction de l'hystérie	hystérie	à l'hystérie. — Mécanisme de l'hystérie hystérie. —
CL 000001 001 hystérie	à l'hystérie. Mécanisme de l'hystérie	hystérie	— Insuffisance de nos connaissances à cet égard. —
CL 000001 001 hystérie	un certain nombre de cas très remarquables d'	hystérie	qui se trouvent également dans nos salles. Il importe
CL 000001 001 hystérie	hystérie. —	hystérie	que le vous vous sachiez. Dès l'abord, se doit
CL 000001 001 hystérie	de la ve-de. fait vider chez les	hystérie	sciences, est due hystérie. Il ne s'agit pas
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	On sait que très communément, en certaine circonstance, se
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	est observée hystérie commode de hystérie. —
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	et oui, du reste, comme l'a fait remarquer avec
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	à son maximum de développement, à l'état de surméditation
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	nous offrait, dans l'hystérie humaine, la reproduction plus ou
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	la hystérie entre l'hystérie et l'hystérie hystérie
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	au moins dans ce qu'elle a d'essentiel. —
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	est mentionnée, ce n'est qu'en cessant, à titre
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	même dans les deux comités et hystérie hystérie. —
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	est mentionnée, ce n'est qu'en cessant, à titre
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	certains d'hystérie sur les enseignements de mon maître. —
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	se rendent courables. Et l'hystérie a à confondre avec
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	hystérie que l'hystérie a l'intention d'être pour le retour au
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	est un phénomène rare, du moins sous sa forme très accident
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	et l'hystérie hystérie hystérie à l'hystérie. —
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	Je ne le crois pas. Messieurs, et l'hystérie
CL 000001 001 hystérie	dans la ve-de. fait vider chez les	hystérie	la succession d'ordre pour que ou hystérie se

3. Le lexique du corpus importé montre une distribution des fréquences lexicales évoquant la loi de Zipf, selon laquelle la fréquence d'un mot est inversement proportionnelle à son rang dans la liste globale des mots après le tri par ordre décroissant de fréquence. Autrement dit, peu de mots apparaissent très souvent, tandis que la majorité des mots apparaissent rarement.
4. Afin de récupérer les fréquences des mots-pivots dérivés du mot « patrie », nous utilisons l'expression régulière `.*patri.*`. Cela nous permet d'extraire les mots *patrie*, *patriote*, *patriotisme*, *compatriotes* et *rapatriés*.
5. `[frlemma="je"] []* [frlemma="souhaiter"] []* [frlemma="année"] within s`

6. L'indice statistique du cooccurrent **convulsive** est 43 et la distance moyenne est 2.1.

Concernant l'indice de spécificité, plus il est élevé, plus la cooccurrence est remarquable.

Pour ce qui est de la distance moyenne, elle indique le nombre de mots entre le cooccurrent et le pivot quand ils sont dans le même voisinage.