**Histoire et Informatique TD 1**

Plutôt base de données, avec notamment ma liste de prisonniers de guerre, pour permettre d’éviter le manuel

Mail : [gaetan.bonnot@univ-paris1.fr](mailto:gaetan.bonnot@univ-paris1.fr)

PIREH

14h sur le semestre 1 (7x2H) – deux méthodes à voir – base de donnés et textométrie

D’abord base de données, on va apprendre à modéliser, faire de la logique, passer de sources à de la donnée exploitable pour l’historien, puis implémentation, mettre dans le logiciel, pas de compétence informatique à avoir, logiciel simple à utiliser et enfin, apprendre à faire des requêtes, pour permettre l’exploitation, langage informatique, SQL, pas langage de programmation, trentaine ou quarantaine de mots, initiation.

Lexicométrie – faire un corpus, voir comment on fait, puis exemples d’analyse lexicale, et enfin, statistiques, expliquer les méthodes par les chiffres, pour amener après à l’analyse multidimensionnelles.

CM de Stéphane Lamassé, jeudi soir de 18h à 20h

S2 🡪 suivi pas exactement individuelle, sous forme de permanence, environ 60 heures, deux demi-journées par semaine, problème possiblement lié à l’informatique.

Logiciel 🡪 libre, gratuit et ouvert. Base de données = Libre office (mettre à jour Java au préalable). Textométrie = lexico 3 (pour sources non occidentales, TXM), mettre Wine pour les macs sinon lexico 3 ne fonctionnera pas.

Validation 🡪 Rendre un dossier, 12-15 pages environ, présentera notre sujet de manière succincte (1-2 pages), parler de nos sources, puis parler de notre modélisation ou projet de modélisation, dessiner base de données ou parler de notre corpus, et enfin, quelques résultats et pistes de résultats 🡪 Rendu entre janvier (si volonté de continuer pour qu’il corrige et améliore) et avril (si juste volonté de valider la matière) 2017.

Le début de l’utilisation des chiffres en histoire se fait avec les annales en 1930 mais dès le début de 1900, cela apparaît.

Voir annotations dans la bibliographie indicative.

Une base de données, c’est comme plusieurs tableaux Excel liées les uns avec les autres. Trois tableaux faits par l’étudiant, information est atomisée, le moins possible dans la case, saisie propre, numéro d’identification donc aucun doublon, atout par rapport à Excel, utilité des bases de données par rapport à la mono table permet d’éliminer les doublons

Chaque tableau est une entité – une entité est une unité de sens, une identité du monde réel.

Chaque entité est caractérisé par des attributs – on lui attribue plusieurs caractéristiques

L’identifiant est la clé primaire, le rappel de la clé primaire dans une autre table est ce qu’on appelle une clé étrangère, clé primaire dans une autre table.

Ces entités sont reliées entre elles, ce sont des associations, elles doivent normalement être sémantisées, permet de ne pas faire de contresens.

Les cardinalités sont des minimums et des maximums théoriques de fois qu’une ligne d’un tableau sera concernée par une association. 3 cardinalités possibles 🡪 1/1, 1/n et n/n (1 livre peut être écrit par plusieurs auteurs et un auteur peut écrire plusieurs livres).

Quelques règles à respecter, si fait, toute modélisation est correcte, cela est bon ou mauvais par rapport à notre sujet mais jamais faux si règle suivi.

Les requêtes permettent de produire des tableaux, base de données relationnelles

Textométrie/Lexicométrie 🡪 nécessité d’avoir des textes, numérisées, pas de textométrie sur situations et extraits. 40 000 mots pour le corpus minimum. La lexicométrie sert à avoir des contrastes dans le corpus, permet aussi de voir les mots les pus souvent associés, joué sur les échelles de mots et de groupes de mots.

La division de l’occurrence avec la forme (ou inverse) permet de voir al richesse de vocabulaire de l’auteur

Courbe de Pareto 🡪 Nombre de fois qu’un même mot apparait et nombre de formes

EN rouge et bleu, apparaissent les mots surreprésentés et sous-représentés. Le corpus est structuré par le temps, l’auteur ou autre

Sur EPI 🡪 activité, modélisation, essayer d’y réfléchir si on veut, acticité faite la semaine prochaine mais permet de s’y préparer.

TD2

PARC\inviteLS

MDP : Pr0v1soire

Base de données, 4 étapes, euristique et non documentaire

1 – Sélectionner l’information, Dictionnaire de données

2 – Entité, unités du réel, décrite par des attributs, variable attaché aux entités, cela constitue les tables, unique grâce à un identifiant

3 – Association sémantisée, on attache un sens à chacune des associations

4 – Cardinalités, un individu dans sa carrière peut avoir n postes

S’intéresser aux personnes, mais aussi possible de s’intéresser aux lois

Un représentant est un individu et un mandat

Table Individu = ID i, nom, sexe, date de naissance, lieu de naissance

Table mandat = ID m, type/électif, circonscription, date de début, date de fin, % de voix, ID i (clé étrangère)

Légende sera occupe, la cardinalité sera 1 individu fait n mandat mais un mandat c’est par 1 individu, un mandat propose n lois, une loi est proposé par 1 seul mandat, mais possible de regrouper en commission, création d’une table commission

Table loi = ID l, date, ID c, voix pour, voix contre (issue possible aussi), domaine

Table commission (relation de n à n avec mandat) = ID c,

Table vote = ID v, ID m, ID l, Voix

Création d’une relation entre des unités du réel, table de liaison où on met les clés étrangères

Initiation au SQL, demander des tableaux a partir de plusieurs tables

Select Nom\_instituteur, Sexe\_instituteur

From TBL\_instituteur

Where (Sexe\_instituteur=’F’) and (Nationalite\_instituteur=’Turque’ or Nationalite\_instituteur=’Française’)

Ou Where (Sexe\_instituteur=’F’) and Nationalite\_instituteur IN (’Turque’,’Française’)

Les parenthèses permettent de bien délimiter, sinon les hommes apparaitront aussi

Select Nom\_instituteur, Annee\_naissance\_instituteur

From TBL\_instituteur

Where Annee\_naissance\_instituteur between ‘1870’ and ‘1890’

Order by Annee\_naissance\_instituteur

Select a.Nom\_I, count (b.ID\_poste) as Effectif

From TBL\_i as a, TBL\_poste as b,

Where a.ID\_instituteur=b.ID\_Instituteur\_poste

Group by a.Nom\_i

Order by Effectif desc (= décroissant)

Order by Nom\_intitueur

(les deux sont possibles)

Existe aussi sum, max, min, avg

Compte la ou tel poste a bien été fait par tel instituteur

LibreOfficeBase Type = integer mettre aussi auto valeur oui

Isoler année, mois, jour car si pas renseigné peut être compliqué

Sexe = boolean

Année = numeric

Saisie requise garder non car sinon obligé de rien avoir, possible de jouer sur la longueur, avec les caractères

Mettre la clé étrangère dans l’autre table, rappeler les individus dans la table poste.

Ne pas utiliser accents, plus, slash et autres signes, et pas d’espaces

Mentionner le type aussi, varchar (alphabétique), booléan (oui/non, deux choix), date, integer, decimal

O

Clé primaire = integer en auto valeur oui

Date pose des problèmes, parce que soit trop, soit pas assez d’infos. Deux solutions alors, séparer jour du mois et de l’année (entier naturel alors)

Ou écris ainsi AAAAMMJJ et si on connaît pas mettre des 00

Base de données = être au plus prêt des base de données

Formulaire = autre manière de formaliser de la donnée, table = manière de représenter la donnée  
le remplissage du tableau est une sorte de SQL

Le formulaire va permettre de gérer toutes les clés en même temps

NAVIGATEUR DE FORMULAIRE est le bouton le plus important dans formulaire, fait la liaison entre le formulaire et les tables, liens avec la donnée structurelle

Mentionner dans Propriétés et données de formulaires que table individu a prendre

Prendre ajouter champ pour ajouter les champs

Mettre barre de navigation avec contrôle supplémentaire sur le coté gauche

Créer ensuite le sous formulaire comme un sous dossier dans le formulaire contenant également tous les autres champs

Lier champs maitre et esclaves, trois petits points et suggérer

Garder les identifiants permet de contrôler Rajouter barre de navigation pour le deuxième formulaire et bien le mettre dans le sous formulaire

Après remplir les formulaires sans toucher aux ID poste et ID identifiant car auto champs font les choses seules

APPUYER SUR F5 POUR AVOIR LA TABLE

Pas possible de faire de la double entrée dans SQL, donc exportation dans Calc

Tableau de pilote, remplir champ de ligne et de données pour faire une double entrée et choisir dans champ de données l’identifiant

Statistiques descriptives sur une colonne permet d’en apprendre plus sur différents choses importantes sur la colonne choisie

Mode = chose la plus partagé, médiane = autant de chaque côté

Plage = étendue = maximum – minimum

Voir écart type et autre, peut être intéressant dans exploitation de base de données

Logiciel d’OCR, avec une photo va rechercher des taches de blancs et autres

20/10/16

Lexicométrie

Antoine Prost, aller voir « Les Mots » pour la lexicométrie, cet article définit ce qu’est un bon corpus.

Damon Mayaffre, pour le corpus et la linguistique, « les corpus politiques ».

Jacques Guilhaumou, un des pionniers de la textométrie et lexicométrie en étant historien, « l’historien du discours et la lexicométrie », intéressant pour la presse.

Alain Guerreau, « pourquoi et comment l’historien doit-il compter les mots ? », *Histoire et mesures*

Ensemble de méthodes, permet d’analyser du vocabulaire qui provient de plusieurs textes rassemblés ensemble qui provient d’un corpus, simplement. Ces méthodes s’appuient sur des statistiques. Apparut en 1960-1970, linguiste Z. Harrin, dépasser le contexte de la phrase pour analyser la phrase dans le texte, ces méthodes ne viennent pas de l’histoire, fruit de linguiste puis de littéraire.

Aujourd’hui, lexicométrie est plutôt de l’histoire culturelle, mais à la base plutôt de l’histoire politique. Réflexion sur l’emploi des mots dans l’histoire, R. KOSELLECK, pas le même sens à l’époque moderne qu’à l’époque contemporaine. On ne parle pas le même langage en fonction de son idéologie, essentiel de savoir dans quel système de communication on est. Savoir quel mot est sur-employé ou sous-employés.

Lexicométrie :

* Corpus homogène
* Contrastif, le but est d’analyser les mots entre eux
* Diachronique, donc être étalé dans le temps
* Texte inscrit dans des situations de communications déterminées

L’intérêt est de sortir d’un rapport de citations, s’intéresser à ce qui est la norme. Nettoyage de corpus à faire, par rapport à ponctuation, J’ et Je, verbe être sous différentes conjugaison et majuscules/minuscules. Lemmatisation à effectuer

André Salem, Lexico 3

TXM, permet de lemmatiser,

Note pad++

Texte Wrangler pour les macs

Délimiteurs permettent juste de dire que l’on veut récupérer les délimiteurs de la langue français

Le texte va alors avoir des variables, on compare ici les textes de Hebert sur l’année 1793-1794, contraste est le temps mais aussi les pages

Forme = suite de caractères identiques, la plus fréquente est « DE », présent 6130 fois dans ce corpus, 6130 occurrences

Hapax = mot qui n’apparait qu’une fois

On va donc s’intéresser à l’entre-deux, dépend de la taille du corpus, du nombre de mots, etc.

Partitions, permet de voir les occurrences du mot sur les mois, si on choisit « mois », « fréquence relative » = partie sur le tout, pondère par rapport à ce qui contient ; « fréquence absolue » = prend seulement en compte l’occurrence

Décrire le corpus pour le devoir, « principales caractéristiques lexicométriques du corpus » = PCLC, voir si corpus homogène

Division des occurrences par les formes, permet de savoir comment utilisation des mots

Si courbe plus faible, signifie que source très stéréotypé, si courbe monte tout le temps, pas d’homogénéité du corpus.

PARETO 🡪 échelle logarithmique, desecrasé des grandeurs énormes, peu de formes sont très utilisée, de plus en plus de formes sont de moins en moins utilisées, à la fin beaucoup de formes sont présentes moins de dix fois, distribution du vocabulaire mais sous forme de courbe.

Droite de Zipf = droite intéressante, constante qui fait que le rang est le mot le Xème fois le plus utilisé, f(x) = c x rang

Spécificités permet de savoir si un mot est spécifique ou pas selon le moment, calcul de probabilité, regarde le nombre de fois ou on dit rois (par exemple)

Tableau de contingence, effectifs, spécificités sur un mois après permet de voir forme sur employée

Permet à la fin de faire une analyse factorielle

Revenir au contexte d’utilisation des mots, possible plutôt alors de s’intéresser aux suites de mots, les segments répétés, ce sont des concordances radicales, sous-parties des concordances (plus facile sur TXM)

OCR = reconnaissance optique de caractères 🡪 cherche dans l’image les lignes, il observe les pixels, puis il recherche les mots, soit les espacements dans les lignes, puis enfin, les signes, les caractères. Il a en plus un dictionnaire de caractères, donc cela fonctionne toujours même si les papiers sont abimés.

Gallica a lui-même un logiciel d’OCR, donc les textes sont disponibles en mode texte.

Documents Léon Blum 🡪 Les balises ne sont pas faites, chaque info que l’on veut garder doit être mise dans la bonne syntaxe

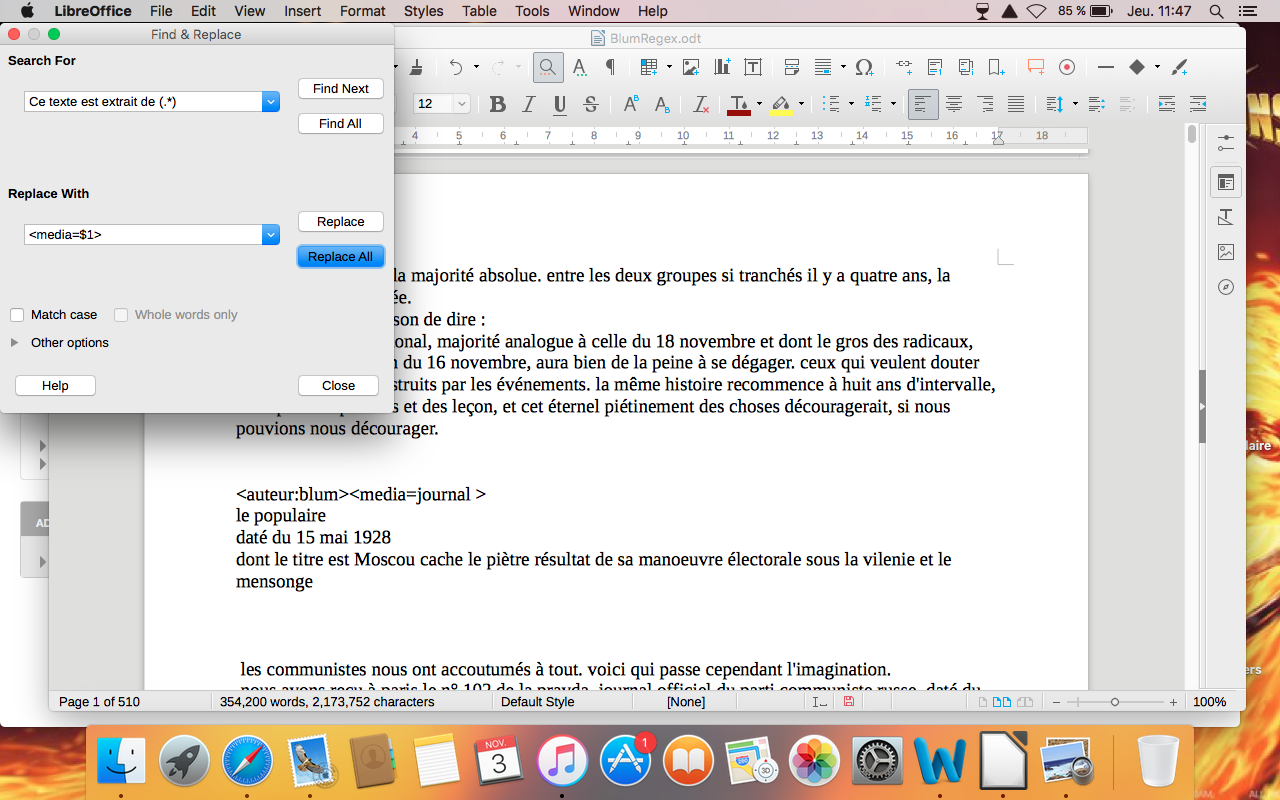
Lexico3 🡪 <auteur=blum> balise = 0 espace, 0 accent

Degré = référencement, bibliographique ou autre, n’a rien faire pour la textométrie ici

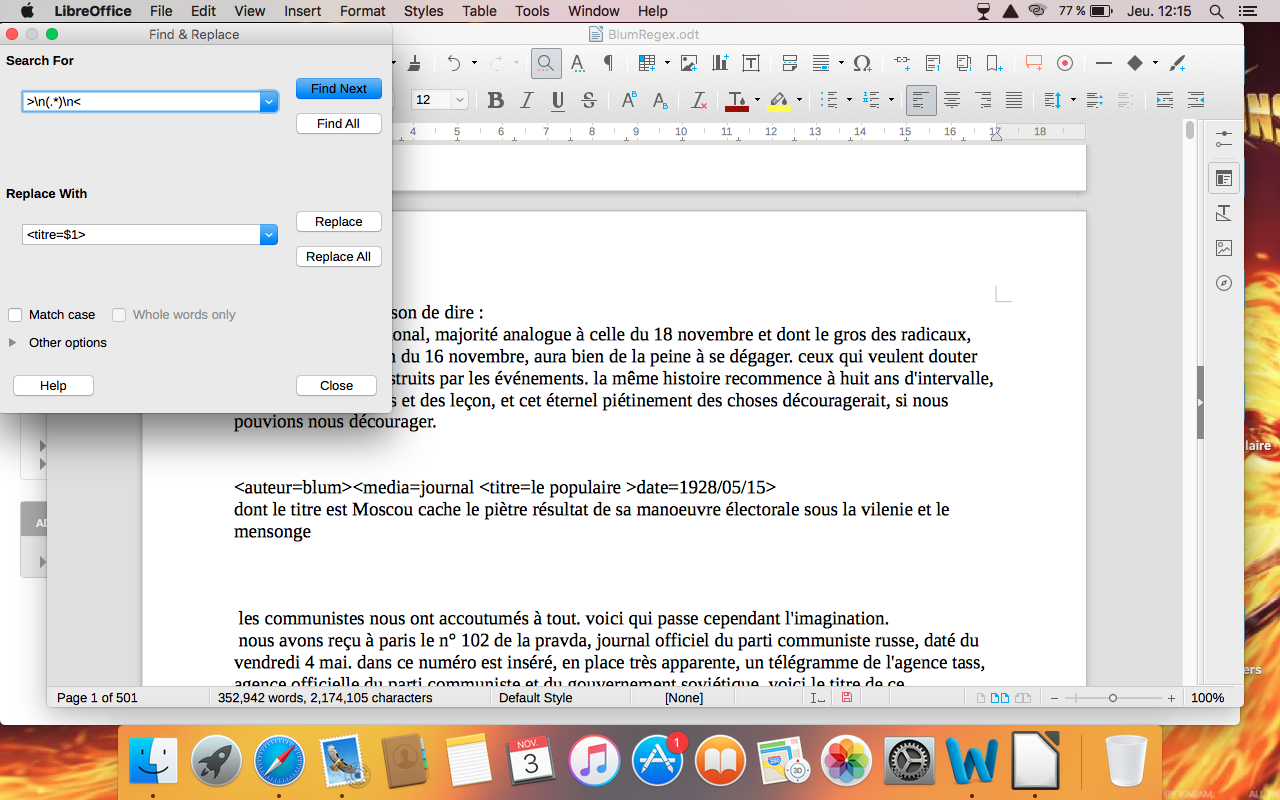
Nettoyage à faire du corpus, notes de bas de pages n’ont pas leur place pour la textométrie

Ch.ch./CH.ch./ch.ch.ch.ch. = les dates donc dans expressions régulières il faut demander \d\d/\d\d/\d\d\d\d mais plus simple de faire alors \d{2}/\d{2}/\d{4}

Savoir après ce qu’on garde ou pas, si les titres d’articles par exemple sont des balises ou déjà du vocabulaire



$1 permet de remplacer les différents signes pour garder ce qui est dans le point étoile entre parenthèses.



Signes les plus employés seront :

.

\*

\d

\r

\n

() 🡪 $

TXM = permet de lemmatiser

Lexico 3 🡪 corpus = 1 seul fichier alors qu’on a le choix dans TXM, 1 fichier = 1 texte

TXM 🡪 balisage = XML, nécessité de dire quand ce n’est plus un texte, avec <\text> à la fin

Toujours choisir dans fichier 🡪 importer 🡪 XML/w + CSV sélectionner le sous dossier et non le dossier qui contient le sous dossier

Ramener à leur racine tous les mots.

Langue principale annoter et français et après mettre lancer le corpus

Corpus, clique droit et après on a tous ce que l’on peut faire avec TXM

Lexico3 est mieux pour faire la description du corpus

.\*isme pour tout ce qui finit par isme

Lexicométrie – Discours Présidentielles 2007

TXM = plateforme de logiciel, utilise plusieurs logiciels, statistiques avec R et utilise Treetager, qui est un lemmatiseur 🡪 double traitement automatique de la langue, met chaque mot à sa racine (lire article de BRUNET, *Qui lemmatise, lime et* *attise ?*) mais cela simplifie énormément le discours politique, réduit les nuances. Associe à chaque mot une fonction morphosyntaxique, comme les verbes et leur conjugaison.

Word = occurrence

[Frpos= « ADJ »], beaucoup de taux d’erreurs pour frpos.

Partition, dans propriétés, changer par ce que l’on veut, ici candidat

Table lexicale 🡺 permet de faire une AFC, distinguer des profils

Sur Lexico 3, possible de voir qui a le langage le plus riche.

Concordances, permet de voir représentations des mots, recontextualiser

Cooccurrence = coprésence, savoir quels sont les mots qui tournent autour de la forme choisie

Aller dans Index pour savoir comment les formes sont utilisées

Lexico3 est bien surtout pour faire les PCLC

AFC

Comparer les lignes et les colonnes entre elles, méthode d’analyse multidimensionnelles, tableau de contingence, faire tout d’un coup avec la méthode

* ACM pour base de données = analyse en composantes multiples, tableau où chaque ligne correspond à une colonne et chaque colonne avec un individu et on va voir s’ils se ressemblent
* AFC = analyse factorielle des correspondances, ne part plus d’un tableau individu/caractère mais d’un tableau de contingence, on va ainsi voir avec les discours si des candidats utilisent les mêmes mots ou autres
* ACP = analyse en composante principale, peu utilisé par les historiens, toutes les variables doivent être quantitatives

ACM 🡪 faire tourner autour des personnages des mots, ne visent pas à l’exhaustivité, première des grande tendance sur axe des abscisses, Bourdieu = première opposition marquante, patrons et petits artisans avec professeurs, pas du tout les mêmes pratiques, axe des ordonnées, les vraies grosses oppositions sont dans les axes, permet de voir des principales tendances

AFC 🡪 Tableau lexical entier 🡺 énorme tableau de contingence, contient tous les mots et les parties choisies, on enlèvera tous les mots très peu utilisé, on ne s’intéressera qu’aux mots les plus employés en s’arrêtant à 20 ou 100

Pour la lexico, intéressant de faire plusieurs AFC sur plusieurs partitions

CSV = format de fichier tableau

Analyse SHS (Paris 1)

Encodage : UTF-8

Choisir Tableau (contingence ici) et type d’analyse (AFC ici)

Nécessité de sélectionner les mots qui apportent le plus de contrastes

10 facteurs 🡺 dix manières qu’il y a de se ressembler et de se différencier, première étape est de dire à cb de facteurs on va s’arrêter, valeur propre = longueur du vecteur, contribution à l’inertie avec les deux autre colonnes, 1er facteur = 1/3 le 2eme = 15% donc à eux seuls 50% de l’info, ici, arrêt aux deux premiers facteurs

Seconde étape est de trier, faire le seuillage, - de X% de l’inertie = virer de l’inertie, logiciel le fait pour nous. Ensuite analyse axe par axe

Modalité par colonne = coordonnées et contribution, dans contribution, voir qui contribue le plus à la fabrication de l’axe, les couleurs permettent de montrer que les données vont ensemble. Ici, plutôt idée d’un discours accusateurs et tourné idéologiquement en rouge mais en bleu, économie en terme de profit, discours du rassemblement

Inkscape, logiciel de dessin, utile pour les tableaux, s’assurer que les infos aient des liens

Probabilité d’indépendance (dans Test Chi 2) ou p valeur, moins de 5% signifie que chance que tableau pas lié au hasard. Au dessus de 10%, ce n’est plus possible d’analyser, faire un autre test, plus adapter comme le test de Fischer

Pour l’ACM, pareil mais juste, il est nécessaire de changer en tableau individu/caractère et ACM après étapes les mêmes.

Devoir 🡺 projet de modélisation ou de base de corpus, rendre entre le 15 janvier et le 15 avril