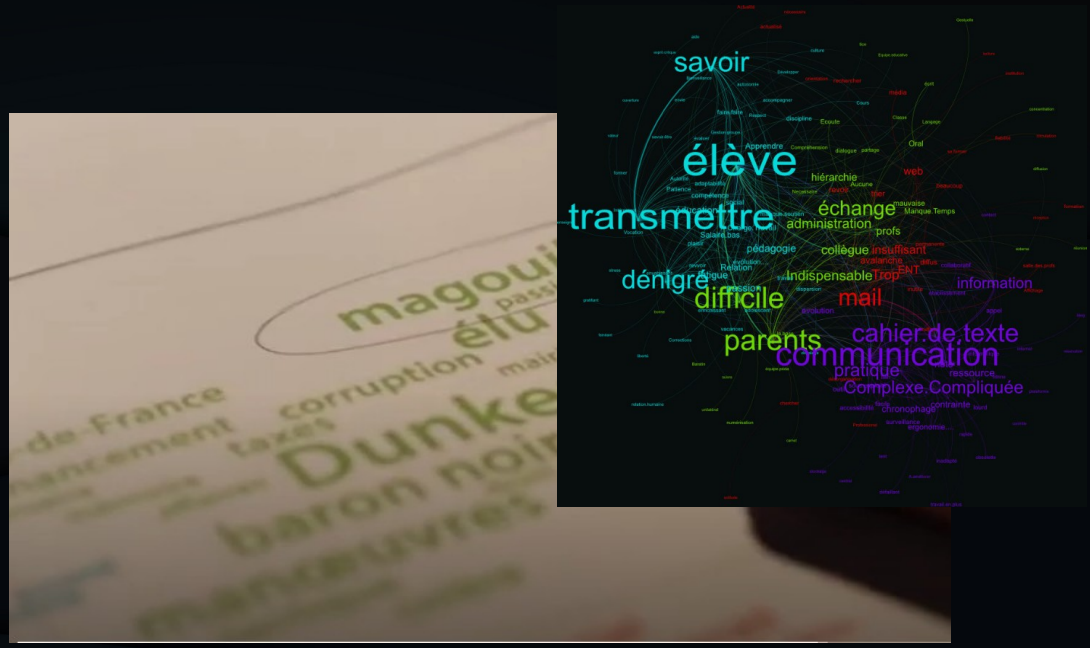
Présentation Iramuteq



Lucie Loubère

2024

Lucie Loubère - 2024

Tour de table : quelle image représente le plus/mieux votre étude ?



Le logiciel Iramuteq

- Interface pour R d'Analyses Multidimensionnelles de Texte et de Questionnaires.
- Développé par P. Ratinaud (Lerass) dans le cadre du LabEx SMS (Toulouse)
- Logiciel libre et gratuit développé sur la base de logiciels libres
  - Python http://www.python.org
  - R (R Development Core Team, 2009): http://r-project.org
  - Lexique 3 (New, Pallier & Ferrand, 2005): http://lexique.org

Le corpus… point de vue technique

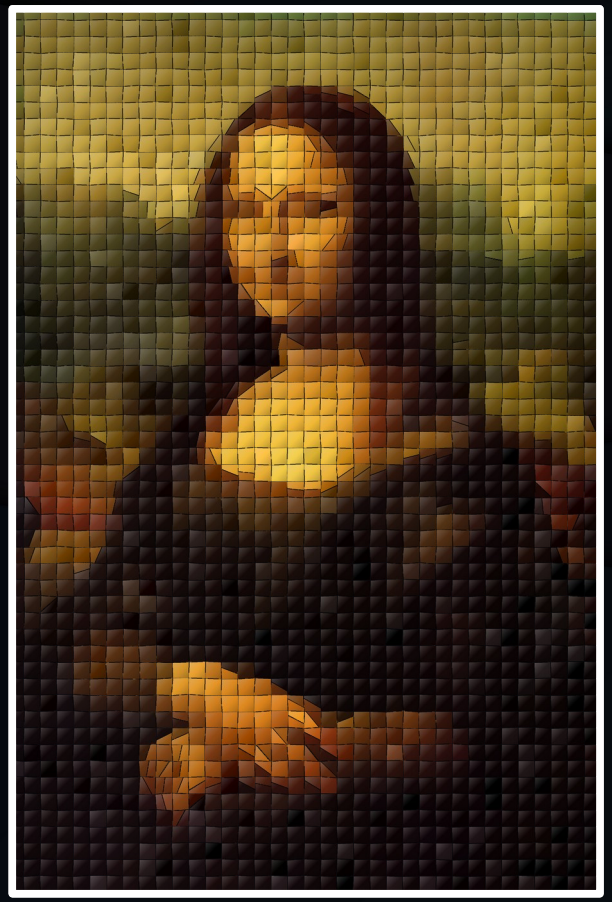
- Vu comme un ensemble
  - 1 fichier .txt
  - Délimitation de chaque texte par une « ligne étoilée »

    **** *variable_modalité *genre_F *region_occi
- Pourra être découpé une fois ouvert dans iramuteq par:
  - Modalité d’une variable,
  - Thématique,
  - Concordancier d’un ou plusieurs mots,

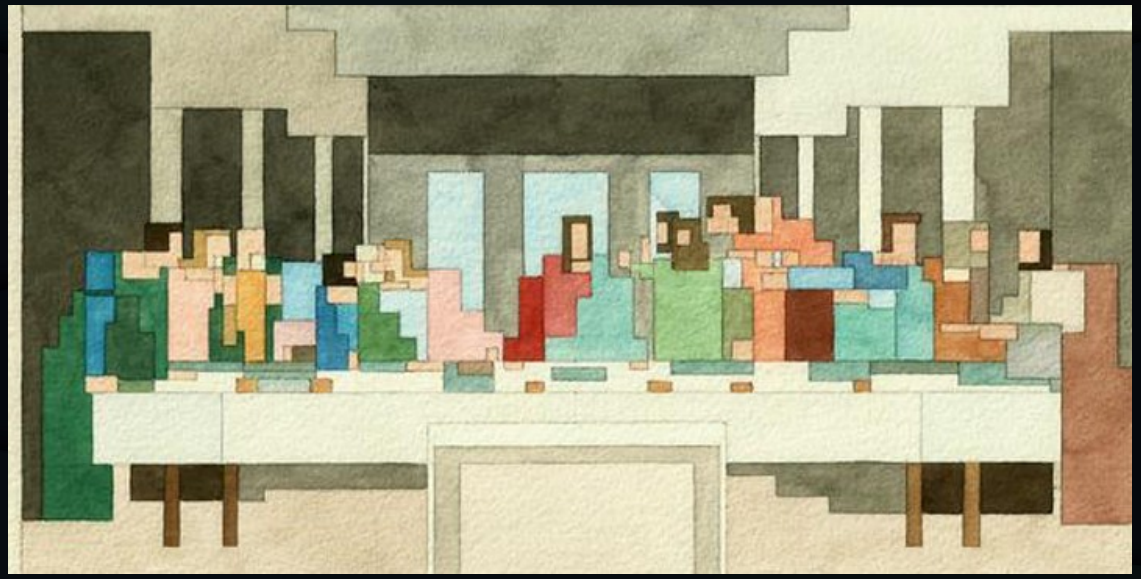
Pour quels types de corpus*…

- Des discours longs (livre, articles, entretiens...)
- Des discours brefs (tweets, questions ouvertes, focus-groups...)
- Des corpus* de toute taille :
  - Quelques entretiens, un discours
  - Grand débat national (167 853 085 occurrences)
    - 111 x *À la recherche du temps perdu*
    - 39 000 x le l'allocution d'E.Macron du 9 novembre 2021

Entrons dans la lexicométrie



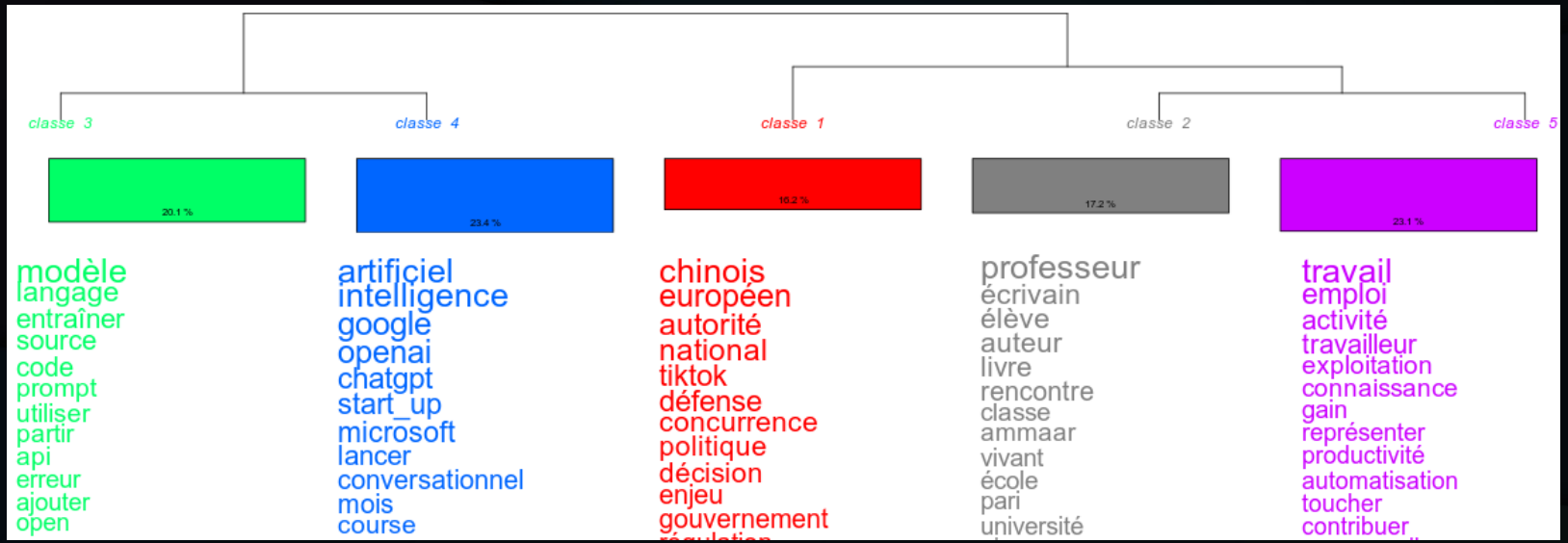
Pixellisation de la Joconde (Léonard de Vinci)



Variation de La cène (Léonard de Vinci) par Adam Lister

Lucie Loubère - 2024

Un exemple sur une thématique

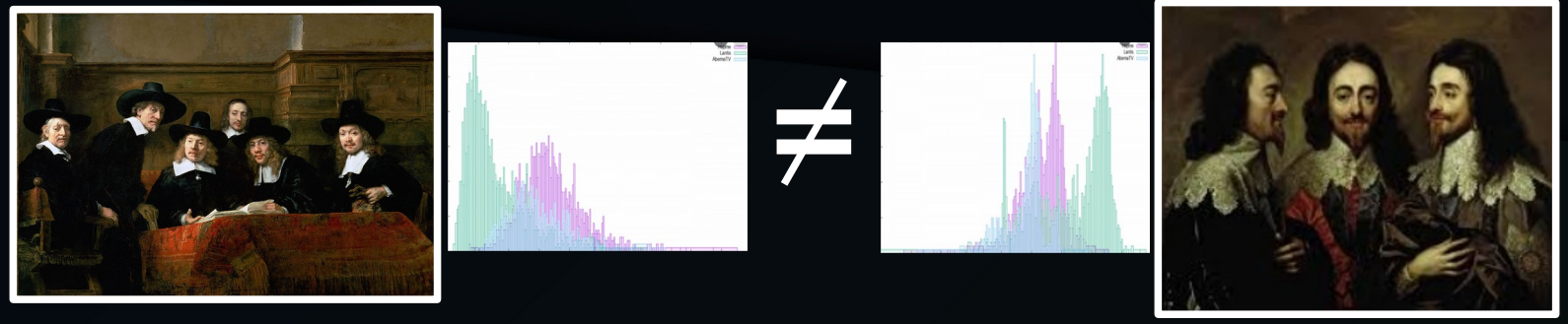


Lebart & Salem (1994)

Supposons (...) que
l'on étudie les histogrammes des **longueurs d'ondes** correspondant aux couleurs d'un tableau de Rembrandt (pour chacun des pixels d'une reproduction).

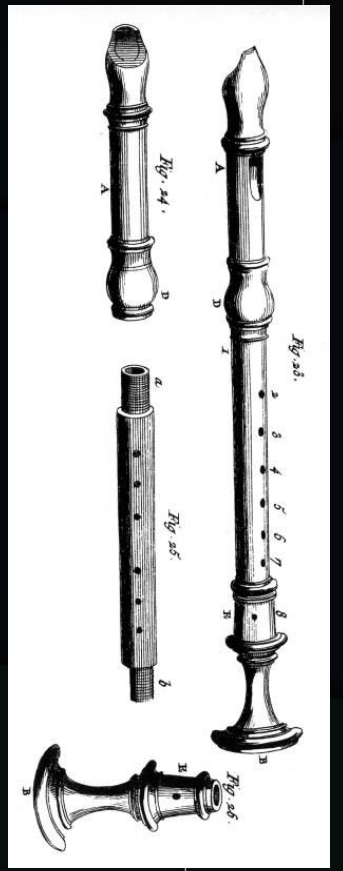
Il va de soi que l'on utilise une **fraction dérisoire** de l'information contenue dans l'image d'origine.

Il est cependant possible que la forme de l'histogramme (ou d'une fonction plus élaborée des mêmes mesures et données de base) **permette de distinguer** un Rembrandt d'un Rubens ou d'un Van Dyck » (p.21).

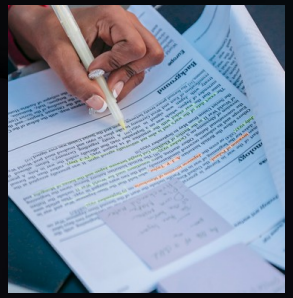


Le Syndic de la guilde des drapiers (Rembrandt)

Triple portrait de Charles 1er (Van Dyck)



Analyses données textuelle



corpus

Grille(s) d'analyse
interprétations

Commentaires

Lexicométrie

Codage

Interprétations





Réorganisations
Mesures
Statistiques
Autres...



résultats

Lucie Loubère - 2024

Pour synthétiser… lexico = statistiques

Les questions que se donne la statistique lexicale sont les suivantes :

« quels sont les textes les plus **semblables** en ce qui concerne le **vocabulaire** et la **fréquence** des formes utilisées ?

Quelles sont les formes qui **caractérisent** chaque texte, par leur présence ou leur absence ? » (Lebart & Salem, 1994, p.135).

Pour quels types d’analyses ?

- Comparaisons lexicales de sous ensembles de textes :
  - Analyses de spécificités
  - AFC
  - Distances de Labbé
- Exploration des mondes lexicaux (Reinert, 1983)
- Exploration de la structure du texte (Analyses de similitudes)

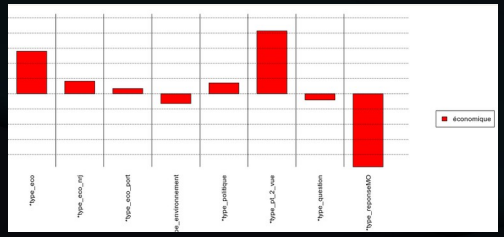
Les comparaisons lexicales (1/2)

- Tableau lexical croisant lemmes (racine grammaticale des mots) et modalités de variables.

	*source_la_croix	*source_le_figaro	*source_le_monde	*source_lhumanité	*source_libération
sécurité	15	46	28	43	15
devenir	33	27	35	17	40
établissement	1	7	12	4	3
tirer	13	4	14	6	12
pencher	0	4	3	5	2
brexit	5	7	4	0	3
reporter	3	5	4	0	0
évolution	4	13	7	3	4
chine	1	5	2	0	2
connaissance	1	2	7	5	1
naturel	0	2	3	1	4
attente	4	5	8	3	3
unir	11	9	15	5	11

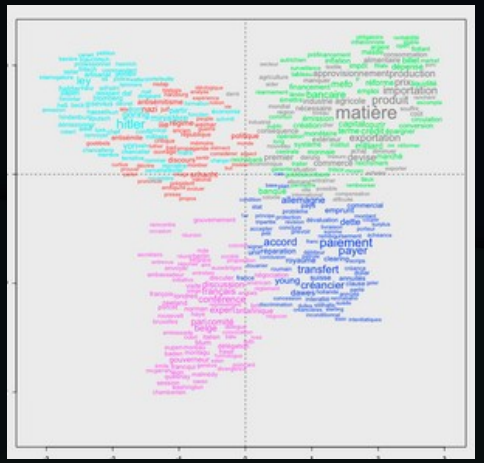
Les comparaisons lexicales (2/2)

Calculs de spécificités :
quel est le vocabulaire spécifique d'une catégorie de textes et/ou répondants ?



Forms	Banal forms	POS	Forms frequencies	POS frequencies	Forms relative frequencies	
formes		*profil_débutant	*profil_débutant	*profil_en_deprise	*profil_engagé	
cancer		12.0343	-0.2663	-5.4833	-1.5261	
sein		11.5894	-0.5908	-3.558	-2.0522	
finalement		9.1673	0.3143	-5.741	-0.8231	
jambe		6.5059	-0.5078	-1.656	-1.2378	
essai		6.4357	-0.7272	-1.6917	-1.0333	
examen		6.1439	-0.4557	-2.6564	-0.5465	
médecin		6.1332	-1.0216	-2.9837	0.3451	
solution		5.7128	-0.5078	-1.656	-0.9198	
traiter		5.3219	-0.3082	-4.6941	0.4895	
attente		4.7981	-0.4336	-1.1439	-1.1039	
irm		4.3342	-0.7272	-1.6917	-0.3629	
au_delà		3.9662	-0.6362	-0.5371	-1.3564	
parkinson		3.7928	1.0671	-5.5307	0.8257	

Analyses Factorielles de Correspondances :
Représentation graphique de la dispersion du vocabulaire des différentes catégories de textes et/ou répondants



Lucie Loubère - 2024

Exploration des mondes lexicaux (1/2)

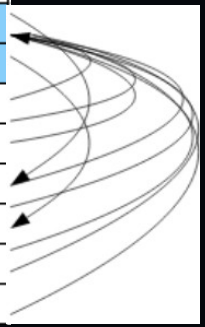
- Classification Reinert
  - Classification hiérarchique descendante
  - Permet à partir d'un ensemble de texte
    - de les scinder
    - regrouper les passages qui se ressemblent sur la présence ET l'absence de mots (forment une classe de discours)
    - Afficher le lexique sur-représenté dans chaque classe de discours
    - Afficher la répartition de chaque modalité (variable étoilée) dans chaque classe
    - Permet d'explorer le texte par ces/ses thématiques

Exploration des mondes lexicaux (2/3)

Classe 1

Classe 2

	éolien	parc	mer	projet	énergie	impact	public	zone
Segment Texte 1	0	0	1	1	0	0	1	1
Segment Texte 7	0	0	1	1	0	0	1	0
Segment Texte 30	1	1	1	0	0	1	0	0
Segment Texte 4	1	0	0	1	1	1	0	0
Segment Texte 12	1	1	0	0	1	1	0	0
Segment Texte 6	1	1	1	0	0	1	0	0
Segment Texte 65	1	0	0	1	0	1	0	0
Segment Texte 22	1	1	1	1	0	0	1	1

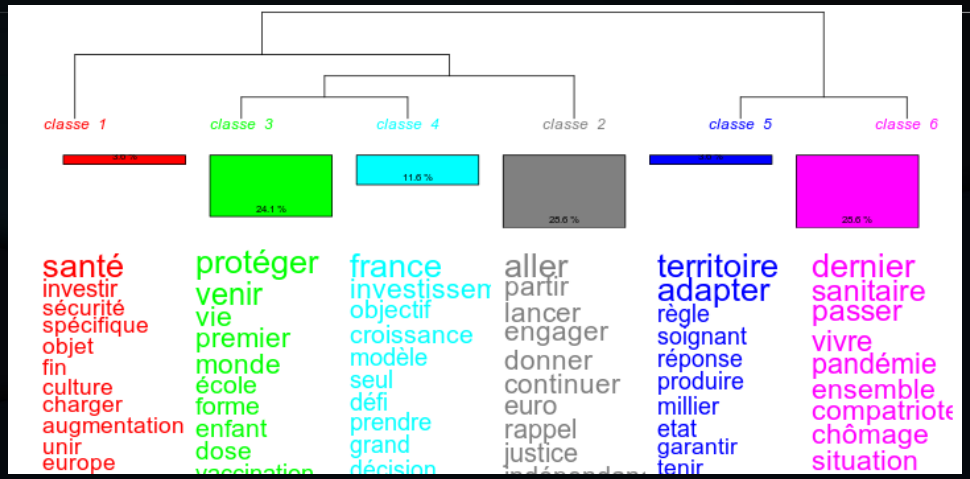


Lucie Loubère - 2024

Exploration des mondes lexicaux (3/3)

1 Classe 1 8045/94858 8.48%	2 Classe 2 699/94858 0.74%	3 Classe 3 3063/94858 3.23%	4 Classe 4 6479/94858 6.83%	5 Classe 5 6677/94858 7.04%	6 Classe 6 4408/94858 4.65%	7 Classe 7 1648/94858 1.74%

num	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	834	2344	35.58	2273.9	nom	chose	< 0,0001
1	387	773	50.06	1736.26	nom	gens	< 0,0001
2	724	2219	32.63	1706.76	ver	penser	< 0,0001
3	530	1631	32.5	1233.0	ver	entendre	< 0,0001
4	1196	6640	18.01	835.59	ver	voir	< 0,0001
5	393	1298	30.28	805.49	adv	vraiment	< 0,0001
6	469	1969	23.82	609.45	ver	travailler	< 0,0001
7	256	787	32.53	591.25	adj	intéressant	< 0,0001
8	340	1292	26.32	536.77	nom	monde	< 0,0001
9	532	2540	20.94	522.35	ver	venir	< 0,0001
10	237	785	30.19	480.66	adv	simplement	< 0,0001
11	226	737	30.66	470.94	nom	idée	< 0,0001
488	2710	15836	17.11	1824.79	pro_p...	je	< 0,0001
489	3599	26591	13.53	1215.71	pro_rel	que	< 0,0001
490	2764	19043	14.51	1117.43	adv_s...	pas	< 0,0001
491	2988	21193	14.1	1109.66	adv_s...	ne	< 0,0001
492	2999	22708	13.21	858.99	pro_p...	on	< 0,0001
493	942	4702	20.03	850.71	pro_p...	j	< 0,0001
494	312	938	33.26	749.55	adj_ind	quelque	< 0,0001
495	2317	16781	13.81	745.14	pro_d...	c	< 0,0001
496	4715	42624	11.06	664.21	aux	avoir	< 0,0001
663	7040	59904	11.75	2241.0		*source_reu	< 0,0001
664	919	6055	15.18	373.67		*db_DieppeTreport	< 0,0001
665	1241	9742	12.74	253.55		*db_SaintNazaire	< 0,0001
666	182	808	22.52	207.07		*3_Président_de_...	< 0,0001
667	319	1871	17.05	180.54		*3_Présidente_de...	< 0,0001
668	293	1709	17.14	168.29		*2_Claude_BRÉVAN	< 0,0001



Segments de texte caractéristiques - Classe 1

****** *db_SaintNazaire *source_reu *2_Béatrice_BUFFON**
***3_directrice_adjointe_EDF_Énergies_Nouvelles_**

score : 6279.29

en revanche j **entends** bien votre **proposition** et je **pense** qu on l a tous **entendue** et c est quelque **chose** à laquelle on pourrait **réfléchir** à la suite et ça pourrait faire partie **effectivement** de **propositions** pour la suite du **débat** concernant le croisic spécifiquement monsieur gaëtan duchene

****** *db_DieppeTreport *source_reu *2_Marie-Thérèse_CASTAY_**

score : 6178.83

pour moi cela a été quelque **chose** de très **important** cette possibilité de **rencontrer** des **gens** qui permettait de **réfléchir** de **façon** un peu plus vaste que ce que j en **pensais** sur l éolien en mer et ses conséquences merci

****** *db_DieppeTreport *source_reu *2_Yannick_POURCHAUX**
***3_Président_du_Comité_régional_des_Pêches_de_Haute-Normandie**

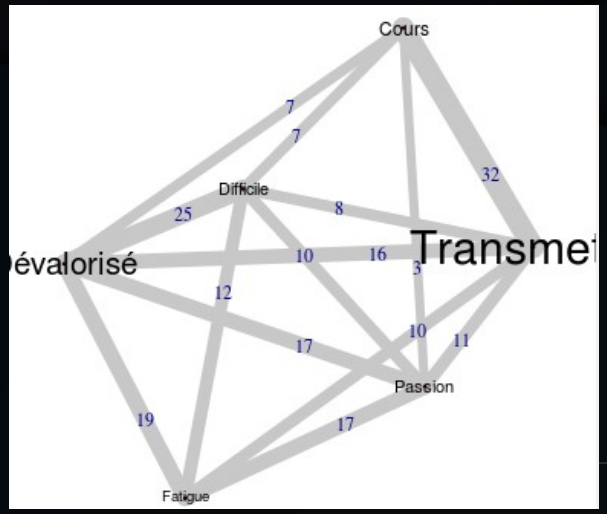
score : 6144.27

marchons nous sur la tête a un **moment donné** il faudra peut_être mettre tout cela en pratique je suis un **spécialiste** de mon secteur mais pas de ce que j ai **entendu** jusque là et des **gens** ont des **choses** très **intéressantes** à dire

Lucie Loubère - 2024

Structure du discours

- Analyse de similitude :
  - Issue de la théorie des graphes (Flament, 1962 ; Flament, 1981 ; Vergès & Bouriche, 2001)
  - Permet de visualiser les relations entre les éléments
  - Exemple avec l'indice « nombre d'occurrences »



Transmettre et *cours* se retrouvent ensemble dans 32 segments de textes

Fatigue et *difficile* dans 12 questionnaires

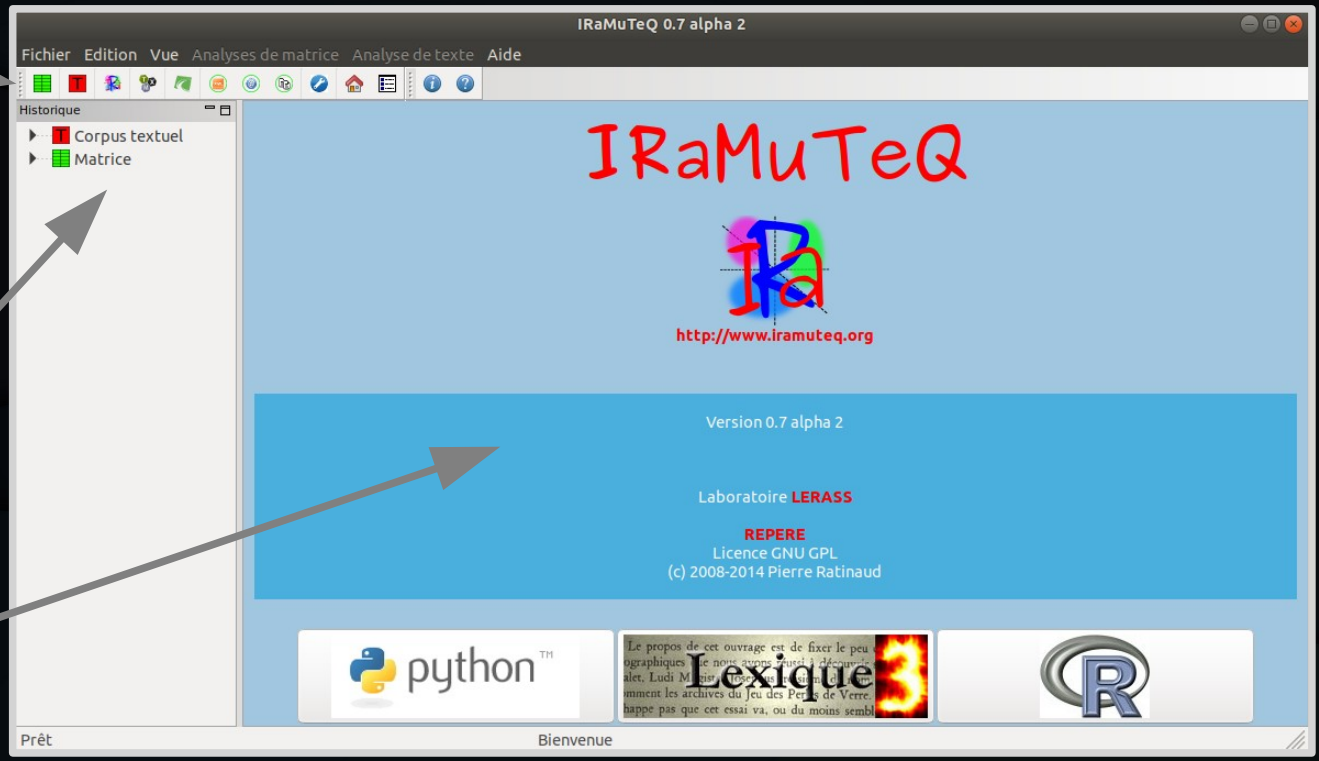
Transmettre et *passion* 11 fois...

Iramuteq : l'interface du logiciel

Menu

Historique
des corpus et analyses

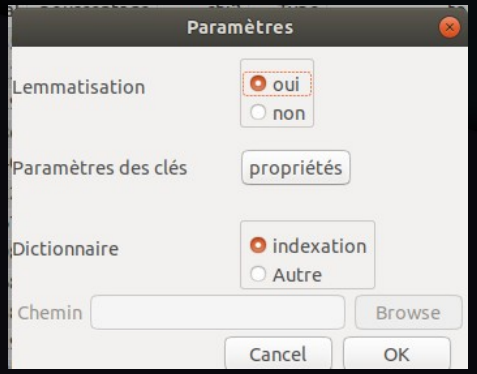
Résultats

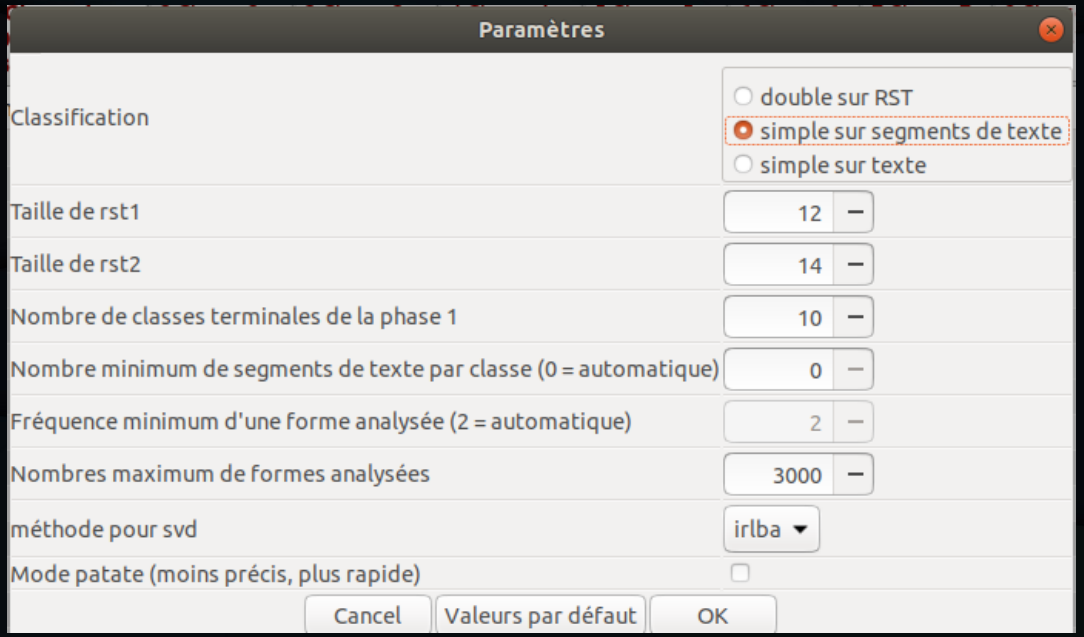


L’ouverture du corpus

- Corpus .txt formaté (ligne étoilée)
- À l’ouverture :
  - Segmentation (par défaut)
  - Nettoyage (par défaut)
  - Indexation avec un dictionnaire (français par défaut)

La classification Reinert : paramètres





La classification Reinert : résultats

Nombre de ST contenant cette forme dans cette classe

Nombre de ST contenant cette forme dans la classification

Pourcentage de ST contenant cette forme dans cette classe

Chi2 de corrélation entre la forme et la classe

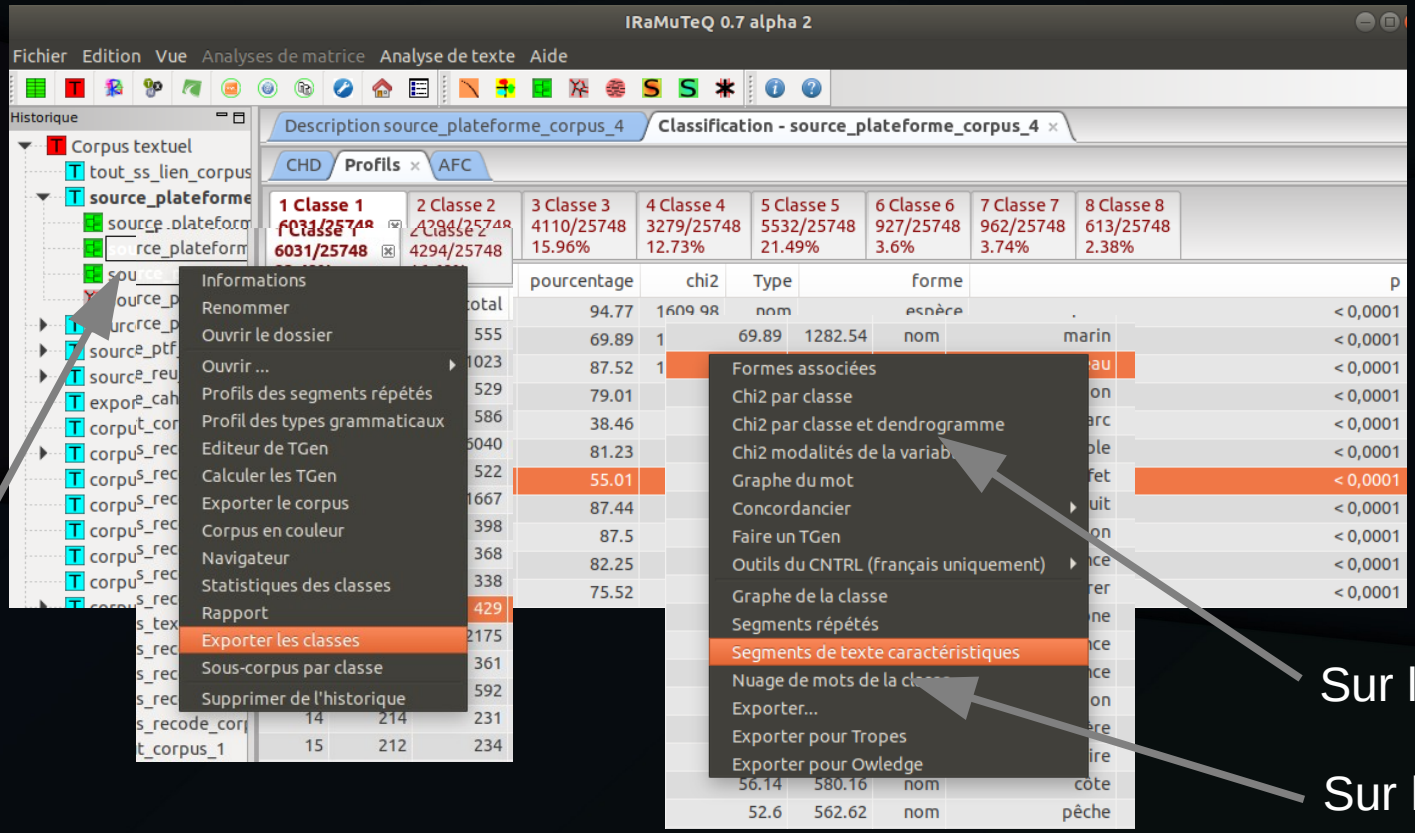
CHD | Profils | AFC

1 Classe 1 6031/25748 23.42%	2 Classe 2 4294/25748 16.68%	3 Classe 3 4110/25748 15.96%	4 Classe 4 3279/25748 12.73%	5 Classe 5 5532/25748 21.49%	6 Classe 6 927/25748 3.6%	7 Classe 7 962/25748 3.74%	8 Classe 8 613/25748 2.38%

num	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	526	555	94.77	1609.98	nom	espèce	< 0,0001
1	715	1023	69.89	1282.54	nom	marin	< 0,0001
2	463	529	87.52	1237.23	nom	oiseau	< 0,0001
3	463	586	79.01	1033.0	nom	fondation	< 0,0001
4	2323	6040	38.46	994.77	nom	parc	< 0,0001
5	424	522	81.23	992.48	nom	câble	< 0,0001
6	917	1667	55.01	991.39	nom	effet	< 0,0001
7	348	398	87.44	923.54	nom	bruit	< 0,0001
8	322	368	87.5	854.59	nom	poisson	< 0,0001
9	278	338	82.25	660.76	nom	présence	< 0,0001
10	324	429	75.52	660.25	ver	montrer	< 0,0001
11	994	2175	45.7	657.35	nom	zone	< 0,0001
12	287	361	79.5	641.93	nom	distance	< 0,0001

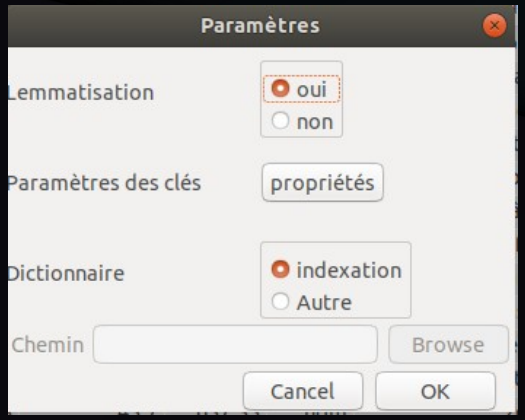
Lucie Loubère - 2024

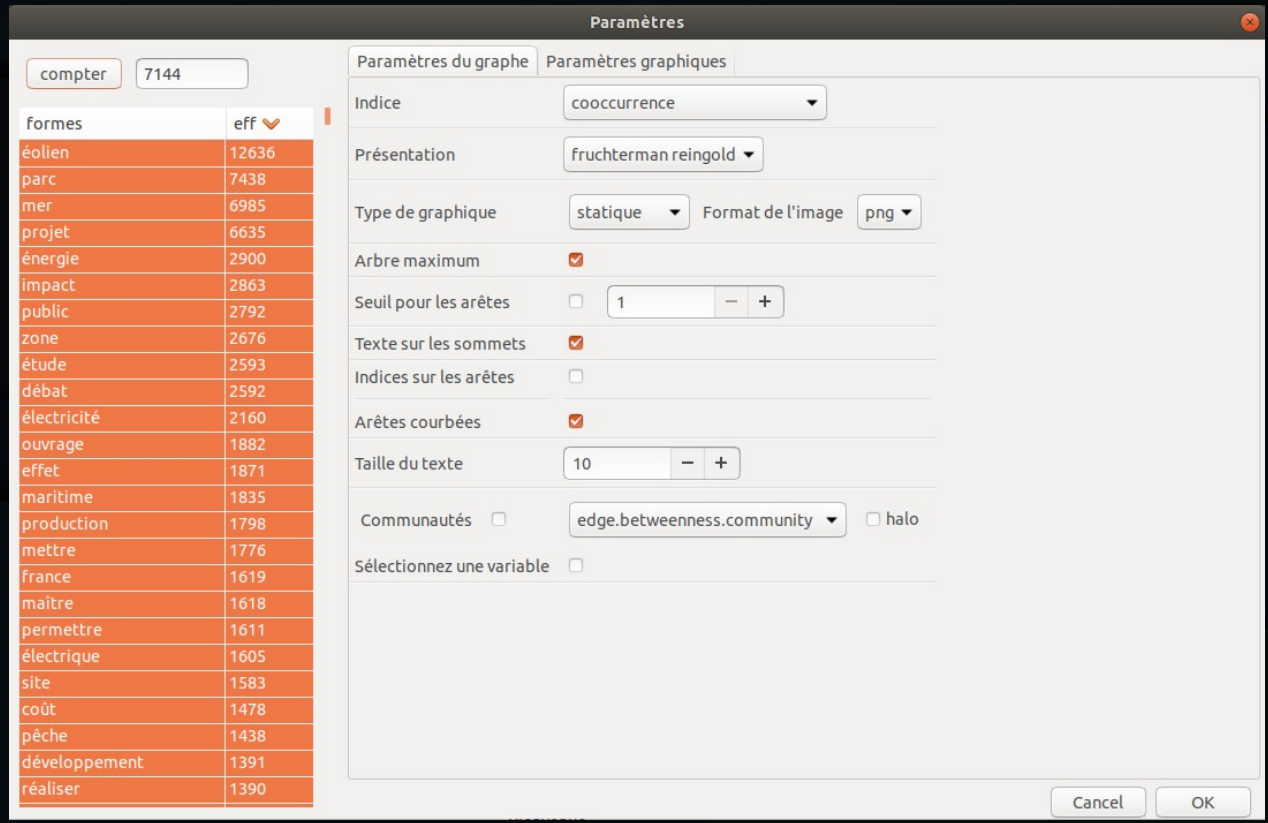
La classification Reinert : options



CC BY NC ND

Analyses de similitudes





Analyses de spécificités

