

PHILOLOGIE COMPUTATIONNELLE

Chap. 1. Naviguer dans un
corpus controversé ou ano-
nyme

Florian Cafiero & Jean-Baptiste Camps
3 novembre 2021

Master Humanités numériques | PSL – ENC, EPHE,
EHESS, ENS



PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

QUELLES DONNÉES POUR LA STYLOMÉTRIE?

PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

DIFFÉRENTES PERSPECTIVES, DIFFÉRENTS TRAITEMENTS

Parmi l'analyse de données textuelles, (au moins) deux approches principales peuvent être distinguées :

- Analyse du contenu / thématique / sémantique / littéraire (content, information retrieval, lexicométrie, ...)
- Identification des auteurs, datation, localisation (authorship attribution, stylométrie, ...).

LE FOND ('SÉMANTIQUE')

Faire ressortir les données relatives à la variété du lexique employé, à son sens, aux thèmes abordés, etc.. Approche lexicométrique, 'distant reading', information retrieval...

Dans cette perspective, les éléments extérieurs aux thèmes évoqués (données scribales, dialectales, morphologiques...) constituent **du bruit**

En pratique, pour ce type d'analyse, on tendra à :

- normaliser (graphies) / lemmatiser (supprimer l'information/le bruit scribal, dialectal, ...);
- privilégier les termes plus rares, porteurs de sens, par rapport aux « mots-vides » (et, ou, je, de, du, est, ...);
- pondérer pour faire ressortir les lemmes les plus caractéristiques des thèmes évoqués;
- travailler sur les cooccurrences et associations de lemmes.

LA FORME (STYLOMÉTRIE)

Dans cette perspective, on cherche à faire ressortir ce qui est **propre** à chaque individu, classe d'âge, genre ou bien à une zone géographique, une période, etc.

On tendra alors :

- essayer de minimiser le bruit relatif aux thèmes évoqués et contenu;
- conserver l'information grammaticale, graphique, ...;
- chercher les éléments les plus stables d'un texte à un autre du même auteur (quand cela est possible);
- éviter les biais dûs à la longueur des textes, etc.

- Solution simple : travailler sur le lexique global.
- Très sensible aux variations de thème, mais efficace si l'on contrôle très sévèrement ce point.
- On a aussi pu préférer utiliser des lemmes, parfois sans très bonne raison.
- Mais peut-être très utile : pour des langues fortement soumises à des variations graphiques, comme l'ancien Français, travailler sur le lemme peut être très utile.
- Exemple : Camps, Clérice, Pinche (2019) Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis, DH 2019. (forthc. DSH)

CAMPS, Jean-Baptiste, CLERICE Thibault, PINCHE, Ariane, Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis, DH 2019. (forthc. DSH ?)

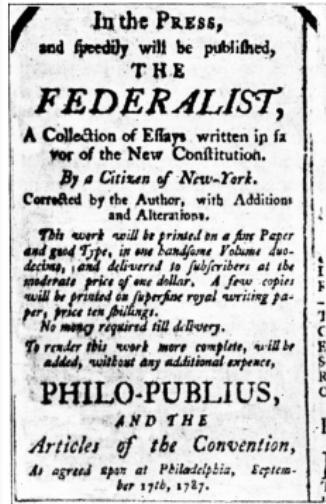
LES MOTS OUTILS - MOSTELLER & WALLACE, 1963

- Idée née dans un article puis une monographie de Frederick Mosteller et David Wallace (1963, 1964) sur les **Federalist Papers**.
- Recueil de 85 articles publiés en 1787 et 1788, dont le but était de promouvoir la nouvelle constitution américaine - et notamment de convaincre les délégués des 13 colonies de la ratifier.
- Tous ne sont pas considérés comme de la même importance : le Federalist n°10 (The Utility of the Union as a Safeguard Against Domestic Faction and Insurrection) ou le n°84 (Certain General and Miscellaneous Objections to the Constitution Considered and Answered), s'opposant à l'apposition d'un Bill of Rights à la constitution, sont par exemple considérés comme des textes fondamentaux de l'histoire des Etats-Unis.
- D'où la question : qui a écrit quoi ?

MOSTELLER, Frederick et WALLACE, David L. Inference in an authorship problem : A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. Journal of the American Statistical Association, 1963, vol. 58, no 302.

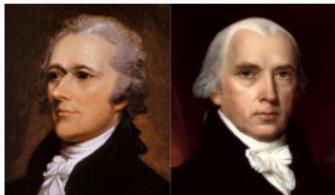
MOSTELLER, Frederick et WALLACE, David L. Inference and disputed authorship : The Federalist. 1964

LES MOTS-OUTILS - THE FEDERALIST



- Tous les articles sont signés d'un même pseudonyme : Publius, en référence au légendaire Publius Valerius Publicola de l'histoire romaine.
- Qui a écrit ces articles ? Trois auteurs potentiels : Alexander Hamilton, James Madison, John Jay.
- Les historiens sont d'accord pour dire que John Jay a en tout et pour tout écrit les FP n°2,3,4,5 et 64 - il ne pose donc pas problème.

THE FEDERALIST - HAMILTON ET MADISON



- Problème : les autres auteurs n'ont pas postérieurement revendiqués tous les textes du Federalist. Notamment parce qu'ils avaient changé d'avis sur certains sujets...
- Hamilton considéré comme l'auteur de 43 essais, Madison comme l'auteur de 14.
- Les articles allant de 49 à 58, ainsi que le 62 et le 63 ne sont attribués à personne.
- Les articles 18, 19, 20 ont été conjointement écrits par Hamilton et Madison – sans que l'on sache la part à attribuer à chacun d'eux.

- A la lecture seule, dur de distinguer les deux auteurs, au point que personne ne s'y était vraiment essayé.
- Style à la fois oratoire et touffu, dans la veine du journal the Spectator.
- Propriétés simples de leurs phrases assez semblables. Exemple de la longueur des phrases : 34,55 mots pour Hamilton contre 34,59 pour Madison , (σ resp. 19,2 et 20,3)
- Idée de l'historien Adair, spécialiste des Federalist Papers : l'usage de while (Hamilton) et de whilst (Madison) semble bien discriminer entre les textes écrits par les deux auteurs.
- Généralisent l'idée : étudier le lexique pour voir si certains mots ou types de mots permettent de distinguer un auteur d'un autre.

THE FEDERALIST - MOTS DE CONTENU

- Quels mots choisir d'étudier?
- Certains mots de contenu semblent permettre de différencier les deux auteurs - comme le mot "war", employé différemment par Hamilton et Madison.
- Mais évidemment dépendant du contexte d'énonciation, de la thématique abordée etc.

TABLE 2.2. FREQUENCY DISTRIBUTION FOR *war*

Rate/1000	H	M
0 (exactly)	23	15
0+2	16	13
2- 4	4	5
4- 6	2	4
6- 8	1	3
8-10	1	3
10-12	—	3
12-14	—	2
14-16	1	2
Totals	48	50

THE FEDERALIST - UN UNIQUE DISCRIMINANT?

- Mot le plus discriminant : upon.
- Se reposer entièrement sur lui malgré tout insuffisant.

TABLE 2.3. FREQUENCY DISTRIBUTION FOR *upon*

Rate/1000	H	M
0 (exactly)	—	41
0 +—1	1	7
1 —2	10	2
2 —3	11	
3 —4	11	
4 —5	10	
5 —6	3	
6 —7	1	
7 —8	1	
Totals	48	50

THE FEDERALIST - MOTS OUTILS

TABLE 2.1. FREQUENCY DISTRIBUTION OF RATE PER THOUSAND WORDS
 FOR THE 48 HAMILTON AND 50 MADISON PAPERS FOR *by*, *from*, AND *to*.
 THE UPPER LIMIT OF A CLASS INTERVAL IS NOT INCLUDED IN THE CLASS

Rate	<i>by</i>		Rate	<i>from</i>		Rate	<i>to</i>	
	H	M		H	M		H	M
1- 3	2		1- 3	3	3	20-25		3
3- 5	7		3- 5	15	19	25-30	2	5
5- 7	12	5	5- 7	21	17	30-35	6	19
7- 9	18	7	7- 9	9	6	35-40	14	12
9-11	4	8	9-11	1		40-45	15	9
11-13	5	16	11-13	3		45-50	8	2
13-15		6	13-15		1	50-55		
15-17		5		—	—	55-60	1	
17-19		3	Totals	48	50	Totals	48	50
Totals	48	50						

LEURS CONCLUSIONS

- Méthodologique :
 - Mots outils discriminent bien entre les textes (tentent trois variations qui marchent toutes bien)
 - Parmi eux, les mots les plus fréquents sont les plus utiles.
- Sur la problématique étudiée :
 - Madison serait l'auteur de la plupart des textes disputés
 - Exception : Hamilton aurait écrit le texte 65
 - Pour les textes écrits à plusieurs mains : Madison aurait écrit la majorité des textes 18 et 19, le texte 20 leur paraît trop compliqué.

D'une efficacité toujours reconnue plusieurs décennies après (Argamon & Levitan, 2005). Les raisons avancées pour expliquer ce phénomène sont diverses (Kestemont, 2014) :

Raisons statistiques

- Plus d'occurrences = plus de fiabilité;
- évite la survalorisation des hapax / contourne la distribution parétienne;

Raisons philologiques et cognitives

- moins soumis aux variations de contenu, thèmes, niveau de langue, genre, versification,...;
- usage inconscient des scripteurs (moins falsifiable, plus caractéristique d'un individu).

ARGAMON, Shlomo et LEVITAN, Shlomo. Measuring the usefulness of function words for authorship attribution. In : Proceedings of the 2005 ACH/ALLC Conference. 2005. p. 4-7.

KESTEMONT, Mike. Function words in authorship attribution. From black magic to theory?. In : Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL). 2014. p. 59-66.

LES SÉQUENCES MORPHOSYNTAXIQUES

- Etudier les différences dans l'usage des natures grammaticales ("Part-of-speech", abrégé en POS) donne d'excellents résultats pour attribuer un texte à un auteur.
- On trouve également les termes "catégorie grammaticale", "classe grammaticale", "espèce grammaticale", ou "partie du discours" pour désigner la même chose.
- Zhao & Zobel (2007) montrent que plus encore que d'étudier la liste des POS, étudier les séquences de deux POS entraîne de bonnes performances.
- D'autres benchmarks ont mis en avant les performances des 3-grammes de POS (Gamon, 2004; Argamon, 1998).

ZHAO, Ying and ZOBEL, Justin. Searching with style : Authorship attribution in classic literature. In : Proceedings of the thirtieth Australasian conference on Computer science-Volume 62. Australian Computer Society, Inc., 2007. p. 59-68.

GAMON, Michael. Linguistic correlates of style : authorship classification with deep linguistic analysis features. In : Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

LES SÉQUENCES DE CARACTÈRES

- Un classique de l'attribution d'autorité est d'utiliser des n-grammes de caractères.
- Permet de conserver de l'information sur l'enchaînement des mots, tout en préservant la simplicité de l'approche "sac de mots".
- La longueur des n-grammes de caractères donnant les meilleures performances pour l'attribution d'autorité varie selon les langues employées.
- "Not All Character N-grams Are Created Equal"(Sapkota et al., 2015) : les n-grammes de caractères ayant le plus de poids dans la différentiation entre textes sont les préfixes / suffixes etc.

SAPKOTA, Upendra, BETHARD, Steven, MONTES, Manuel, et al. Not all character n-grams are created equal : A study in authorship attribution. In : Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics : Human language technologies. 2015. p. 93-102.

Affixes

- Préfixe : un n-gramme de caractères couvrant les n premiers caractères d'un mots de longueur au moins n+1
- Suffixe : un n-gramme de caractères couvrant les n derniers caractères d'un mots de longueur au moins n+1
- Espace-préfixe : un préfixe commençant par un espace
- Espace-suffixe : un suffixe finissant par un espace

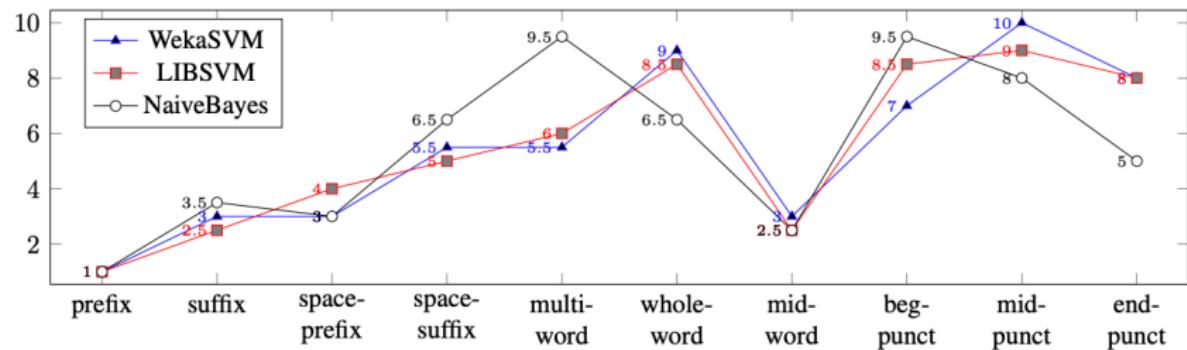
Mots

- Mot entier (whole word)
- Milieu de mot (midword)
- Multi-mots (multi-word) : un n-gramme couvrant plus d'un mot, incluant donc un espace

Ponctuation

- Ponctuation au début (begpunct) : un n-gramme commençant par une ponctuation
- Ponctuation au milieu (midpunct) : un préfixe incluant une ponctuation en dehors de ses extrémités
- Ponctuation à la fin (end-punct) : un n-gramme finissant par une ponctuation.

SC	Category	Character <i>n</i> -grams
affix	<i>prefix</i>	act wan pac see lik fas
	<i>suffix</i>	ors ted act med ike ned
	<i>space-prefix</i>	.ac .wa .to .se .if .th .pa .li .an .ol .on
	<i>space-suffix</i>	he_ rs_ ed_ to_ ee_ if_ ct_ ke_ an_
word	<i>whole-word</i>	The see the old one
	<i>mid-word</i>	cto tor ant nte eem eme ash shi hio ion one
	<i>multi-word</i>	e_a s_w d_t o_s e_i f_t e_p t_s d_l n_o d_o
punct	<i>beg-punct</i>	-fa
	<i>mid-punct</i>	d-f
	<i>end-punct</i>	ld- ne.



LES MOTS CHOISIS

- A l'inverse, on peut considérer que certains, mots certes rares, mais choisis avec soin, sont particulièrement révélateurs de l'auteur d'un texte.
- Dans un texte en vers, c'est le cas des mots à la rime.
- Moins souvent utilisé, car moins fiable statistiquement dans la plupart des cas : pas toujours assez de mots à la rime pour obtenir des résultats significatifs.
- Mais peut être intéressant dans certains cas

Mike Kestemont, Walter Daelemans, Dominiek Sandra, "Robust Rhymes? The Stability of Authorial Style in Medieval Narratives", Journal of Quantitative Linguistics, 2012

- Etude sur **Jacob van Maerlant** (ca. 1230 - ca. 1288), prolifique auteur néerlandais.
- Laisse derrière lui un vaste corpus versifié (plus de 200.000 vers)
- Mais certains textes qui lui ont été attribués pourraient ne pas être de sa main.
- Son contemporain **Lodewijk van Velthem** (ca. 1260/1275) - ca 1317) pourrait être l'auteur de certains textes qu'on lui prête.

Table 1. An example of the variation in medieval text transmission: one line from the *Rijmbijbel* shows variant readings in a series of parallel manuscripts.

Manuscript	Variant reading for 1 line from <i>Rijmbijbel</i> (‘On that moment and the same time’)
D	Ter stont ende ter seluer vren
E	Tier stont ende ter seluer vren
F	TIere stont enter seluer vren
G	Tottien standen en ter uren
H	TEn standen ende ter seluer vren
I	Tjerst stont ende tier veren
J	Tyer stont ende tier seluer vren
N	TJer stont tier seluer vre

- Idée : la transmission de textes médiévaux peut être chaotique, et certains mots du texte ont pu être affectés au cours des copies successives, notamment les mots outils.
- Mais les mots à la rime, porteurs de sens, et devant par définition rimer avec un mot suivant, sont moins susceptibles d'avoir subi de graves variations.
- Travaillent par conséquent sur les mots à la rime les plus fréquemment utilisés par l'auteur.
- Cette méthode laisse transparaître les thèmes évoqués et l'évolution des préoccupations de van Maerlant, mais permet toutefois d'identifier assez bien l'auteur d'un texte.

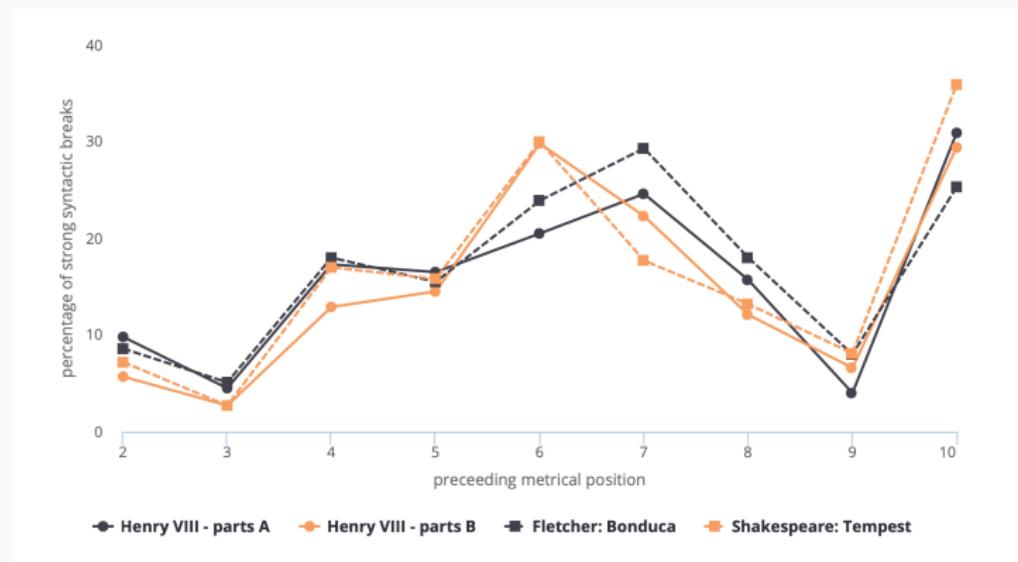


POÉSIE ET MÉTRIQUE : UN STYLE DENSE

- La plupart des événements observés en stylométrie sont des "événements rares" : employer tel mot, ou tel n-gramme de caractère etc., est événement qui a une probabilité relativement faible d'arriver.
- Il faut donc de grandes quantités de texte pour arriver à des calculs fiables.
- Problème : comment faire par exemple, quand on étudie de la poésie, où les textes peuvent être relativement courts ?
- Avantage de la métrique : booléen (accentué/pas accentué) ou nombre limité de valeurs (type de rythme).
- La métrique est par ailleurs suffisamment indépendante du contenu, dans la plupart des cas (même si variation selon genres etc.)
- Permet donc des analyses sur des textes plus courts.

MÉTRIQUE : EXEMPLE D'HENRY VIII

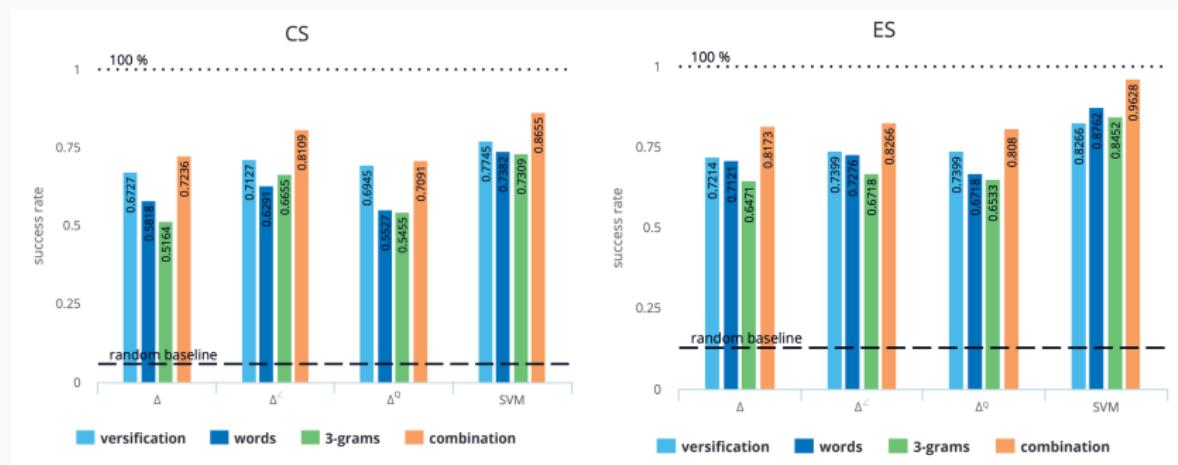
Exemple d'intuition : Shakespeare semble placer plus souvent ses pauses syntaxiques après l'hémistiche, quand Fletcher les placerait plus souvent après le septième pied.



TARLINSKAJA, Marina. *Shakespeare and the Versification of English Drama, 1561-1642*. Routledge, 2016.

MÉTRIQUE : BANC D'ESSAI

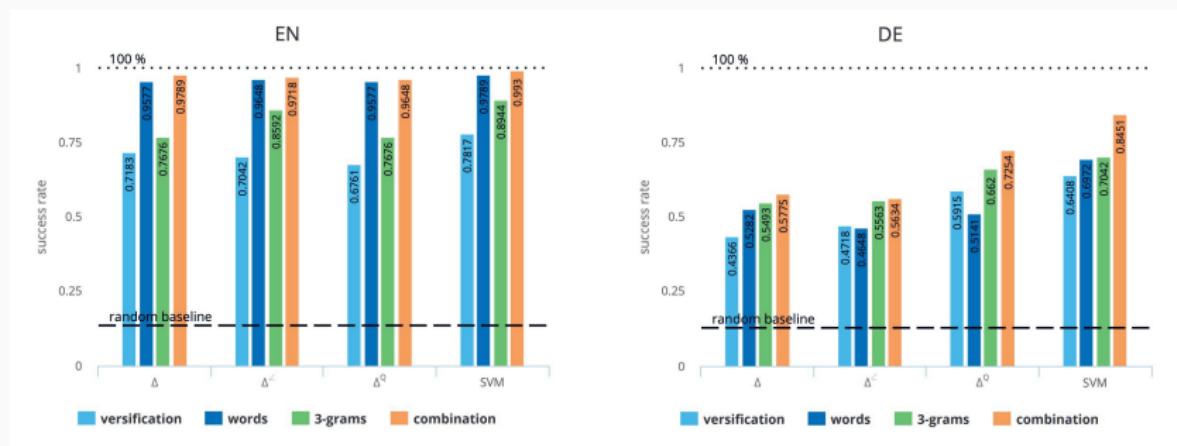
Très bonne performance en tchèque ou en espagnol par rapport à d'autres méthodes.



PLECHÁČ, Petr, BOBENHAUSEN, Klemens, et HAMMERICH, Benjamin. Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry. *Studia Metrica et Poetica*, 2018, vol. 5, no 2, p. 29-54.

MÉTRIQUE : BANC D'ESSAI 2

Moins bonne performance en anglais ou en allemand.



- Étude prometteuse sur le latin - appliquée à la Guerre punique de Silius Italicus

NAGY, Benjamin. Metre as a stylometric feature in Latin hexameter poetry. arXiv preprint arXiv:1911.12478, 2019.

COMBINER LES APPROCHES

- En pratique, on ne choisit pas nécessairement un angle plutôt qu'un autre.
- Juola : multiplier les méthodes, les mettre en parallèles. Si toutes vont dans le même sens, la probabilité de se tromper devient infinitésimale.
- Combiner les caractéristiques dans le même calcul, pour maximiser la précision. Pratique extrêmement courante.

PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

ÉCHANTILLONNER ?

- Bien d'autres caractéristiques du texte peuvent être étudiées selon le contexte de l'étude.
- Nombre de mots dans une phrase, nombre de proposition par phrase, apparition d'un certain type de vocabulaire après un type de mots etc.
- Rodionova et Marusenko (2010) : annotent environ 50 caractéristiques des textes étudiés pour leur calcul d'attribution d'autorité.
- Mais dans ce cas, pas toujours aussi simple d'obtenir automatiquement l'information.
- Besoin de faire beaucoup de travail d'annotation soi-même.
- Dans ce cas, possibilité d'échantillonner

MARUSENKO, Mikhail et RODIONOVA, Elena. Mathematical methods for attributing literary works when solving the “Corneille–Molière” problem. Journal of Quantitative Linguistics, 2010, vol. 17, no 1, p. 30-54.

LOI DE HEAPS (OU LOI DE HERDAN)

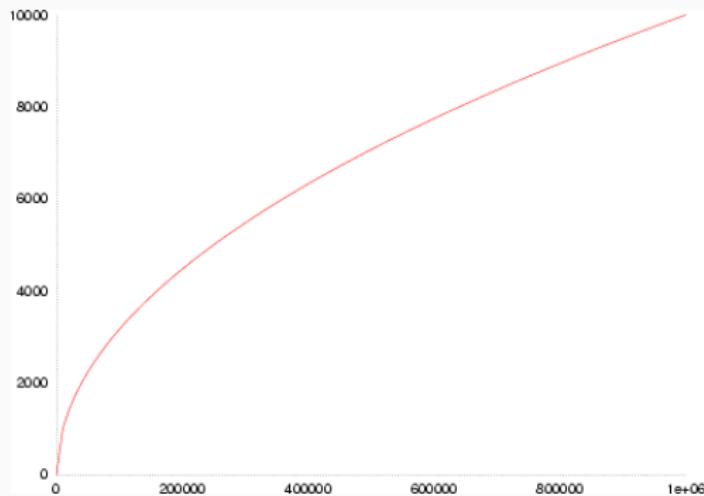
Heaps, Harold Stanley (1978), Information Retrieval : Computational and Theoretical Aspects, Academic Press.

Herdan, Gustav (1960), Type-token mathematics, The Hague : Mouton.

- Observation relativement intuitive : la probabilité de trouver un nouveau mot est plus faible en fin de texte qu'en début de texte.
- Mais les nouveaux mots apparaissent tout de même sans cesse!

LOI DE HEAPS (OU LOI DE HERDAN)

- Mathématiquement : $V_R(n) = Kn^\beta$
- $V_R(n)$ nombre de mots dans les n premiers mots d'un texte R
- K et β : paramètres. En pratique, K entre 10 et 100, β entre 0.4 et 0.6



CONSÉQUENCE : ÉCHANTILLONNAGE

- Quand et comment peut-on légitimement échantillonner?
- Grâce à la loi de Heaps-Herdan, on sait qu'avec un échantillon (relativement court) d'un texte, on aura rapidement l'essentiel du vocabulaire d'un auteur.
- On sait par contre que l'on n'a aucune chance d'avoir l'intégralité du vocabulaire dans notre échantillon.
- Rassurant : on peut travailler de manière assez fiable avec des échantillons courts; inquiétant : il est impossible d'avoir un échantillon parfaitement représentatif du vocabulaire d'un auteur.

COMMENT ÉCHANTILLONNER ?

- Le plus simple : choisir un passage (on choisit les 10000 premiers mots etc.)
- Peut être tout à fait efficace dans une majorité de cas, et plus praticable pour certains types d'annotations.
- Mais ce n'est pas la meilleure méthode dans des cas simples comme l'étude de mots, POS, ou n-grammes de caractères (Eder, 2015, in *Literary and Linguistic Computing*).
- Le plus efficace est de tirer au sort des mots / items des textes que l'on étudie.
- Poserait problème dans certains cas d'annotations, demandant un contexte plus ample.
- Minimum d'items pour avoir des résultats fiables dépendant de la langue, mais indépendant de la méthode d'après Eder : 2500 mots en prose latine, 5000 mots en anglais, hongrois ou polonais.

SÉLECTION DES VARIABLES (*CULLING*)

Les n plus fréquents

Les 100, 500, 1000, 5000 plus fréquents (combien?). Cf. Eder et al.

Présence

Variables présentes dans n% des textes.

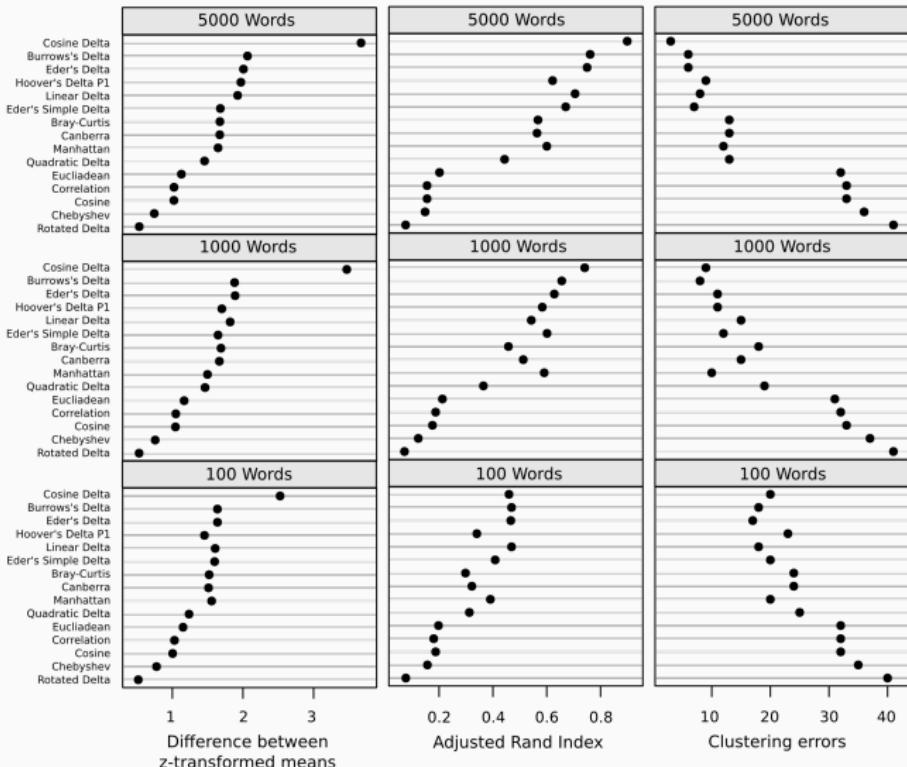
Fiabilité statistique (méthodes d'enquêtes)

Critère de Moisl (2011).

$$n = \bar{p}(1 - \bar{p}) \left(\frac{z}{e}\right)^2$$

Si distribution normale; sinon, il faut d'abord corriger, avec une variable miroir,

$$v_{\text{mirror}_{ji}} = (\max_v + \min_v) - v_{ji}$$



PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

FRÉQUENCES ET PONDÉRATIONS

FRÉQUENCES

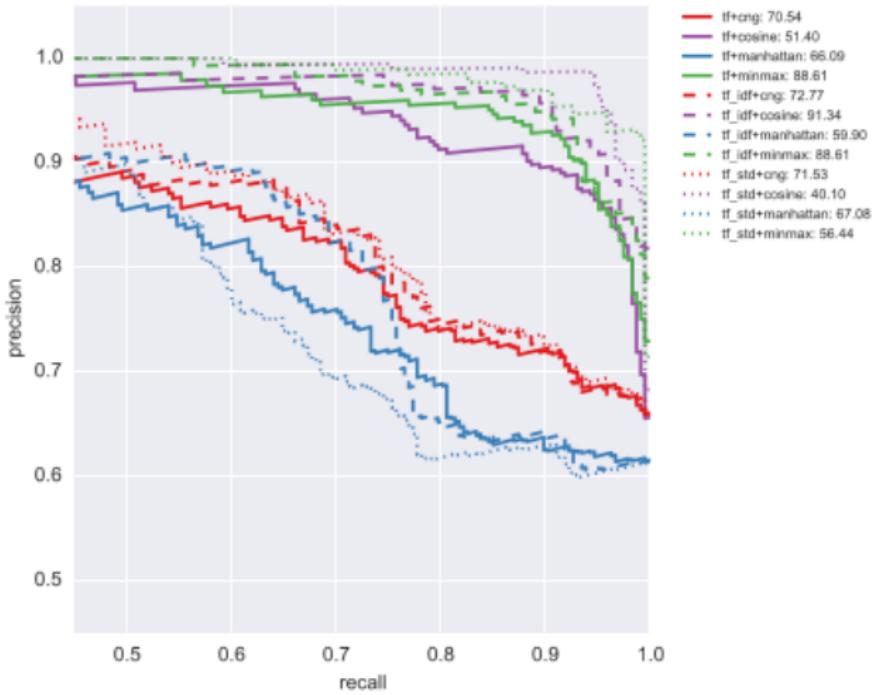
- fréquence absolue
(nombre d'occurrences);
- tf, fréquences relatives

$$\frac{tf_{ij}}{\sum_i^n tf_{ij}}$$

- booléen (1/0).

PONDÉRATIONS

- aucune;
- fréquence rapportée à l'écart-type (cf. delta de Burrows et d'Argamon);
- tf-idf (term frequency by inverse document frequency).



M. Kestemont, J. Stover, M. Koppelaar, F. Karsdorp, et W. Daelemans, « Authorship Verification with the Minmax Metric », Proceedings of the Digital Humanities 2016 conference, 2016, http://www.dhbenelux.org/wp-content/uploads/2016/05/108_KestemontEtAl_FinalAbstract_DHBenelux2016_long.pdf.

Ce qu'on veut :

pour chaque colonne (i.e. texte) du tableau, diviser chaque valeur par la somme des valeurs de la colonne (i.e., le nombre total de mots de la phrase).

Ce qu'on veut :

pour chaque colonne (i.e. texte) du tableau, diviser chaque valeur par la somme des valeurs de la colonne (i.e., le nombre total de mots de la phrase).

```
1 #Créons une copie de l'objet
2 CornMolRetraite = CornMol[1:100,]
3 #Et mettons en œuvre notre pondération
4 for(i in 1:ncol(CornMolRetraite)){
5     CornMolRetraite[,i] =
6     CornMolRetraite[,i]/sum(CornMolRetraite[,i])
7 }
```

CRÉER UNE FONCTION GÉNÉRALISTE DU MÊME OBJET

```
1 "ponderation-maison" = function(x) {  
2     X = x ;  
3     for(i in 1:ncol(x)){  
4         X[,i] = X[,i]/sum(X[,i])  
5     }  
6     return(X)  
7 }
```

Ensuite, je peux l'appeler via

`ponderation-maison(x)`

Si j'ai besoin de l'utiliser fréquemment, je peux aussi la stocker dans un fichier de script (.R) et le sourcer quand j'en ai besoin, via

`source(<monfichierdescript.R>)`

- Si on ne veut pas travailler sur les fréquences d'apparition d'un mot ou autre phénomène textuel, on peut choisir de binariser.
- Si un phénomène est présent, on le note 1, sinon, on note 0.
- Ceci permet de contourner les biais pouvant apparaître suite aux phénomènes de contagion ("clumping") : un mot a en soi très peu de chance d'apparaître dans un texte; mais une fois qu'il a été utilisé, la probabilité qu'il réapparaisse devient bien plus forte.

Avec une approche très particulière à R, en évitant d'utiliser une boucle :

```
1 CornMolRetraite[CornMolRetraite > 0] = 1
```

ANALYSE EXPLORATOIRE DE DONNÉES

PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

ANALYSE PAR RÉDUCTION DES DIMENSIONS

- Nous avons vu lors du cours précédent comment des représentations graphiques simples permettaient de résumer une grande partie de la relation entre deux variables.
- L'objectif des méthodes d'analyse de données que nous allons présenter est de remplir les mêmes offices dans le domaine plus délicat des statistiques multivariées.
- Ces analyses de données, qui visent à simplifier l'information pour la rendre lisible, sont appelées **analyse par réduction des dimensions**.
- Au-delà d'usages descriptifs que nous évoquerons ici, certaines de ces méthodes ont un usage prédictif. Elles feront l'objet de notre dernière séance.

ANALYSE PAR RÉDUCTION DES DIMENSIONS : OBJECTIFS

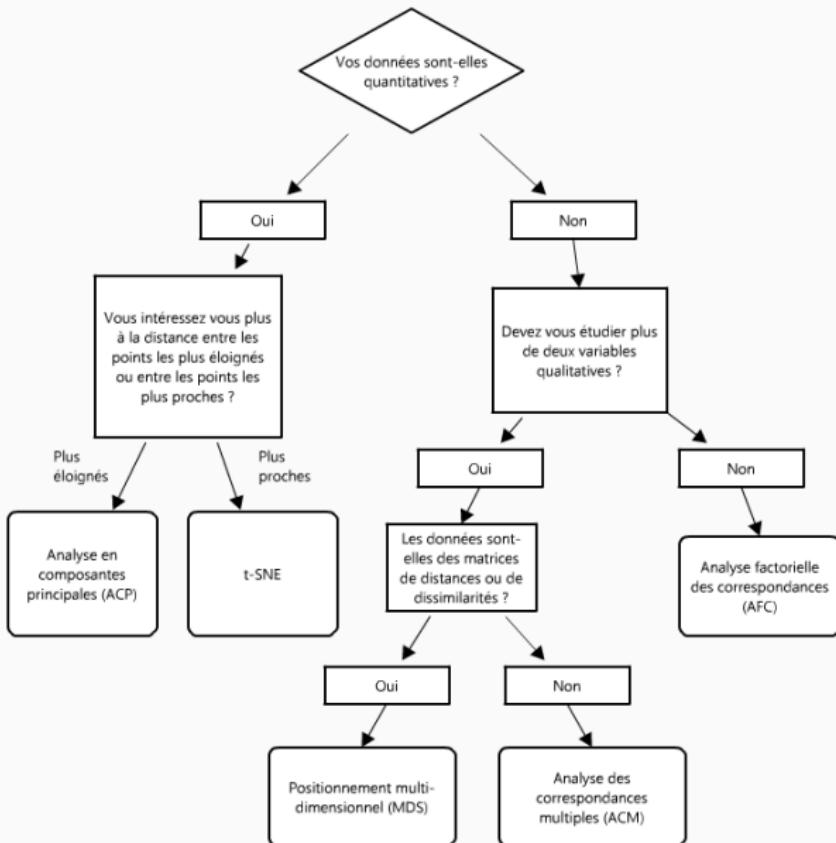
La valeur de ces méthodes est pour nous triple :

- **Visualiser** des données complexes pour y discerner des regroupements, des régularités, des typologies
- **Débruiter** les données, en supprimant de l'analyse les dimensions qui peuvent être considérées comme à négliger
- **Décorrélérer** : les axes créés au cours de ces analyses ne sont pas nécessairement des variables de la base, mais des construits n'ayant aucune corrélation entre eux.

ANALYSE PAR RÉDUCTION DES DIMENSIONS : MÉTHODE

- Dans une analyse factorielle, on cherche à déterminer les axes qui absorbent le plus d'inertie possible par rapport au centre de gravité du nuage de point.
- Concrètement, on va passer d'un nuage de points de grande dimension à un sous-espace de plus petite dimension, sur lequel on pourra réunir la plus grande quantité d'information possible.
- On choisit de déterminer des axes, définissant un plan sur lequel les points du nuage seront projetés.
- Ces axes sont choisis pour que les points projetés soient les plus dispersés possibles.

ANALYSE PAR RÉDUCTION DES DIMENSIONS : TYPOLOGIE



ANALYSE PAR RÉDUCTION DES DIMENSIONS : TYPOLOGIE

Trois de ces méthodes sont particulièrement usitées. Elles sont utiles dans des circonstances différentes :

1. Quand les variables sont quantitatives, on peut réaliser une **Analyse en Composantes Principales (ACP)** ou utiliser un algorithme t-SNE.
2. Quand les variables sont qualitatives, on utilise des **Analyses des Correspondances** - la correspondance étant "l'équivalent" de la corrélation pour des variables qualitatives :
 - Quand les individus sont décrits par deux variables qualitatives, on peut construire un tableau de contingence et réaliser une **Analyse Factorielle des Correspondances (AFC)**.
 - Quand les individus sont décrits par un jeu plus de deux variables qualitatives, on peut réaliser une **Analyse des Correspondances Multiples (ACM)**.

ANALYSE PAR RÉDUCTION DES DIMENSIONS : VENTILATION

- Nécessité parfois de se débarrasser de modalités peu pertinentes, car elles brouillent le calcul, ou la visualisation
- Plutôt que la suppression, possibilité de ventiler : on remplace les modalités ne dépassant pas un effectif minimal par la valeur moyenne de l'échantillon sans ces modalités. On les neutralise ainsi.

PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

- Suite à des idées de Pearson, formalisée par Harold Hotelling en 1933¹
- Principe général, commun à ces trois analyses : trouver des variables combinaisons linéaires d'autres variables, qui formeront les "facteurs".
- Les points sont visualisés en fonction de ces facteurs, la variance du nuage de points étant maximisée - intuitivement : on veut que les points soient le plus séparés possibles, pour ne pas obtenir un amas de points sans séparation visible.

1. Hotelling H., « Analysis of a Complex of Statistical Variables with Principal Components », 1933, Journal of Educational Psychology

ALTERNATIVE AUX ACP : T-SNE

- t-SNE : t-Distributed Stochastic Neighbor Embedding
- Développée par Laurens van der Maaten et Geoffrey Hinton, en 2008 puis perfectionnée jusqu'en 2014²
- Implémentations disponibles sous R (mais aussi Python, Matlab etc.)³

2. L.J.P. van der Maaten and G.E. Hinton. "Visualizing High-Dimensional Data Using t-SNE", Journal of Machine Learning Research 9 (Nov) : 2579-2605, 2008; L.J.P. van der Maaten, "Accelerating t-SNE using Tree-Based Algorithms", Journal of Machine Learning Research, 3221-3245, 2014.

3. Voir : <https://lvdmaaten.github.io/tsne/>

PRINCIPES DU T-SNE

- Principe : préserver les distances les plus courtes, plutôt que les distances les plus longues comme dans les ACP.
- On veut pour chaque point que la distance avec les points les plus proches soit bien préservée.
- Perplexité : indique le nombre de voisins du point avec lesquels on va vouloir préserver la distance. Les valeurs standard s'étalent entre 5 et 50.

LIMITES DU T-SNE

- Attention à la complexité de l'algorithme. En cas de grande base de données (>10.000 individus), il est préférable d'utiliser la méthode du Barnes-Hut t-SNE, reposant sur l'approximation éponyme, et qui permet de travailler avec de très grandes quantités de données.
- Il est possible d'éviter un temps d'exécution trop long en définissant un nombre maximal d'itérations (max_iter dans le package tsne de R par exemple).
- Les résultats peuvent être différents à chaque fois que l'on calcule un t-SNE. Il peut être utile de relancer plusieurs fois la procédure, et de sélectionner celle qui donne le meilleur résultat⁴.

4. Ce qui veut dire que l'on choisit la solution qui minimise la fonction objectif.

ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

- L'AFC s'applique à des **tableaux de contingence**⁵ c'est-à-dire des tableaux croisant deux variables qualitatives.
- L'AFC est en cela très distincte de l'ACP : les lignes et les colonnes jouent ici des rôles symétriques alors que la distinction entre lignes et colonnes, i.e. entre individus et variables, est majeure en ACP.
- Mais elle découle en fait de l'ACP : une AFC réalise une ACP sur le profil "ligne", une autre ACP sur le profil "colonne", et superpose les deux graphiques.

5. Voir Pearson K., "On the Theory of Contingency and Its Relation to Association and Normal Correlation", Mathematical Contributions to the Theory of Evolution, London : Dulau & Co, 1904

ANALYSE DES CORRESPONDANCES MULTIPLES (ACM)

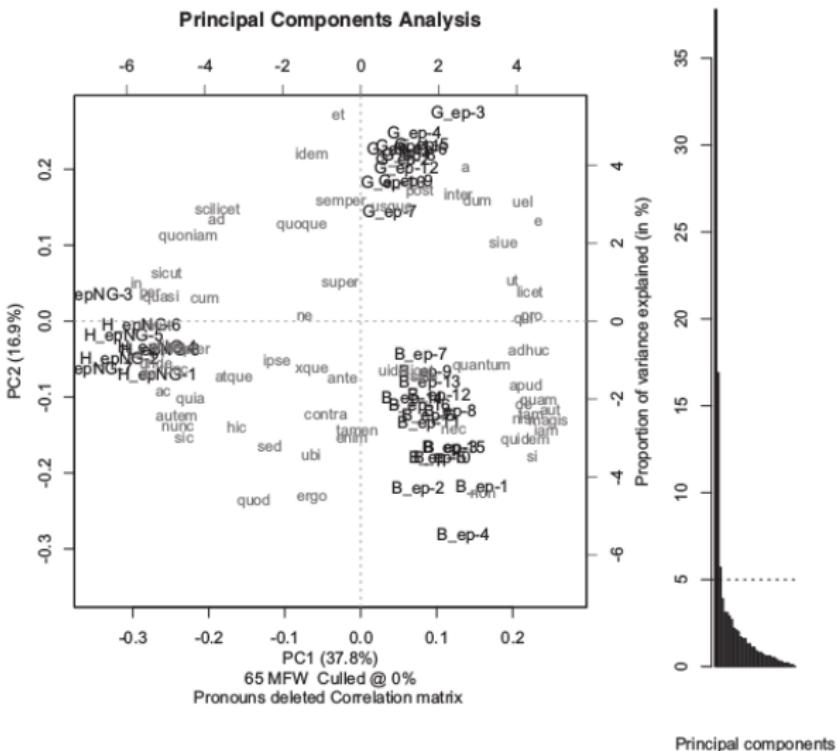
- Elle s'applique à des tableaux dans lesquels un ensemble d'individus est décrit par un ensemble de variables qualitatives.
- Le type de tableau utilisé ici est voisin de celui analysé en ACP, les variables quantitatives étant remplacées par des variables qualitatives.
- L'ACM est sensible aux effectifs faibles. Avant toute ACM, il est indispensable de réaliser une analyse préliminaire de chaque variable, afin de voir si toutes les classes sont aussi bien représentées ou s'il existe un déséquilibre.

Il existe différents packages pour réaliser des Analyses factorielles sous R. Nous choisissons ici d'utiliser **FactoMineR()**, un des packages de référence, français de surcroît, et régulièrement mis à jour.

- Pour les ACM, on utilise `mca()`, et `plot.mca()`
- Pour les ACP, on utilise `pca()` et `plot.pca()`
- Pour les AFC, on utilise `ca()` et `plot.ca()`

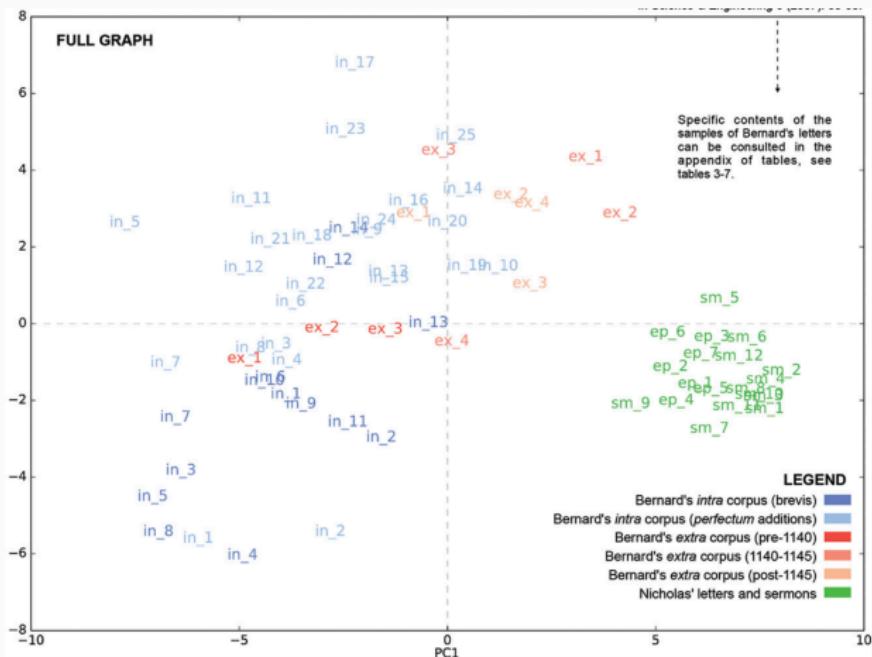
options invisible et axes

HILDEGARDE DE BINGEN ET SES SECRÉTAIRES



ACP des epistolaria de Hildegarde de Bingen, Guibert de Gembloux et Bernart de Clairvaux (Kestemont et al., 2015)

BERNARD DE CLAIRVAUX ET NICOLAS DE MONTIÉRAMEY



ACP du corpus épistolaire de Bernard de Clairvaux comparé aux lettres et sermons de Nicolas de Montiéramey. (Jeroen De Gussem, in Speculum, 2017)

POSITIONNEMENT MULTIDIMENSIONNEL

- Le positionnement multidimensionnel (ou Multidimensional scaling, MDS en anglais)
- Méthode issue de la biologie, créée pour évaluer les différences et les évolutions génétiques.
- Principe : on affecte à une position à chaque individu dans un espace à N dimensions, en fonction des similarités/dissimilarités entre individus.
- Options majeures : Classical MDS et Non metric MDS.
- Classical MDS : MDS classique, parfois appelé MDS de Torgerson. Transforme les distances en similarités et réalise une ACP.

POSITIONNEMENT MULTIDIMENSIONNEL NON MÉTRIQUE

- MDS non métrique : approche différente. On définit a priori un nombre d'axes. Les données sont compressées pour être affichées dans un référentiel composé de ces axes.
- Choix du nombre d'axes important :
 - trop peu d'axes forcera plusieurs axes de variation à être exprimés sur une seule dimension.
 - Mais trop de dimensions n'est pas mieux! Une seule source de variation pourrait alors être mécaniquement dispersée sur plusieurs dimensions.

POSITIONNEMENT MULTIDIMENSIONNEL NON MÉTRIQUE (2)

- Indice de déformation : le "stress". Il augmente mécaniquement avec le nombre d'individus et le nombre de variables (gare à ne pas comparer cette valeur entre différentes bases).⁶
- On peut faire un graphique montrant l'évolution du stress en fonction du nombre de dimensions choisies pour comprendre quel serait le nombre de dimensions nécessaire pour une lecture optimale ("scree plot")
- Relancer plusieurs fois les MDS (possibilités d'être "coincé" sur un optimum local)

6. Des débats nombreux existent quant aux manières de juger un niveau de stress acceptable. Moins de 0.1 est souvent vu comme excellent.

CAS PARTICULIER : POSITIONNEMENT UNIDIMENSIONNEL

- Exemple : si on choisit de projeter vers une dimension, on obtient un axe.

Patrick Mair, Jan de Leeuw, Patrick J. F. Groenen

Configurations Unidimensional Scaling

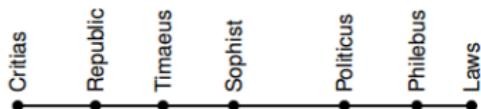


Figure 9: Unidimensional scaling on Plato's works.

CAS PARTICULIER : POSITIONNEMENT UNIDIMENSIONNEL

- Ce type de visualisation peut toutefois poser problème, la déformation (stress) induite par la projection vers une seule dimension pouvant être très importante.
- La lecture de ce type de graphique demande beaucoup de prudence (tentation de lire cela comme un axe chronologique par exemple, ce qui n'est pas toujours justifié).

QUESTIONS?

DISTANCIATION LITTÉRAIRE ET CLUSTERS

REGROUER DES TEXTES : SUPERVISÉ VS NON SUPERVISÉ

- Après avoir exploré nos corpus, on est tenté de vouloir regrouper nos textes entre eux.
- Lesquels sont de la même époque ? Ont le même style ? Ont le même auteur ?
- Deux options pour cela :
 - Je donne des données brutes à mon algorithme, sans aucune métadonnée, et je lui demande comment il regrouperait les différents textes : c'est une démarche dite **non supervisée**
 - J'étiquette mes données avec des métadonnées (auteur du texte, lieu d'écriture, époque de création, genre etc.) Je fais reconnaître à l'algorithme les caractéristiques liées à ces métadonnées. Puis je lui demande d'étiqueter des textes sans métadonnées. C'est une démarche **supervisée**.
- Nous nous intéressons aujourd'hui au premier cas.

PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

QUELLE DISTANCE ENTRE DEUX TEXTES ?

- Pour regrouper des textes qui nous semblent proches, il nous faut d'abord définir ce que nous appelons "proche".
- Nécessité de définir ce qu'est la distance entre deux textes.
- Idée la plus simple :
 - On choisit un critère d'intérêt (nombre d'occurrences de mots, de mots-outils, de 3-grammes de POS)
 - On calcule la différence, pour chaque mot/mot-outil etc., entre leur proportions dans le texte A et dans le texte B : le texte A utilise "de" dans 0.5% des cas, le texte B dans 0.3 % des cas -> différence de 0.2.
 - On somme ces différences, élevées au carré pour avoir toujours des valeurs positives, pour chaque mot. Le résultat obtenu est la distance entre les deux textes.

MÉTRIQUE (1) : DISTANCE EUCLIDIENNE

- Agir de la sorte revient à utiliser la notion de distance la plus courante dans les mathématiques standards : la **distance euclidienne**.

$$\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

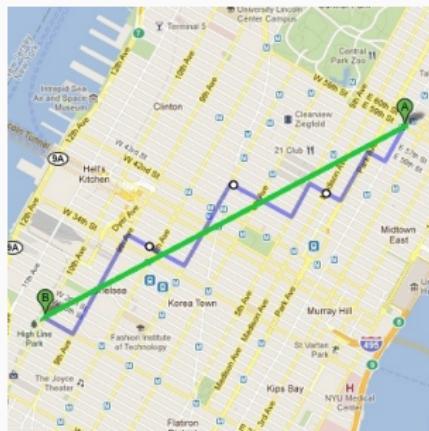
- C'est la mesure habituelle de la distance, celle du monde physique.
- Mais pourquoi la distance entre deux textes se mesurerait comme la distance entre une table et une chaise ?

MÉTRIQUES : UNE GRANDE VARIÉTÉ...

Distance/Similarity measure	Formula
Manhattan Distance	$\sum_{i=1}^n x_i - y_i $
Euclidean Distance	$\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
Canberra Distance	$\sum_{i=1}^n \frac{ x_i - y_i }{ x_i + y_i }$
Cosine Distance	$\frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$
Burrows' Delta	$\frac{1}{n} \sum_{i=1}^n \left \frac{x_i - \mu_x}{\sigma_x} - \frac{y_i - \mu_y}{\sigma_y} \right $
Argamon's Linear Delta	$\frac{1}{n} \sum_{i=1}^n \sqrt{\left \frac{(x_i - y_i)^2}{\sigma_i^2} \right }$
Eder's Delta	$\frac{1}{n} \sum_{i=1}^n \left(\left \frac{x_i - y_i}{\sigma_i} \right \cdot \frac{n-n_i+1}{n} \right)$
Eder's Simple Delta [9]	$\sum_{i=1}^n \sqrt{x_i} - \sqrt{y_i} $
Argamon's Quadratic Delta	$\frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_i^2} (x_i - y_i)^2$
Bray-Curtis Dissimilarity	$\frac{\sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n (x_i + y_i)}$
Kulczynski Distance	$\frac{\sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n \min(x_i, y_i)}$
Jaccard Index	$\frac{2 \sum_{i=1}^n x_i - y_i }{2 \sum_{i=1}^n (x_i + y_i)} \cdot \frac{1 + \sum_{i=1}^n x_i - y_i }{1 + \sum_{i=1}^n (x_i + y_i)}$
Gower Similarity	$\frac{1}{n} \sum_{i=1}^n \frac{ x_i - y_i }{\max_i - \min_i}$
Alternative Gower Similarity	$\frac{1}{n_0} \cdot \sum_{i=1}^n x_i - y_i $
Horn's modification of Morisita's Overlap Index	$\frac{2 \sum_{i=1}^n x_i y_i}{\left(\frac{\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i)^2} + \frac{\sum_{i=1}^n y_i^2}{(\sum_{i=1}^n y_i)^2} \right) \sum_{i=1}^n x_i \sum_{i=1}^n y_i}$
Mountford Index	$\frac{1}{\alpha}$, where α is the parameter of Fisher's log-series
Binomial Index [5]	$\sum_{i=1}^n \frac{x_i \cdot \ln \frac{x_i}{2n} + y_i \cdot \ln \frac{y_i}{2n} - 2n \cdot \ln \frac{1}{2}}{2n}$

MÉTRIQUE (2) : DISTANCE DE MANHATTAN

Krause, E. F., Taxicab Geometry : An Adventure in Non-Euclidean Geometry. New York : Dover, 1986.



Somme des valeurs absolues des différences entre coordonnées .

$$\sum_{i=1}^n |x_i - y_i|$$

MÉTRIQUE (3) : DISTANCE DE CANBERRA

Beaucoup de mesures de distance utilisées dérivent de la distance de Manhattan.

- Distance de Canberra⁷ :

Distance de Manhattan pondérée.

$$D_{\text{Canb}}(x, y) = \sum_{i=1}^n \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

- Utilité : permet d'atténuer l'importance de certaines différences majeures d'un point de vue numérique. On s'intéresse plus à l'existence de différences qu'à l'importance quantitative de chacune de ces différences.

7. Lance, G. N.; Williams, W. T. (1966). "Computer programs for hierarchical polythetic classification ("similarity analysis").", Computer Journal, 9 (1) : 60–64.

MÉTRIQUE(4) : DELTA DE BURROWS

Le delta de Burrows⁸ fait partie des distances les plus usitées en stylométrie. Considérés comme des outils aussi performants pour la prose que pour la poésie⁹, Leur efficacité a cependant été récemment remise en cause (Iannidis et al., 2015)

Delta de Burrows : combine la standardisation des comptes d'occurrences (z-transformation → on normalise les moyennes à 0 et l'écart type à 1) et d'un changement de métrique (distance de Manhattan) :

$$\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - B_i}{\sigma_i} \right|$$

8. John Burrows, ‘Delta’ : a Measure of Stylistic Difference and a Guide to Likely Authorship, *Lit Linguist Computing* (2002) 17 (3).

9. David L. Hoover, “Testing Burrows’s Delta”, *Literary & Linguistic Computing* (2004) 19 (4).

POURQUOI CETTE MÉTRIQUE ?

- En réalité, peu d'intuitions sur pourquoi une mesure de distance serait plus efficace qu'une autre.
- Argamon (2008) : delta de Burrows incohérent mathématiquement : la z-transformation a du sens dans le cas d'un monde euclidien. Mais elle est combinée à une distance de Manhattan...
- Propose alors deux alternatives pour rendre plus cohérent ce calcul.
- Delta linéaire d'Argamon : distance de Manhattan, mais normalisation non plus selon la moyenne arithmétique et l'écart-type, mais selon la médiane et la dispersion.
- Delta quadratique d'Argamon : distance Euclidienne et z-transformation.

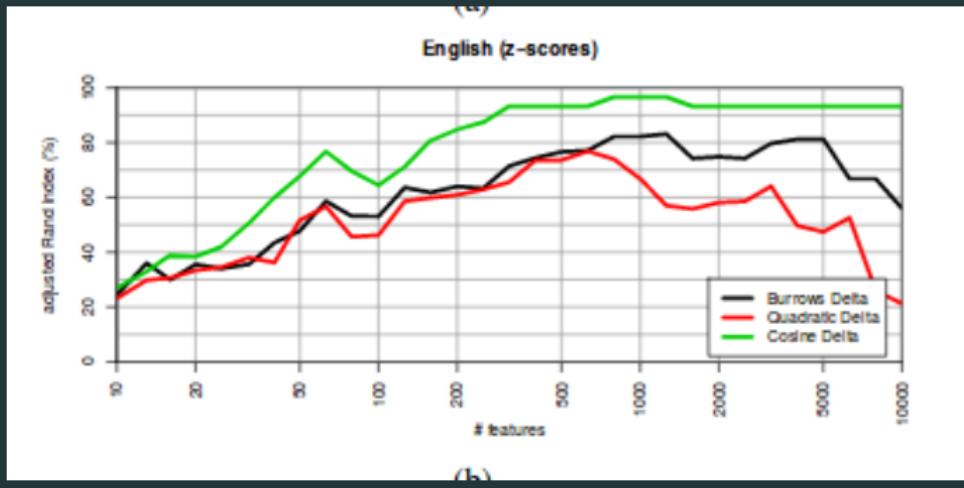
MÉTRIQUE (5) SIMILARITÉ COSINUS

Pour mesurer la similarité entre deux textes, on peut représenter chacun des textes comme un vecteur. On calcule alors le cosinus de l'angle formé par les deux vecteurs.

$$\cos(\theta) = \frac{\mathbf{A} \times \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|}$$

Cette valeur est théoriquement comprise dans $[-1; 1]$. Mais pour des valeurs toujours positives ou nulles, comme c'est le cas en textométrie, la valeur de cette mesure sera toujours comprise dans $[0; 1]$.

MÉTRIQUES : EFFICACITÉ (EVERT ET AL., 2017)



MÉTRIQUE (6) MINMAX

Nouvelle proposition : Koppel et Winter, 2014 : distance minmax.

$$\text{minmax}(\vec{A}, \vec{B}) = 1 - \left(\frac{\sum_{i=1}^n \min(\text{tf}(A_i), \text{tf}(B_i))}{\sum_{i=1}^n \max(\text{tf}(A_i), \text{tf}(B_i))} \right)$$

NORMALISATIONS DES FRÉQUENCES

Z-scores

Le delta de Burrows fait en réalité usage d'un type de normalisation qui peut s'employer plus généralement, les **z-scores**

Normalisation de longueur des vecteurs

L'efficacité de la distance cosinus peut être reproduite par la normalisation de longueur des vecteurs (Jannidis et al. 2017)

L1 Norm (Manhattan)

L2 Norm (Euclidienne, l plus couramment utilisée)

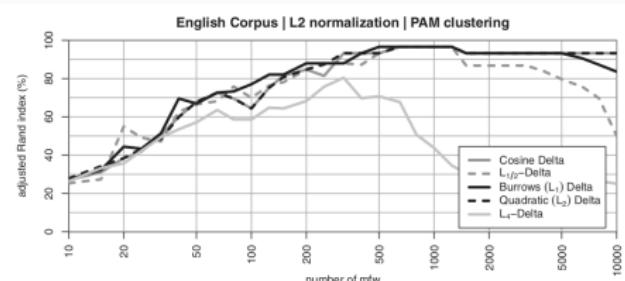


Fig. 9 Clustering quality of different Delta measures with length-normalized vectors (according to the Euclidean norm) in the English Corpus

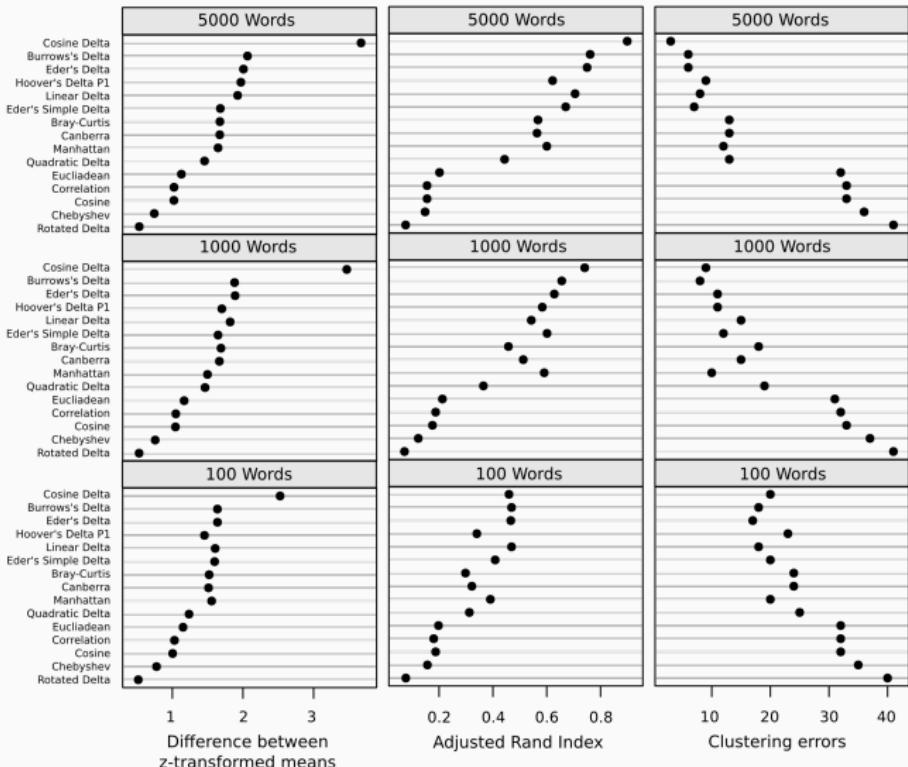
EFFICACITÉ COMPARÉE DES DIFFÉRENTES MESURES DE DISTANCE

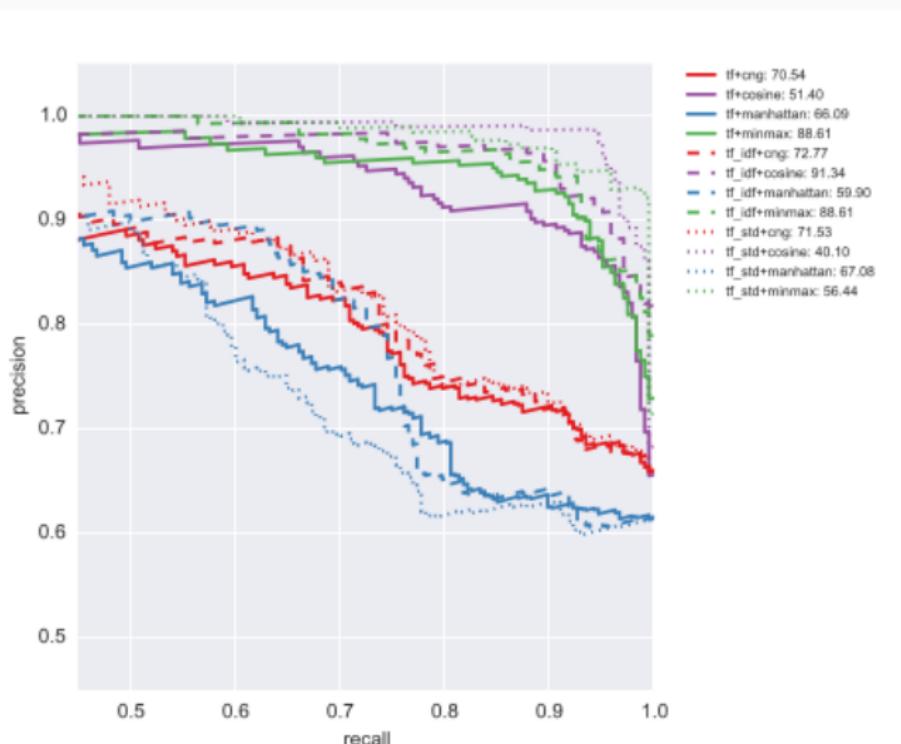
Deux éléments à prendre en compte :

- absence d'un cadre théorique expliquant la pertinence relative des différentes métriques;
- nécessité d'avoir recours à l'expérience, sur des corpus donnés, donnant parfois (souvent?) des résultats contradictoires.

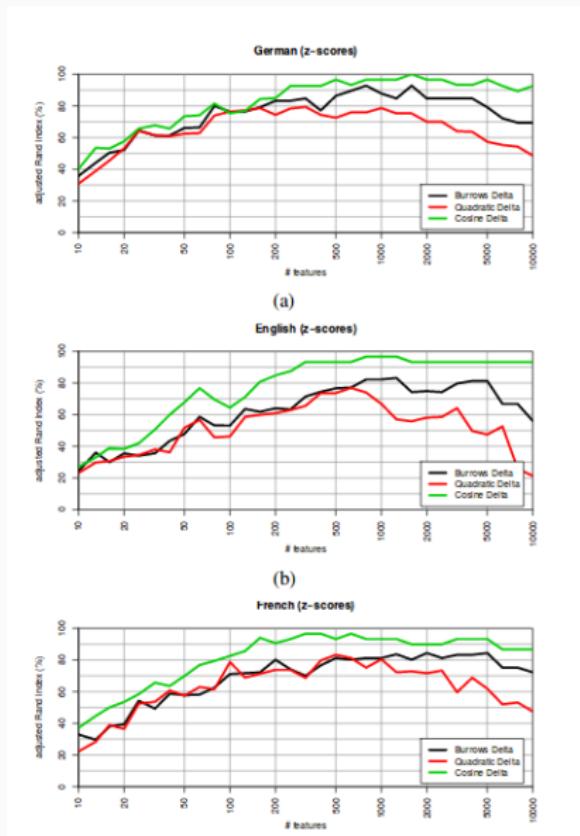
Articles récents sur ce sujet :

- Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, et Thorsten Vitt. « Understanding and explaining Delta measures for authorship attribution ». *Digital Scholarship in the Humanities* 32-2 (2017), ii4–ii16, https://academic.oup.com/dsh/article/32/suppl_2/ii4/3865676.
- M. Kestemont, J. Stover, M. Koppel, F. Karsdorp, et W. Daelemans, « Authorship Verification with the Minmax Metric », *Proceedings of the Digital Humanities 2016 conference*, 2016, http://www.dhbenelux.org/wp-content/uploads/2016/05/108_KestemontEtAl_FinalAbstract_DHBenelux2016_long.pdf.





VARIABILITÉ SELON LES LANGUES - EVERET AL., 2017



COMMENT DÉCOUPER ?

- Une fois la notion de distance éclaircie, reste à choisir une méthode pour regrouper ses données de manière pertinente.
- Le champ du **partitionnement de données** (en anglais, cluster analysis ou data clustering) regroupe un ensemble de techniques visant à regrouper des données éparses en sous-ensembles homogènes.
- Ces méthodes sont extrêmement nombreuses (probablement plus d'une centaine d'algorithmes utilisés), et plus ou moins adaptées à telle ou telle problématique spécifique.
- Elles varient selon les manières d'appréhender ce qu'est l'homogénéité, et selon les types de regroupement (hiérarchiques ou « à plat »).

PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

CENTROÏDE / CENTRE DE MASSE

- **Centroïde** : aussi appelé centre de masse.
- Notion issue de la physique, en particulier de la mécanique classique : c'est un point d'équilibre pour un certain objet.
- Pour un cercle ou une sphère, le centre de masse est tout simplement ce que l'on appelle le centre.
- De manière plus générale, c'est le barycentre des points pondérés.

LES K-MOYENNES

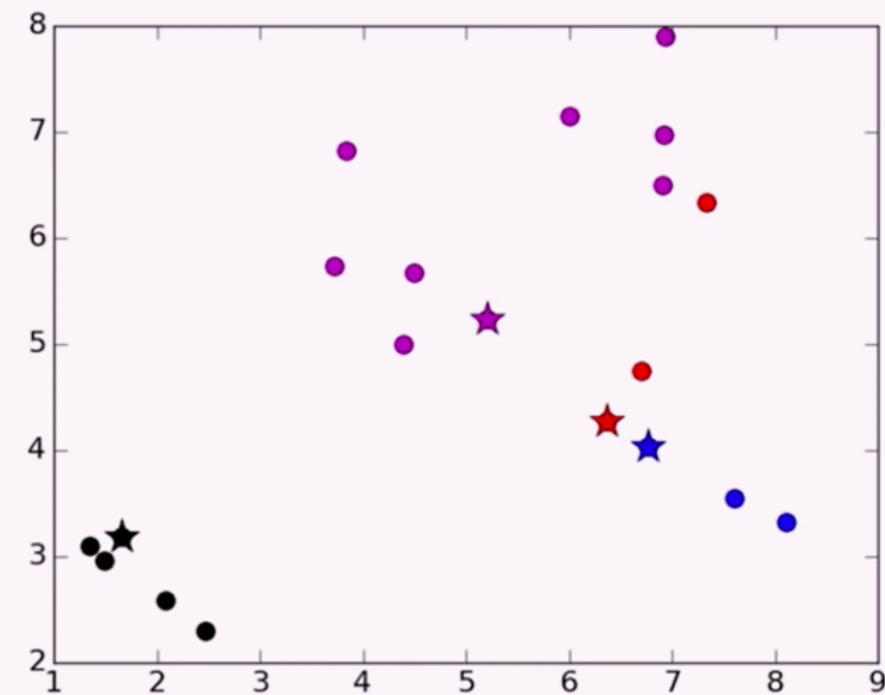
- La méthode dite des **k-moyennes** (en anglais k-means vise à regrouper en un nombre k de catégories des individus décrits par des variables quantitatives.
- Minimiser la dissimilarité à l'intérieur de chaque cluster.
- Mais sous contrainte bien sûr... D'où la nécessité de choisir le nombre de clusters.
- Comment choisir ce nombre ?
 - Connaissance a priori : je sais que les textes sont écrits par trois auteurs; à deux époques; selon quatre genres principaux etc. (danger : plusieurs manières d'être un texte d'une époque e.g., et dans ce cas, besoin d'un k plus élevé.
 - Recherche d'une bonne valeur pour k : en tester plusieurs et chercher la configuration qui, statistiquement ou intuitivement fait le plus sens.

L'ALGORITHME DES K-MOYENNES

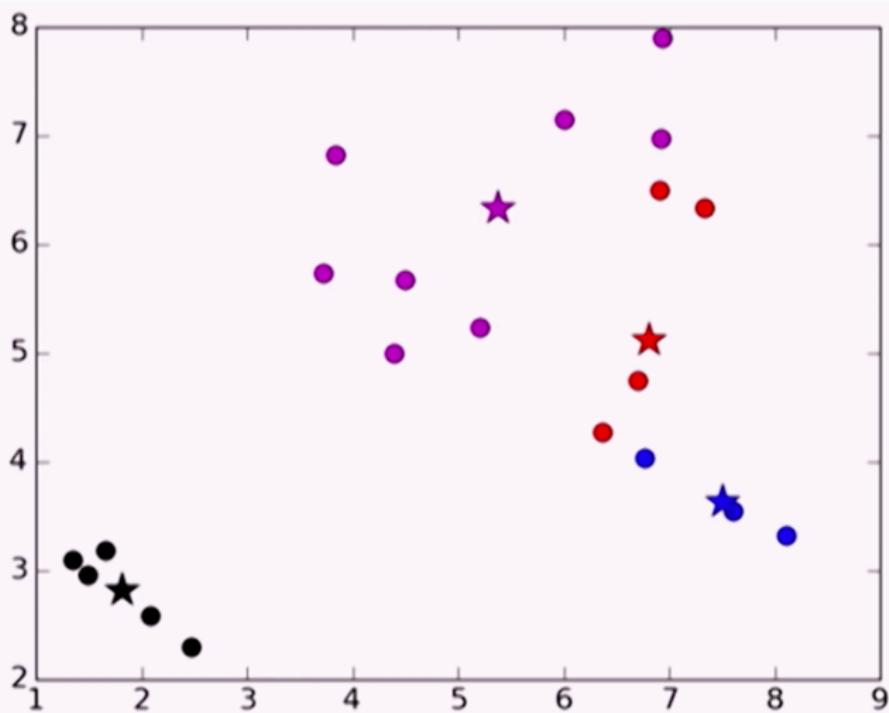
Algorithme :

1. Choisir k points qui représentent la position moyenne des partitions $m_1(1), , m_k(1)$ initiales (au hasard par exemple)
2. Assigner chaque observation à la partition la plus proche
3. Mettre à jour la moyenne de chaque cluster.
4. Recommencer.

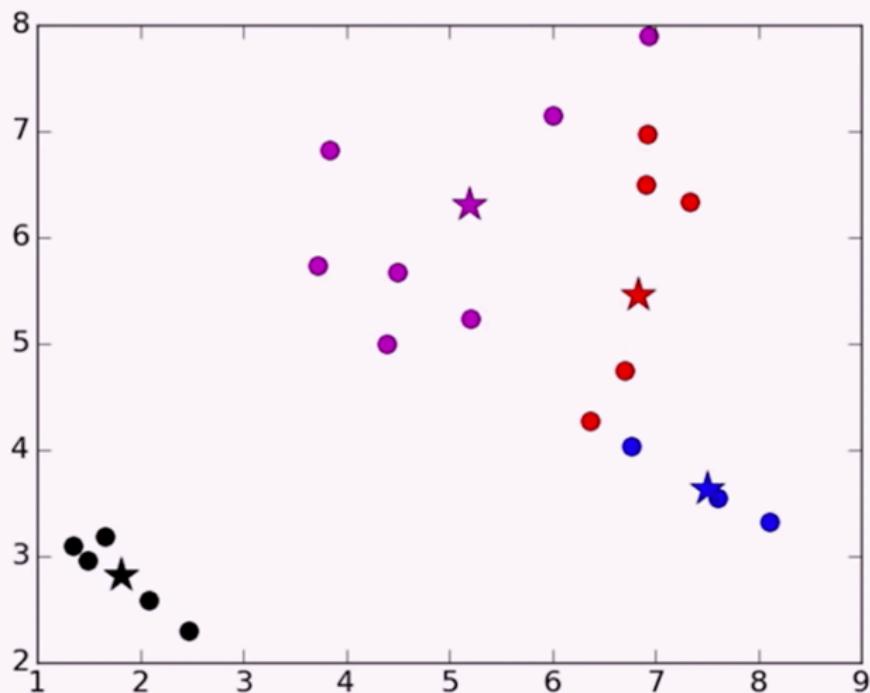
LES K-MOYENNES : K=4; POSITION ALÉATOIRE INITIALE



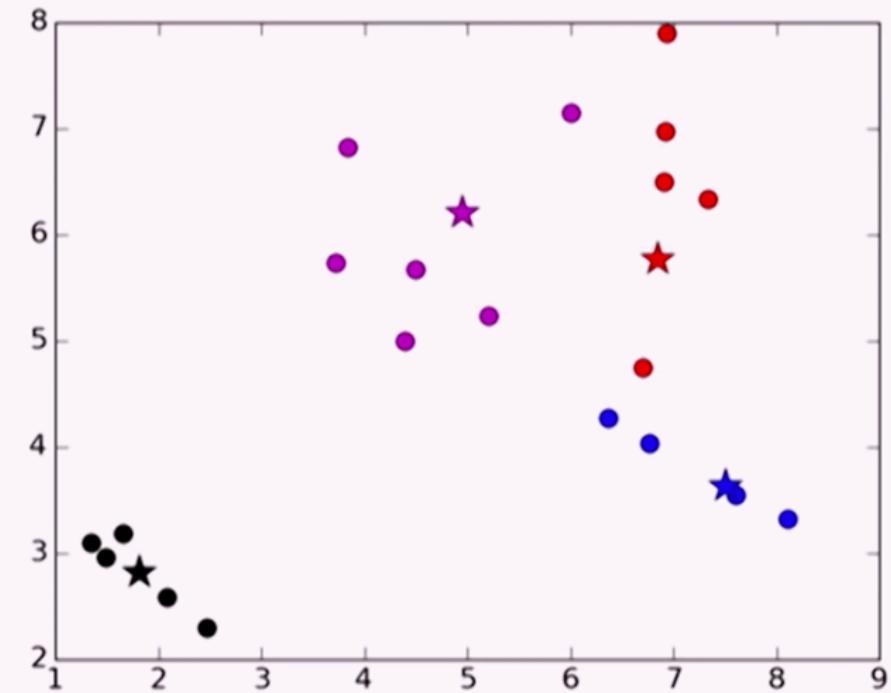
LES K-MOYENNES : K=4; PREMIÈRE ITÉRATION



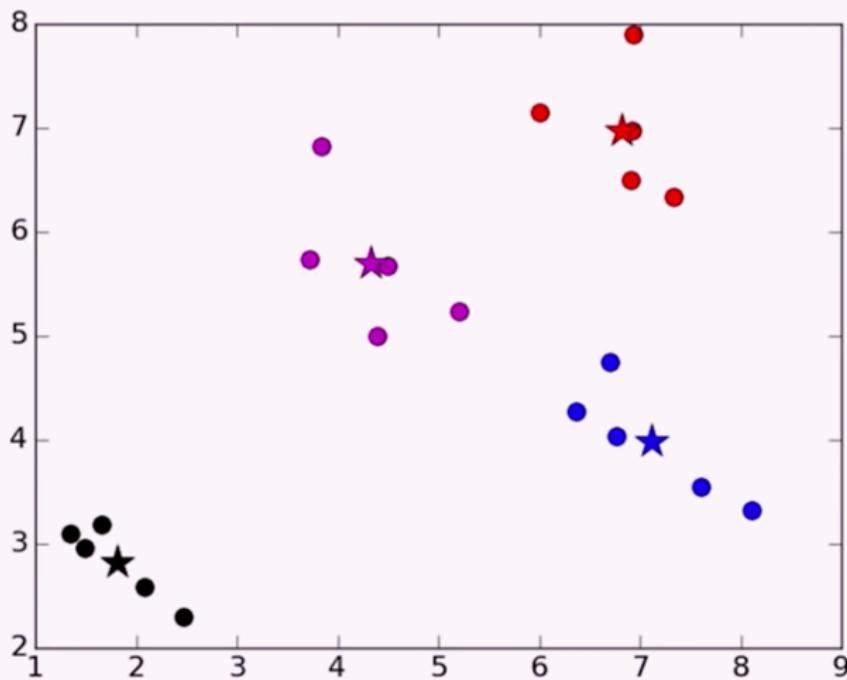
LES K-MOYENNES : K=4; SECONDE ITÉRATION



LES K-MOYENNES : K=4; TROISIÈME ITÉRATION



LES K-MOYENNES : K=4; ÉTAT FINAL



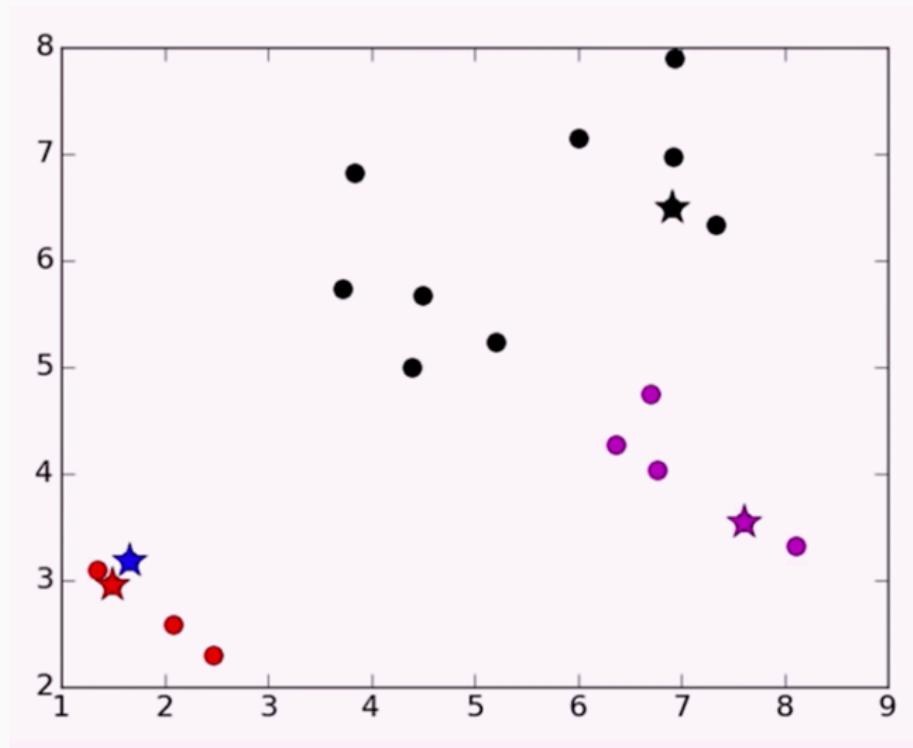
- Algorithme couramment implémenté peut prendre du temps pour le traitement de grandes bases de données.¹⁰
- Les positions aléatoires des centroïdes ont un double impact :
 - Sur le temps de calcul : s'ils "tombent mal", il faudra plus d'itérations pour que l'algorithme converge.
 - Sur le résultat : **les k-means ne suivent pas un algorithme déterministe**. Des positions initiales différentes pourront donc amener à un résultat différent.

Définition : algorithme déterministe

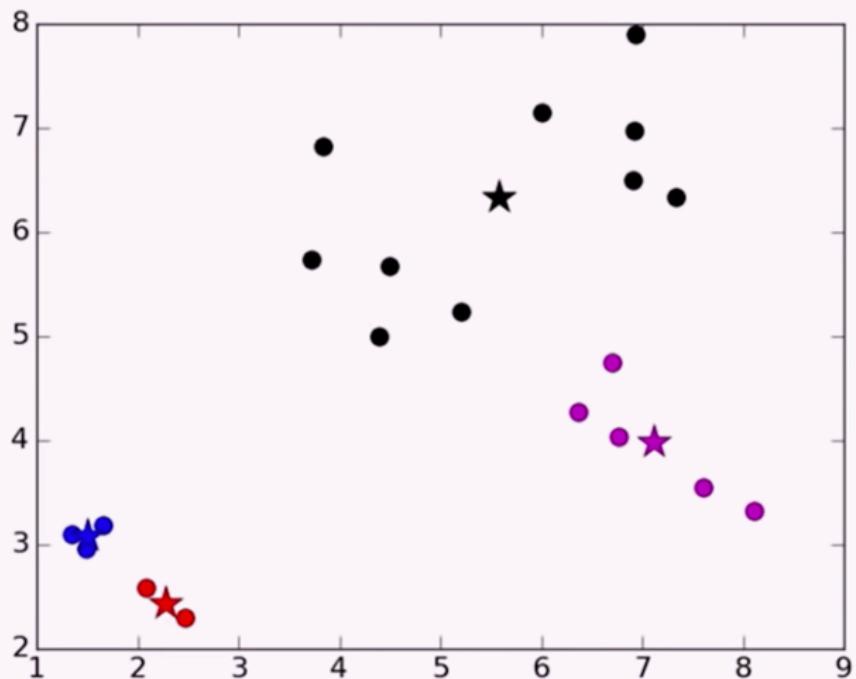
Un algorithme déterministe est un algorithme qui, pour une entrée particulière, produira toujours la même sortie, la machine sous-jacente passant toujours par la même séquence d'états.

10. David Arthur et Sergei Vassilvitskii, "Worst-Case and Smoothed Analysis of the ICP Algorithm, with an Application to the k-Means Method", SIAM J. Comput., vol. 39, n° 2, 2009, p. 766-782.

LES K-MOYENNES : AUTRE ALLOCATION INITIALE; K=4



AUTRE ALLOCATION INITIALE; K=4; ÉTAT FINAL



UN ALGORITHME GLOUTON

- Algorithme glouton, ou algorithme gourmand (greedy algorithm en anglais) : cherche, étape par étape, un optimum local.
- N'atteint pas toujours un optimum global.
- Mais ce type d'algorithme permet de faire baisser la complexité du calcul - ce qui diminue le temps de calcul, voire permet sa faisabilité.

LES K-MÉDOÏDES

- Méthode voisine des k-means, dans laquelle on calcule cependant la distance des points de la classe au "médoïde", c'est à dire au point central de la classe.
- Permet d'utiliser les méthodes d'agrégation et les mesures de distance que l'on souhaite, contrairement aux k-means (distance euclidienne, agglomération selon moyenne des distances)¹¹.
- Cette méthode est plus appropriée en cas de données présentant un effectif important d'individus aberrants. (On peut également utiliser les k-median à cet effet.)

11. Kaufman, L. et Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1 -Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416

UN ALGORITHME PLUS COMPLEXE

- Elle implique par contre des algorithmes plus **complexes**.
- Pour mesurer le temps mis pour résoudre un problème grâce à un algorithme donné, on effectuait au départ des mesures temporelles : combien de temps tel calcul prend etc.
- Mais contingent à une machine particulière...
- Solution : évaluer théoriquement la rapidité d'un algorithme.
- la théorie de la complexité hiérarchise la difficulté entre les problèmes algorithmiques en «classes de complexité ».

Pour $i \in \mathbb{N}$: nombre d'itérations de l'algorithme :

- Complexité des k-means : $O(n \times k \times i)$
- Complexité des k-médoïdes : $O(n^2 \times k \times i)$

CLASSES DE COMPLEXITÉ

Ordre de grandeur du temps nécessaire à l'exécution d'un algorithme d'un type de complexité

Temps	Type de complexité	Temps pour n = 5	Temps pour n = 10	Temps pour n = 20	Temps pour n = 50	Temps pour n = 250	Temps pour n = 1 000	Temps pour n = 10 000	Temps pour n = 1 000 000
$O(1)$	complexité constante	10 ns	10 ns	10 ns	10 ns	10 ns	10 ns	10 ns	10 ns
$O(\log(n))$	complexité logarithmique	10 ns	10 ns	10 ns	20 ns	30 ns	30 ns	40 ns	60 ns
$O(\sqrt{n})$	complexité racinaire	22 ns	32 ns	45 ns	71 ns	158 ns	316 ns	1 μ s	10 μ s
$O(n)$	complexité linéaire	50 ns	100 ns	200 ns	500 ns	2.5 μ s	10 μ s	100 μ s	10 ms
$O(n \log^*(n))$	complexité quasi-linéaire	50 ns	100 ns	200 ns	501 ns	2.5 μ s	10 μ s	100,5 μ s	10,05 ms
$O(n \log(n))$	complexité linéarithmique	40 ns	100 ns	260 ns	850 ns	6 μ s	30 μ s	400 μ s	60 ms
$O(n^2)$	complexité quadratique (polynomiale)	250 ns	1 μ s	4 μ s	25 μ s	625 μ s	10 ms	1 s	2.8 heures
$O(n^3)$	complexité cubique (polynomiale)	1.25 μ s	10 μ s	80 μ s	1.25 ms	156 ms	10 s	2.7 heures	316 ans
$2^{\text{poly}(\log(n))}$	complexité sous-exponentielle	30 ns	100 ns	492 ns	7 μ s	5 ms	10 s	3.2 ans	10^{20} ans
$2^{\text{poly}(n)}$	complexité exponentielle	320 ns	10 μ s	10 ms	130 jours	10^{59} ans
$O(n!)$	complexité factorielle	1.2 μ s	36 ms	770 ans	10^{48} ans
$2^{2^{\text{poly}(n)}}$	complexité doublement exponentielle	4.3 s	10^{278} ans

- On parle de **robustesse** d'un estimateur statistique pour décrire sa capacité à ne pas varier en fonction de petites variations dans les données, ou dans les paramètres choisis.
- **Un exemple simple : moyenne vs médiane :**
 - 1,2,3,4,5 : moyenne = 3; 1,2,3,4,5 : médiane = 3
 - 1,2,3,4,1000 : moyenne= 202; médiane=3
- Avantage des statistiques robustes : moins sensible à la présence de points aberrants (outliers)
- Variété d'outils robustes : médiane, k-médiannes/k-médoïdes, régressions robustes (sous R : `rlm` dans le package MASS)
- Cependant pas toujours nécessaire, et parfois coûteux en termes de calcul.

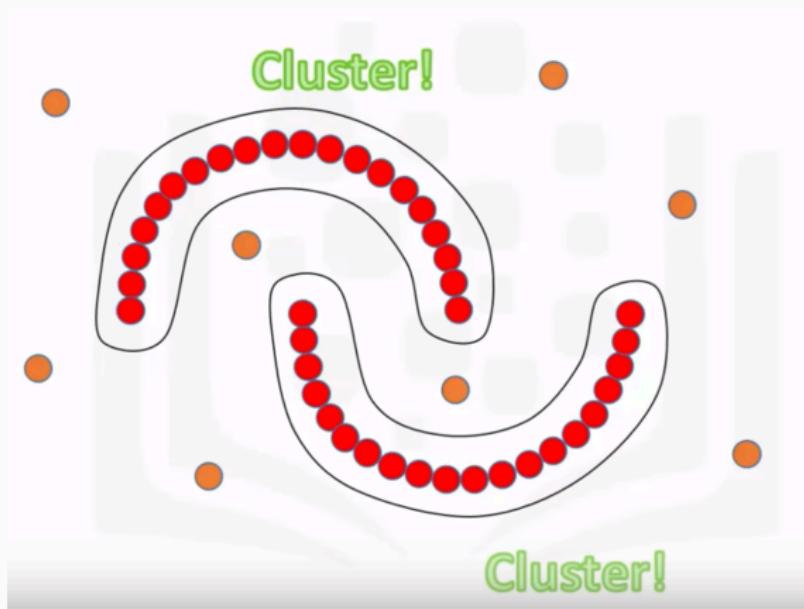
ESTER, Martin, KRIEGEL, Hans-Peter, SANDER, Jörg, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In : Kdd. 1996. p. 226-231.

- Le partitionnement en fonction de critères de densité démarre il y a seulement une vingtaine d'années, avec l'algorithme DBSCAN (Ester et al., 1996)
- **DBSCAN** : Density-Based Spatial Clustering and Application with Noise
- Méthodes nées pour d'autres usages, et très utiles pour les données spatiales.
- Avantage par rapport aux k-moyennes : mieux gérer le bruit.
- Peut-être utilisé dans le cas d'un corpus avec beaucoup d'hapax, de textes aberrants.

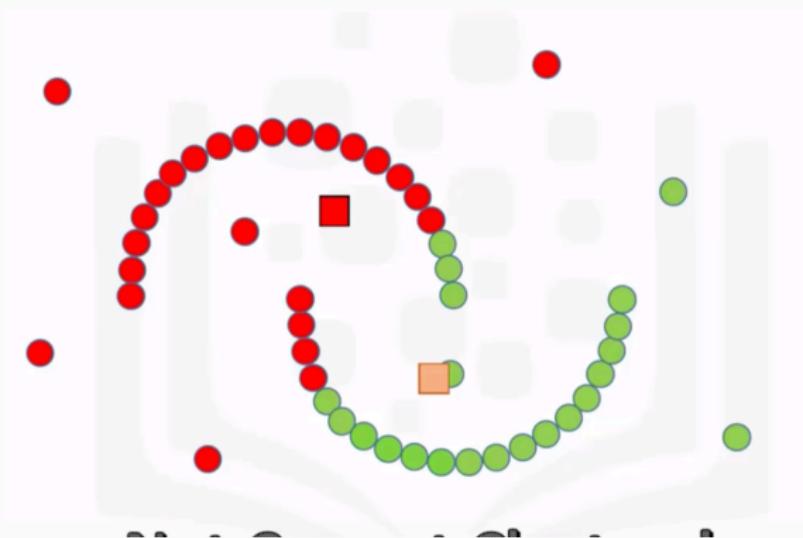
CONFIGURER DBSCAN

- Demande par contre de faire des choix, qui peuvent avoir un impact très important sur le résultat final.
- Il faut choisir combien de points dans un certain rayon son considérés comme suffisament denses.
- Deux paramètres à choisir : dans R (package dbscan), MinPts (nombre de points minimaux pour une zone dense) et EPS (distance maximale atteignable)
-

RÉSULTATS D'UN DBSCAN



RÉSULTATS D'UN K-MOYENNES SUR LES MÊMES DONNÉES



CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

- La **classification ascendante hiérarchique** réalise un regroupement sous forme de dendrogramme entre les différents individus d'un jeu de données.
- Elle est utilisable dans le cadre d'individus décrit par des **variables quantitatives**
- On peut toutefois tricher, en transformant des données qualitatives en données quantitatives - ce que nous verrons un peu plus tard.
(La méthode la plus employée est de réaliser une **analyse factorielle** à partir des données quantitatives, et de se servir des coordonnées des points dans les axes factoriels pour réaliser la CAH.)

MÉTHODE D'AGGLOMERATION

On doit ensuite choisir comment on regroupe les individus entre eux. Par défaut, le package propose d'utiliser la

Distance moyenne - "average" Calcule toutes les distances entre les différents points et en fait la moyenne

On peut aussi raisonner en terme d'extrémités

"Complete linkage" : calcule la distance maximale entre deux points

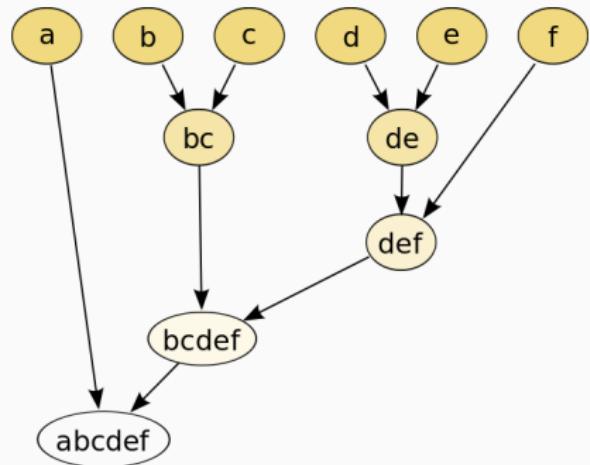
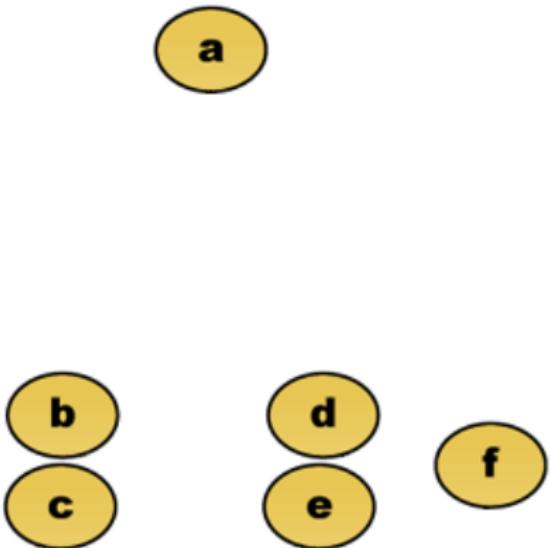
"Single linkage" : calcule la distance minimale entre deux points

L'algorithme le plus couramment utilisé est la :

Méthode de Ward : on calcule la distance entre les centres de gravité

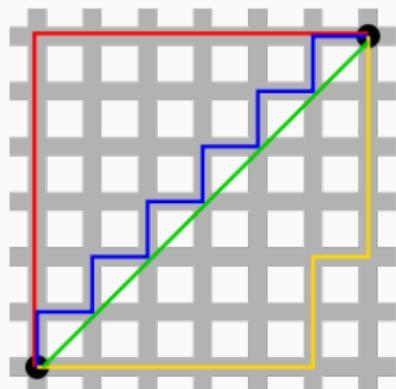
Ces méthodes se rejoignent seulement dans des cas très spécifiques.

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE



Distance de Manhattan

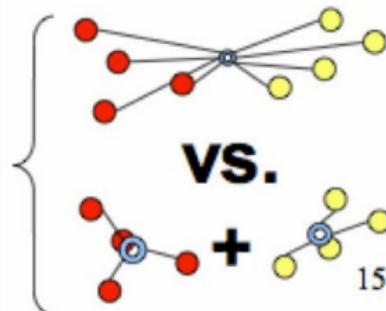
$$\sum_{i=1}^n |x_i - y_i|$$



Méthode de Ward

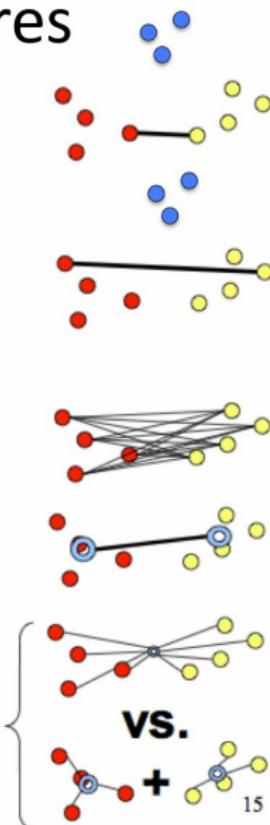
Minimise la distance aux centroïdes.

Favorise la constitution de groupes homogènes.

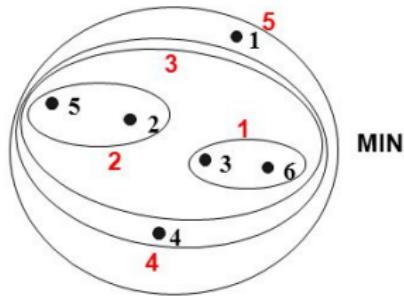


Cluster distance measures

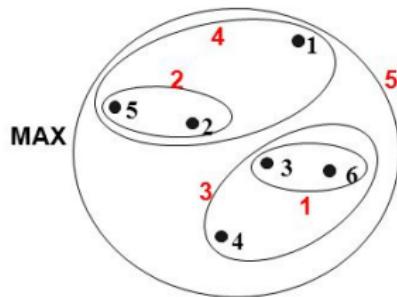
- Single link: $D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$
 - distance between closest elements in clusters
 - produces long chains a→b→c→...→z
- Complete link: $D(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$
 - distance between farthest elements in clusters
 - forces "spherical" clusters with consistent "diameter"
- Average link: $D(c_1, c_2) = \frac{1}{|c_1|} \frac{1}{|c_2|} \sum_{x_1 \in c_1} \sum_{x_2 \in c_2} D(x_1, x_2)$
 - average of all pairwise distances
 - less affected by outliers
- Centroids: $D(c_1, c_2) = D\left(\left(\frac{1}{|c_1|} \sum_{x \in c_1} \vec{x}\right), \left(\frac{1}{|c_2|} \sum_{x \in c_2} \vec{x}\right)\right)$
 - distance between centroids (means) of two clusters
- Ward's method: $TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$
 - consider joining two clusters, how does it change the total distance (TD) from centroids?



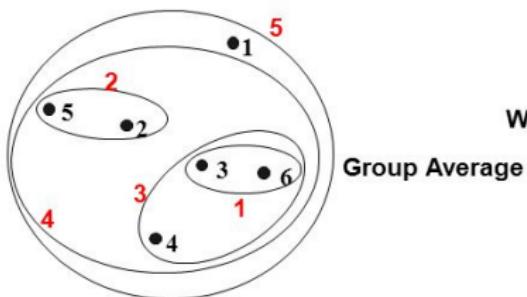
Hierarchical Clustering: Comparison



MIN

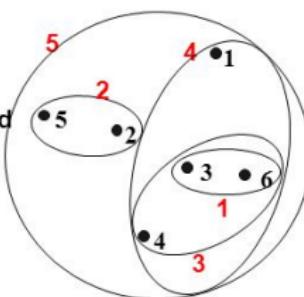


MAX

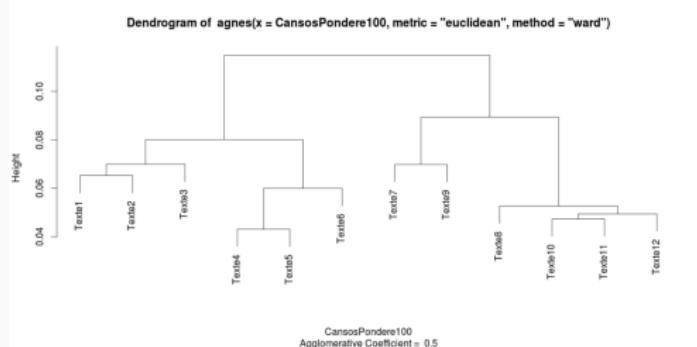


Group Average

Ward's Method



LIRE LES RÉSULTATS



hauteur (*height*)

Distance qui séparent les individus ou groupes.

Coefficient d'agglomération (ac)

Mesure de la qualité du partitionnement : pour chaque individu i , soit $m(i)$ la distance entre lui et le premier groupe auquel il s'agglomère divisé par la hauteur totale (à laquelle se fait la dernière fusion de groupe), on prend simplement la moyenne arithmétique des $m(i)$, $ac = \frac{1}{n} \sum_{i=1}^n m(i)$

N.B. : varie en fonction du nombre d'observations et de variables; ne pas utiliser pour comparer sur données différentes

PURETÉ D'UN CLUSTER

- La pureté d'un cluster donne le pourcentage d'individus classés "correctement", i.e., conformément aux attentes, selon un critère donné.
- Dans notre cas d'attribution d'autorité pour des pièces de théâtre : soit N le nombre de pièces, k le nombre de clusters, c_i un cluster observé, t_j un cluster théoriquement attendu. La pureté (PC) s'écrit alors

$$PC = \frac{1}{N} \sum_{i=1}^k \max_j |c_i \cap t_j|$$

INDICE DE DUNN

DUNN Joseph C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 1973.

- Autre mesure de la qualité : l'indice de Dunn
- Purement statistique cette fois-ci.
- Rapport entre la distance maximum qui sépare deux éléments classés ensemble et la distance minimum qui sépare deux éléments classés séparément.
- Varie entre 0 (pire classification) et l'infini (classification parfaite).

CAH AVEC R ET CLUSTER

```
agnes(x, diss = inherits(x, "dist"),
      metric = "euclidean | manhattan",
      stand = FALSE,
      method = "average|single|complete|ward|weighted",
      par.method,
      keep.diss = n < 100, keep.data = !diss)

1 #importer la bibliotheque cluster
2 library(cluster)
3 #calculer la CAH
4 maCAH = agnes(theatre,
5                 metric ="manhattan", method="ward")
6 #consulter les resultats
7 summary(maCAH)
8 #les tracer
9 plot(maCAH)
```

PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

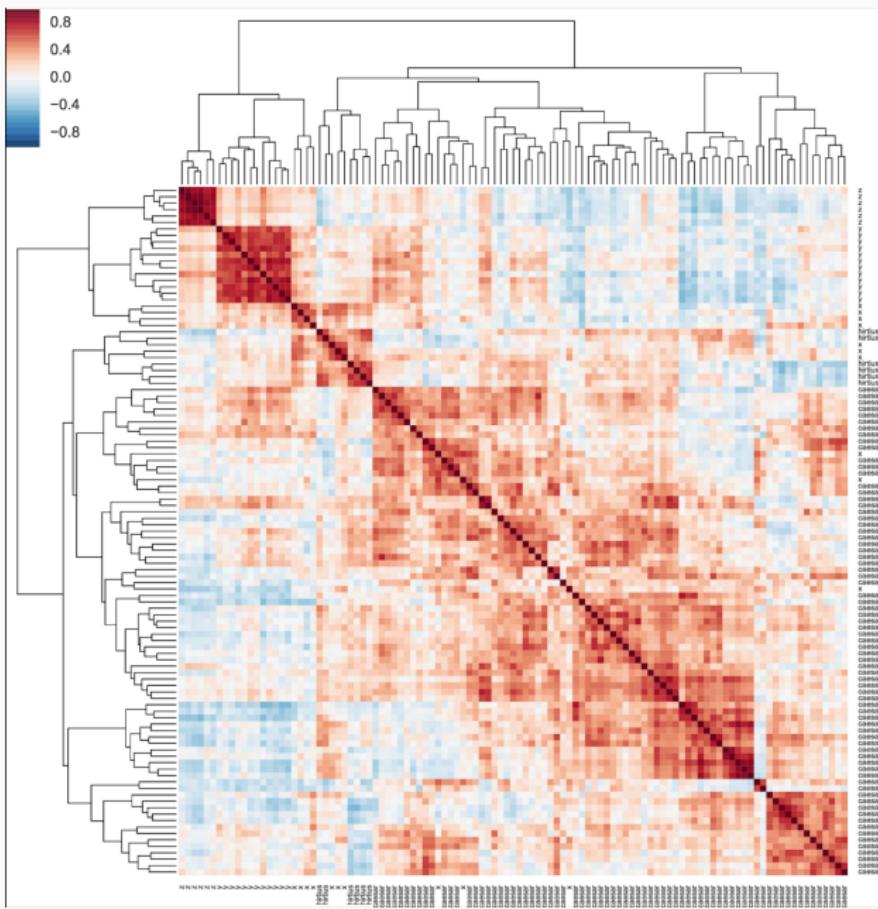
Cartes thermiques

Réseaux

CARTE THERMIQUE

- Aussi appelée carte de fréquentation, ou heatmap
- Représentation graphique faisant correspondre à l'intensité d'une grandeur variable un nuancier de couleurs sur une matrice à deux dimensions.
- Souvent utilisée dans le cadre de cartes par exemple.
- Mais bien sûr pas obligatoire.
- Possibilité de combiner CAH et cartes thermiques en une seule visualisation.
- Exemple d'utilisation : KESTEMONT, Mike, STOVER, Justin, KOPPEL, Moshe, et al. Authenticating the writings of Julius Caesar. Expert Systems with Applications, 2016, vol. 63, p. 86-96.





PLAN

Quelles données pour la stylométrie ?

Caractéristiques utilisées

Échantillonnage et sélection

Fréquences et pondérations

Analyse exploratoire de données

Analyse par réduction des dimensions

Présentation des différentes analyses

Distanciation littéraire et clusters

Mesures les distances

Méthodes de partitionnement

Cartes thermiques

Réseaux

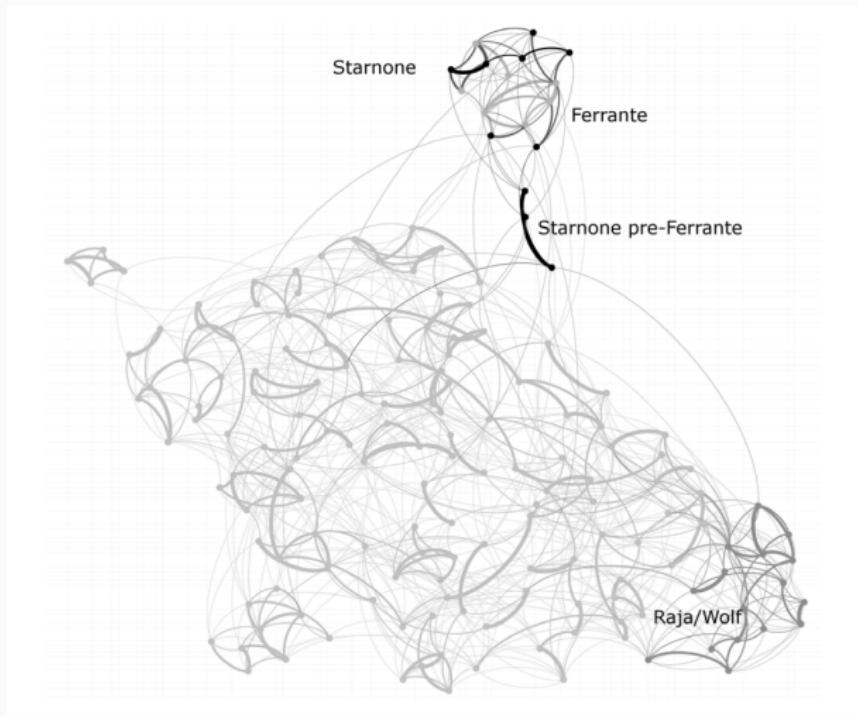
EDER, Maciej. Visualization in stylometry : Cluster analysis using networks. *Digital Scholarship in the Humanities*, 2017, vol. 32, no 1, p. 50-64.

Le principe de cette visualisation proposée par Maciej Eder est simple :

- On calcule, selon un critère de distance, les textes les plus proches de chaque texte.
- On choisit les k textes les plus proches d'un texte, ou les textes les suffisamment proches (on définit un seuil de distance maximale entre deux textes)
- On trace à chaque fois un lien allant du texte vers eux.
- On répète l'opération pour tous les textes de la base de données.

QUI EST ELENA FERRANTE?

RYBICKI, Jan. "Partners in Life, Partners in Crime?". UPPADO, 2018, p. 111.



ANNEXES

RÉFÉRENCES BIBLIOGRAPHIQUES

Sur les propriétés du lexique et l'approche sémantique :

Voir Crédit et utilisation de corpus de textes médiévaux, Minorque, 16 - 24 Septembre 2014, en ligne :

<http://www.glossaria.eu/minorque/programme.html>, et particulièrement, Alain Guerreau, « Les propriétés numériques étranges du vocabulaire : Introduction à la statistique lexicale », ainsi que Id., « Sémantique historique et statistique ».

Voir aussi : Alain Guerreau, Statistiques pour historiens : Méthodes pratiques de statistique et de cartographie, <http://elec.enc.sorbonne.fr/statistiques/stat2004.pdf>,

part., chap. 9. «Distributions lexicales», 10.«Sémantique et formalisation», 11. «Statistique lexicale et érudition».

RÉFÉRENCES : ÉTUDES ATTRIBUTIONNISTES

Champ de la stylométrie et de l'authorship attribution (historique ou légiste) extrêmement vaste. **Quelques travaux de synthèse :**

KOPPEL (Moshe), Schler (Jonathan) et Argamon (Shlomo), «Computational methods in authorship attribution», *Journal of the American Society for Information Science & Technology*, 60-1, (2009), p. 9-26.

RUDMAN (Joseph), «The State of Non-Traditional Authorship Attribution Studies—2012 : Some Problems and Solutions », dans *English Studies*, 93-3, (2012), p. 259-274, en ligne : <http://www.tandfonline.com/doi/abs/10.1080/0013838X.2012.668785>.

STAMATATOS (E.), «A Survey of modern authorship attribution methods», *Journal of the American Society for information Science and Technology*, 60-3, (2009), p. 538-556.

Un exemple désormais assez fameux :

KESTEMONT (Mike), Moens (Sara) et Deploige (Jeroen), «Collaborative authorship in the twelfth century : A stylometric study of Hildegard of Bingen and Guibert of Gembloux», *Literary and Linguistic Computing*, (2013), en ligne, <http://llc.oxfordjournals.org/content/early/2013/10/26/llc.fqt063>.

BIBLIOGRAPHIE

Création et utilisation de corpus de textes médiévaux, Minorque, 16 - 24 Septembre 2014, en ligne :

<http://www.glossaria.eu/minorque/programme.html>.

DUMAIS (Susan T.), « Improving the retrieval of information from external sources », Behavior Research Methods, Instruments, & Computers, 23-2, 1991, 229-236, <http://link.springer.com/article/10.3758%2FBF03203370>.

Guerreau (Alain), Statistiques pour historiens : Méthodes pratiques de statistique et de cartographie, <http://elec.enc.sorbonne.fr/statistiques/stat2004.pdf>,

part., chap. 9. «Distributions lexicales», 10.«Sémantique et formalisation», 11. «Statistique lexicale et érudition».

KESTEMONT (Mike), Moens (Sara) et Deploige (Jeroen), «Collaborative authorship in the twelfth century : A stylometric study of Hildegard of Bingen and Guibert of Gembloux», Literary and Linguistic Computing, (2013), en ligne, <http://llc.oxfordjournals.org/content/early/2013/10/26/llc.fqt063>.

BIBLIOGRAPHIE (SUITE)

KOPPEL (Moshe), Schler (Jonathan) et Argamon (Shlomo),
«Computational methods in authorship attribution», Journal of the
American Society for Information Science & Technology, 60-1, (2009),
p. 9-26.

RUDMAN (Joseph), « The State of Non-Traditional Authorship
Attribution Studies—2012 : Some Problems and Solutions », dans
English Studies, 93-3, (2012), p. 259-274, en ligne :
[http://www.tandfonline.com/doi/abs/10.1080/
0013838X.2012.668785](http://www.tandfonline.com/doi/abs/10.1080/0013838X.2012.668785).

STAMATATOS (E.), «A Survey of modern authorship attribution
methods», Journal of the American Society for information Science
and Technology, 60-3, (2009), p. 538-556.