Analyse de données textuelles

François Husson

Unité pédagogique de mathématiques appliquées - l'institut Agro

husson@agrocampus-ouest.fr

Analyse de données textuelles

L'analyse textuelle

Le texte comme matériel statistique

Analyse textuelle : analyse statistique de corpus, ou ensemble de documents, en langage naturel.

Objectifs :

- extraire les thèmes
- évaluer et visualiser proximités entre documents et entre mots
- classer / catégoriser les documents
- étudier l'évolution temporelle d'un corpus, l'organisation d'un texte long

et ainsi :

- synthétiser les documents par des mots/phrases caractéristiques
- organiser/classer des bases de documents
- déterminer la structure/forme d'un corpus ou d'un document
- découvrir et dater les changements
- connecter les documents avec des données complémentaires

L'analyse textuelle

Format des données textuelles en entrée (= point de vue)

- corpus de questions ouvertes : une ligne = une réponse
  Questions ouvertes permettent de :
  - recueillir une information spontanée
  - comprendre la réponse à une question fermée (aimez vous les maths ? Pourquoi ?)
  - explorer des domaines mal connus
- analyse d'un discours : une ligne = un paragraphe
- un corpus de sentences au tribunal : une ligne = une sentence

Point de départ :

- le corpus est divisé en documents (a priori ou de façon « artificielle »)
- un document est une séquence d'occurrences de « mot »
- identifier les différents mots et compter leur fréquence
- comparer la distribution des mots dans les documents du corpus

Qu'est-ce qu'un mot ?

Cela pose la question de la définition de « mot »
Exemple : L'enfant lit une bande dessinée. Il aime lire dans son lit.

- forme fléchie (ou forme graphique) l enfant lit (. . .) lit (. . .) lire
- lemme (=entrée du dictionnaire) : leAr enfantNm lireVb (. . .) litNm
- segment répété : il aime, aime lire, il aime lire
- unité lexique complexe : bande dessinée

Prétraitements en analyse textuelle

Se pose la question des prétraitements :

- correction orthographique
- réduction des majuscules en minuscules (parfois conservation des majuscules correspondant aux noms propres)
- harmonisation des graphies : Edf versus Electricité de France
- dans les interviews, séparer les questions des réponses, etc.
- différentier les suites de chiffres en date, montant monétaire, ...
- éventuelle lemmatisation (regrouper singulier et pluriel, les différentes formes d'un même verbe)
- éventuelle stemmatisation : regroupement des formes graphiques de même racine ou stem (ex. malade, malades, maladie et maladies regroupés sous le stem malad)
- conserver ou non les mots-outils

Analyse de données textuelles

AFC pour l'analyse textuelle

AFC directe du tableau lexical :

- conserver les mots cités suffisamment souvent
- visualiser les mots dont la contribution est supérieure à la moyenne sur le plan
- Mettre les segments répétés en supplémentaires

Analyse d'un tableau lexical agrégé

Tableau lexical agrégé : regrouper les documents en catégories :

- pour des questions ouvertes : catégories sont les combinaisons sexe-âge
- pour des œuvres littéraire : courants littéraires
- etc.

AFC du tableau lexical agrégé :

- même démarche mais sur un tableau avec beaucoup moins de lignes
- des tableau lexicaux agrégés peuvent avoir les mêmes lignes, même si les individus n'utilisent pas les mêmes mots (par exemple répondants de nationalités différentes). On pourra alors comparer la position des catégories d'un pays à l'autre en faisant des AFC séparées ... ou en utilisant l'AFM (cours suivant)

Analyse de données textuelles

Un exemple en linguistique

- Aragon (23 textes) : FeuJoie, Perpétuel, Destinées, Snark, Peinture, ...
- Balzac (49 textes) : *Chouans*, *Physiologie*, *Vendetta*, *Gobseck*, ...
- Corneille (34 textes) : *Mélite*, *Clitandre*, *Veuve*, *Gelerie*,*Suivante*, ...
- ...

Un exemple en linguistique

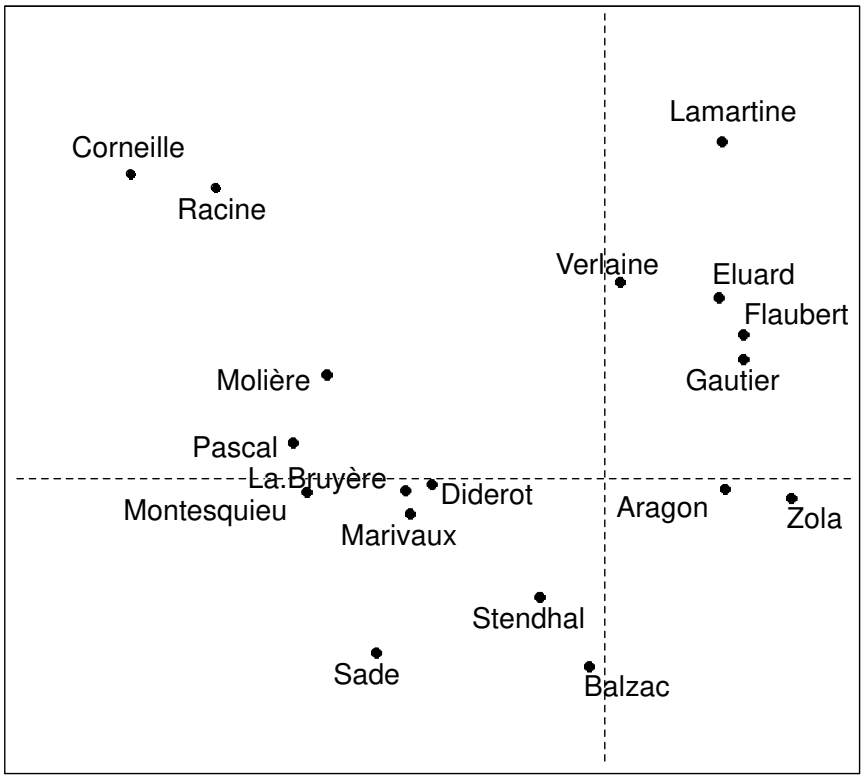
- Aragon (23 textes) : FeuJoie, Perpétuel, Destinées, Snark, Peinture, ...
- Balzac (49 textes) : *Chouans*, *Physiologie*, *Vendetta*, *Gobseck*, ...
- Corneille (34 textes) : *Mélite*, *Clitandre*, *Veuve*, *Gelerie*,*Suivante*, ...
- ...

On conserve les mots cités au moins 100 fois

accord	264	0	88	44	...
affaire	1029	2040	74	154	...
âge	545	629	92	108	
ah	219	0	0	0	
air	2093	2009	95	191	
allemagne	366	0	0	0	
allemand	476	0	0	0	
amant	303	760	566	0	
âme	478	2190	1101	240	
ami	1090	2583	307	407	
amour	1374	3286	1791	167	
an	1812	3009	112	182	
anglais	315	0	0	0	
. . .					

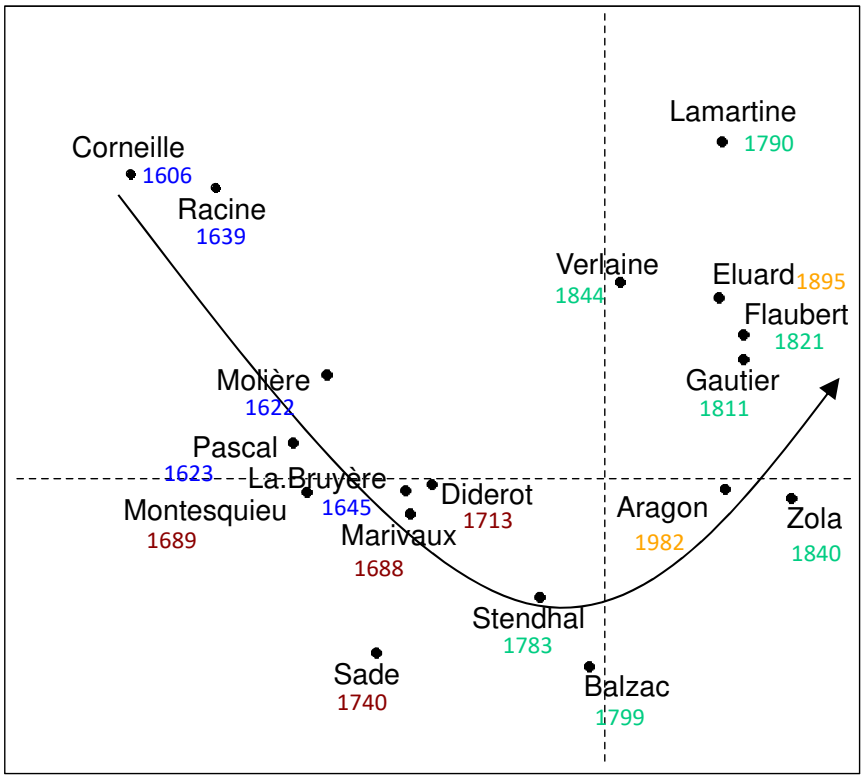
978 mots

Analyse des correspondances : visualisation des auteurs



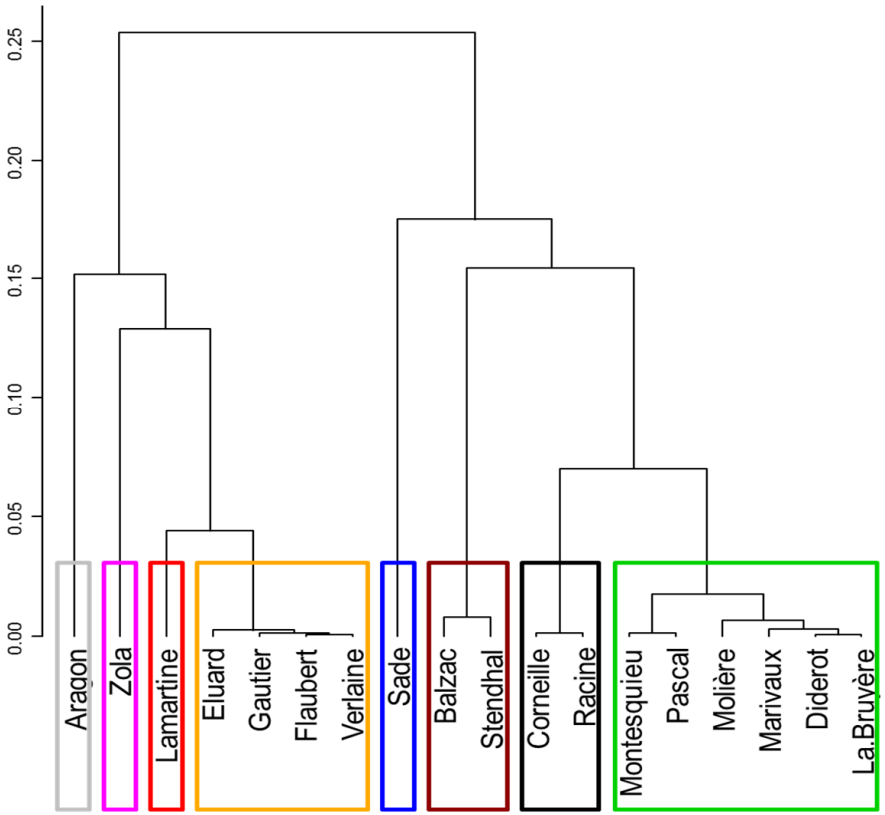
Avec l'AFC, les auteurs sont d'autant plus proches qu'ils emploient les mots dans les mêmes proportions, i.e. qu'ils s'intéressent aux mêmes sujets et ont les mêmes préoccupations

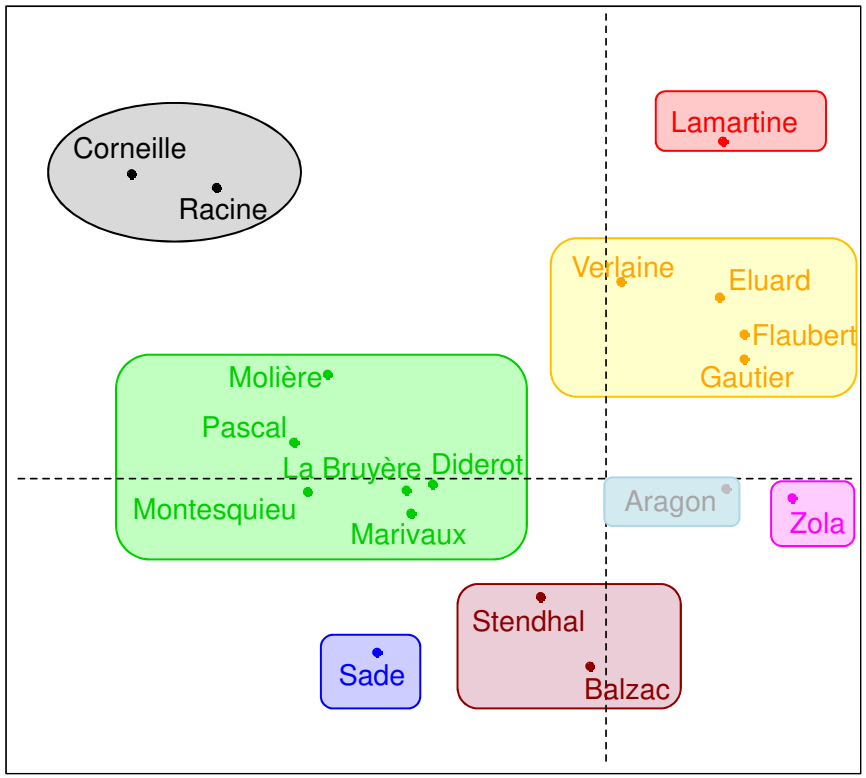
Analyse des correspondances : visualisation des auteurs



- Corneille et Racine sont proches et très éloignés de Zola.
  Ce sont 2 auteurs classiques du 17ème tandis que Zola est un naturaliste du 19ème
- Évolution du vocabulaire selon les siècles

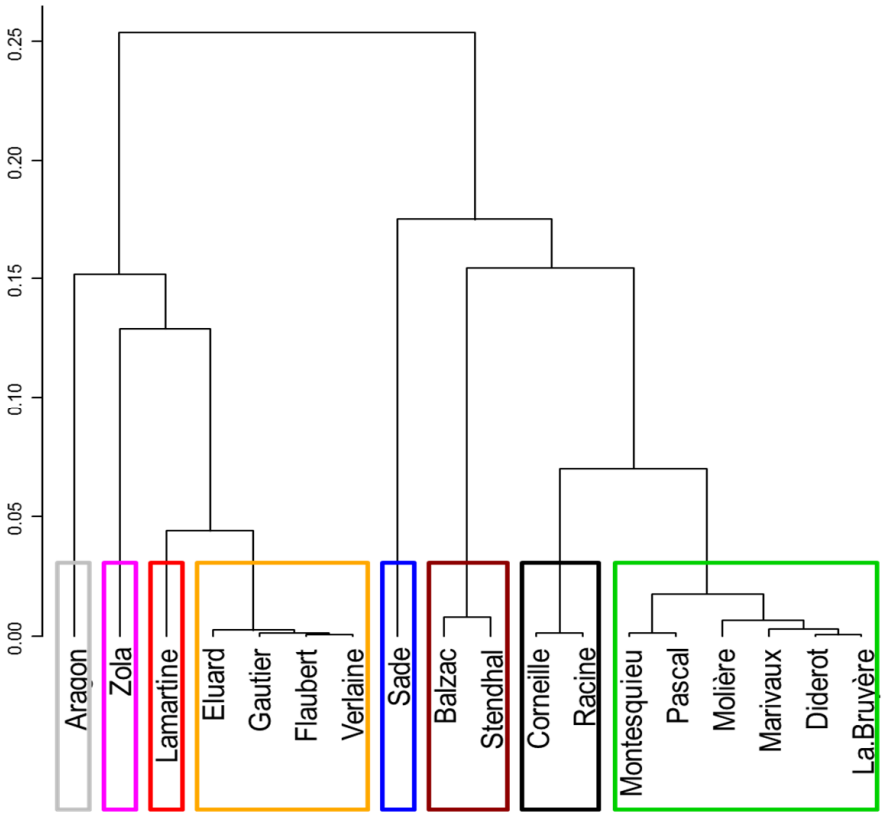
Classification des auteurs

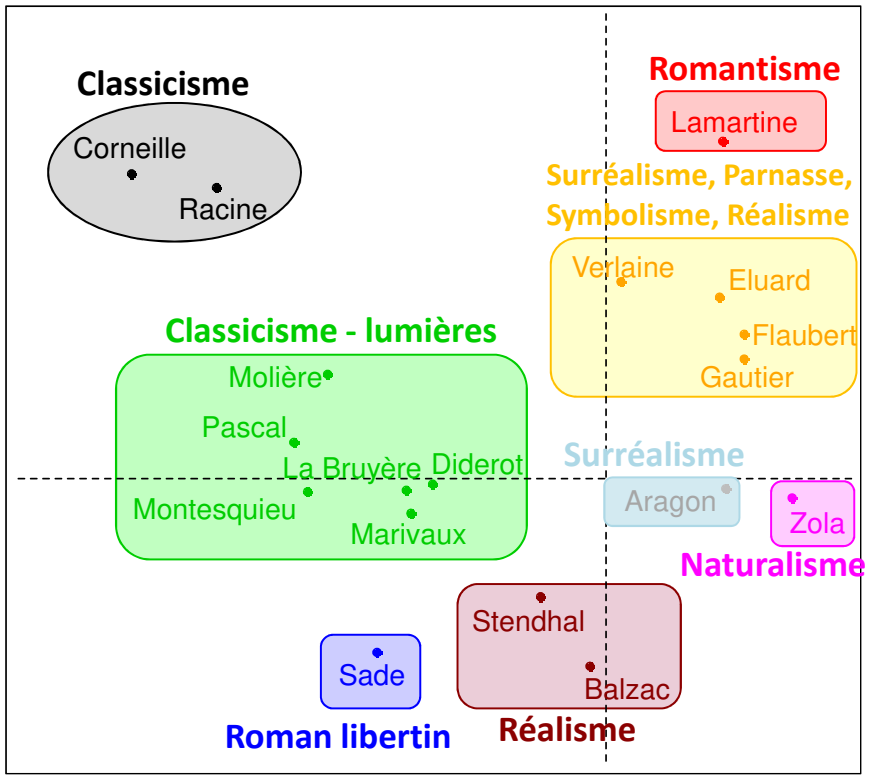




- La classification retrouve des classes d'auteurs connues

Classification des auteurs





- Stendhal et Balzac (réalistes) sont très éloignés de Lamartine (romantique). On retrouve ici que les auteurs réalistes ont un point commun : s'éloigner des excès romantiques !
- Points communs naturalistes / réalistes : montrer la société telle qu'elle est, le roman devient le miroir de la société

Analyse de données textuelles

Mots caractéristiques

Le mot *Victoire* caractérise-t-il les œuvres de Corneille-Racine ?

Mots caractéristiques

Le mot *Victoire* caractérise-t-il les œuvres de Corneille-Racine ?

	Corneille-Racine	Pas Corneille -Racine	Total
Victoire	248	55	303
Pas Victoire	80 817	1 507 008	1 587 825
Total	81 065	1 507 063	1 588 128

Principe : une urne contient 1 588 128 boules, sur 303 boules est écrit le mot *Victoire*, on tire 81 065 boules.

Mots caractéristiques

Le mot *Victoire* caractérise-t-il les œuvres de Corneille-Racine ?

	Corneille-Racine	Pas Corneille -Racine	Total
Victoire	248	55	303
Pas Victoire	80 817	1 507 008	1 587 825
Total	81 065	1 507 063	1 588 128

Principe : une urne contient 1 588 128 boules, sur 303 boules est écrit le mot *Victoire*, on tire 81 065 boules.

H_0 : la fréquence F du mot *Victoire* suit une loi $\mathcal{H}(1\,588\,128, 303, 81\,065)$

Mots caractéristiques

Le mot *Victoire* caractérise-t-il les œuvres de Corneille-Racine ?

	Corneille-Racine	Pas Corneille -Racine	Total
Victoire	248	55	303
Pas Victoire	80 817	1 507 008	1 587 825
Total	81 065	1 507 063	1 588 128

Principe : une urne contient 1 588 128 boules, sur 303 boules est écrit le mot *Victoire*, on tire 81 065 boules.

H_0 : la fréquence F du mot *Victoire* suit une loi $\mathcal{H}(1\,588\,128, 303, 81\,065)$

Peut-on remettre en cause cette hypothèse ?
$\Longrightarrow$ 248 provient-il d'une loi hypergéométrique $\mathcal{H}(1\,588\,128, 303, 81\,065)$?

Mots caractéristiques

Le mot *Victoire* caractérise-t-il les œuvres de Corneille-Racine ?

	Corneille-Racine	Pas Corneille -Racine	Total
Victoire	248	55	303
Pas Victoire	80 817	1 507 008	1 587 825
Total	81 065	1 507 063	1 588 128

Principe : une urne contient 1 588 128 boules, sur 303 boules est écrit le mot *Victoire*, on tire 81 065 boules.

H_0 : la fréquence F du mot *Victoire* suit une loi $\mathcal{H}(1\,588\,128, 303, 81\,065)$

Peut-on remettre en cause cette hypothèse ?
$\Longrightarrow$ 248 provient-il d'une loi hypergéométrique $\mathcal{H}(1\,588\,128, 303, 81\,065)$?

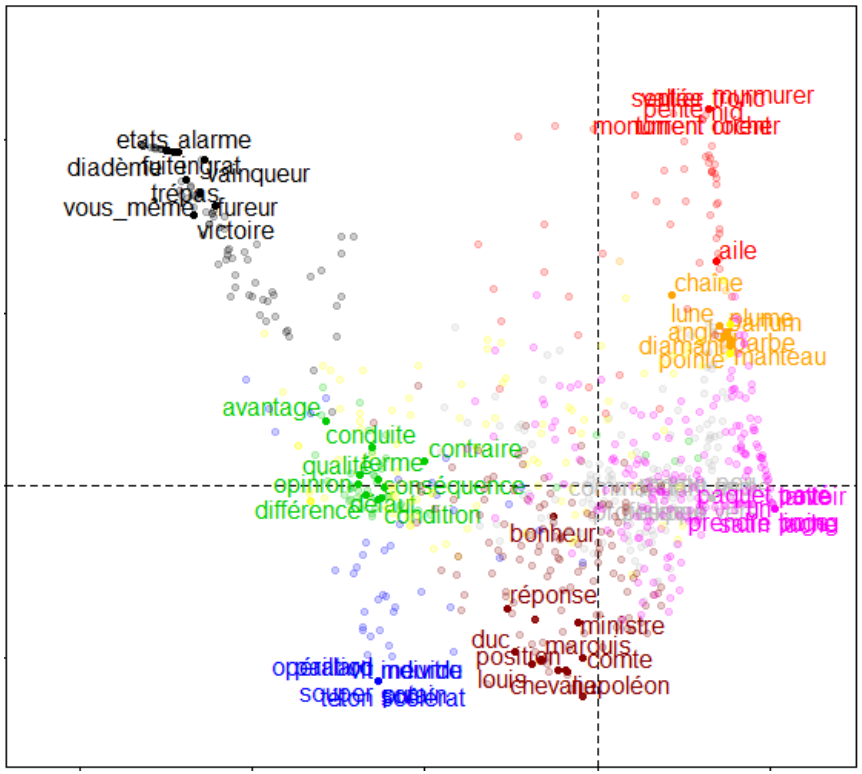
Corneille-Racine

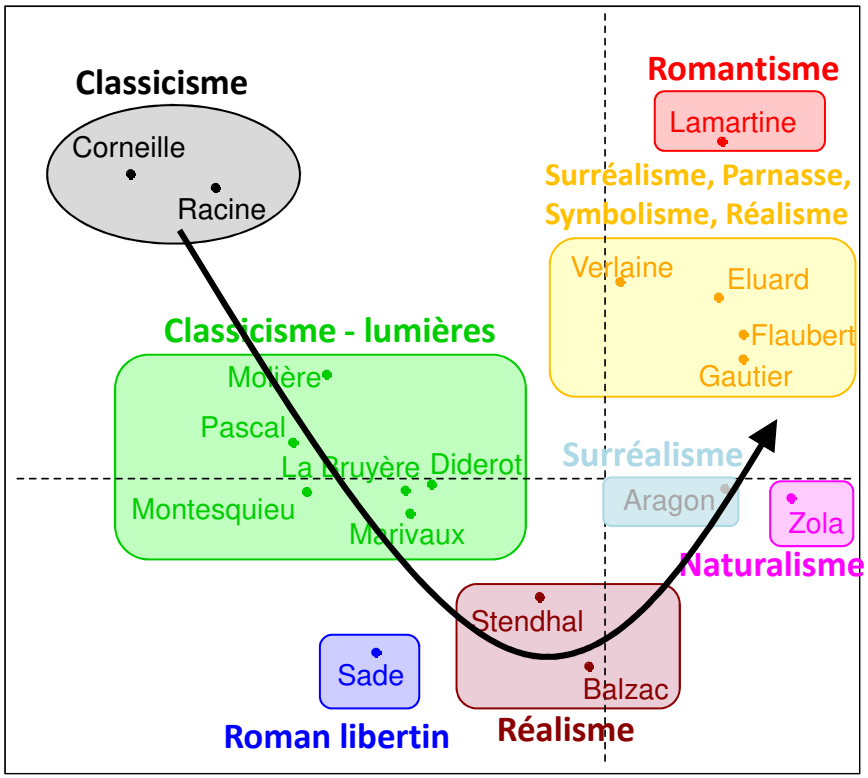
	Intern	% glob	% Intern freq	Glob freq	p.value	v.test
Victoire	0.306	0.019	248	303	2.28e-06	4.727

$$\frac{248}{81\,065} = 0.00306\,;\ \frac{303}{1\,588\,128} = 0.0001908\,;\ P\left[F \geq 5 \mid F \sim \mathcal{H}(1\,588\,128, 303, 81\,065)\right]$$

$\Longrightarrow$ Rejet de H_0 : *Victoire* est sur-employé par Corneille-Racine

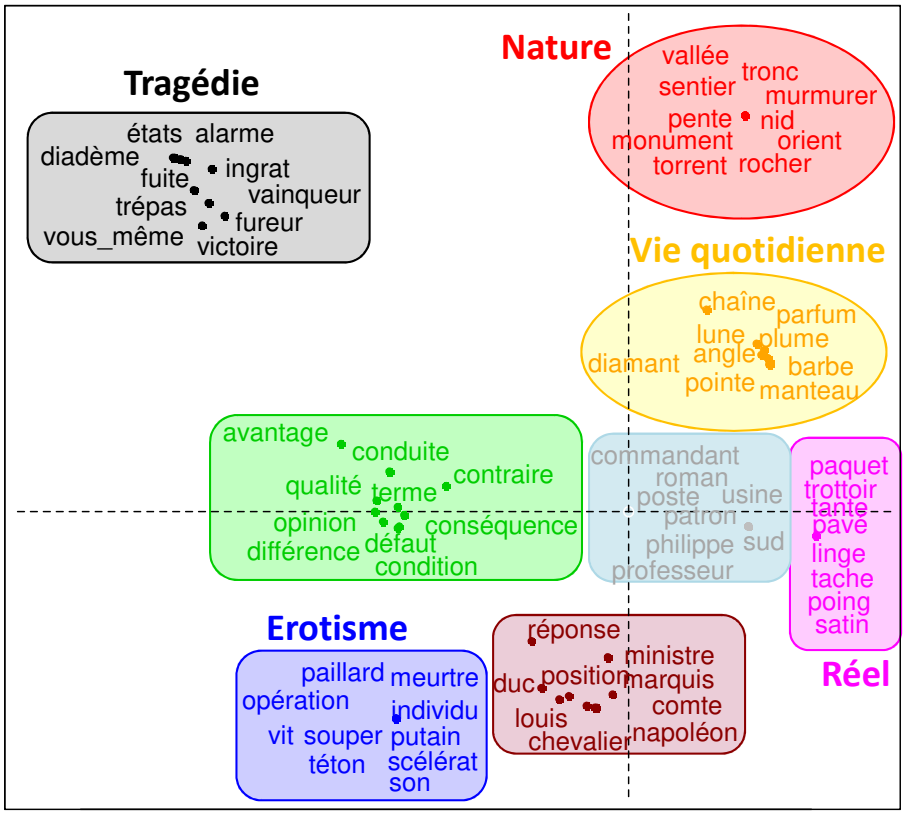
Caractérisation par les mots

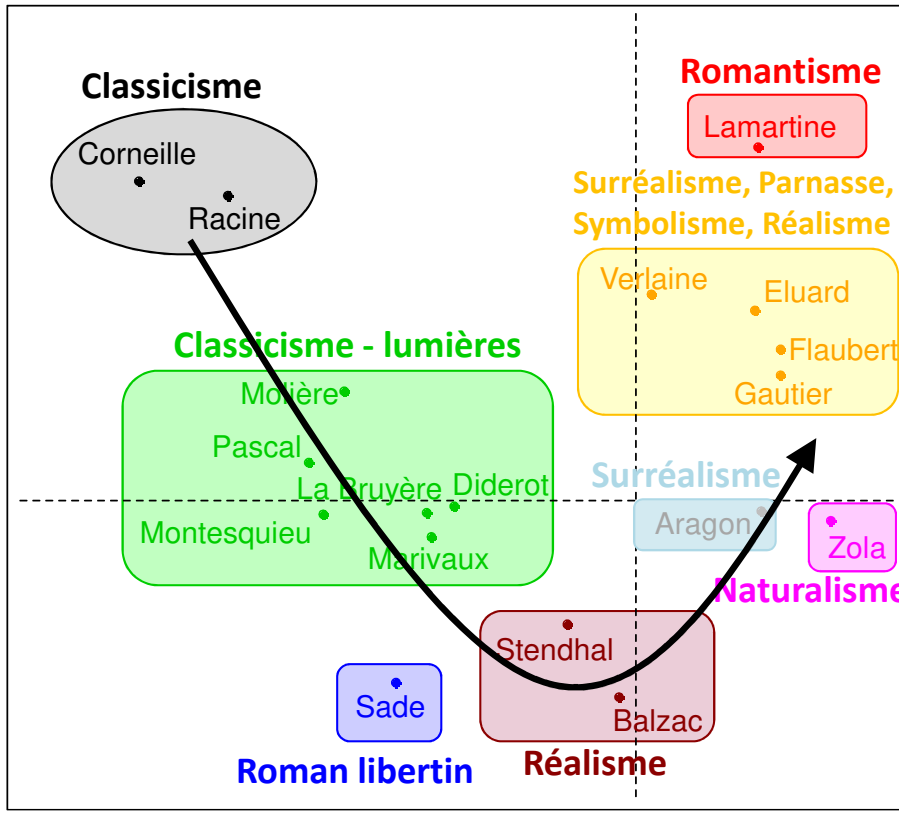




Les mots permettent de caractériser les sujets de prédilection des auteurs et les courants littéraires

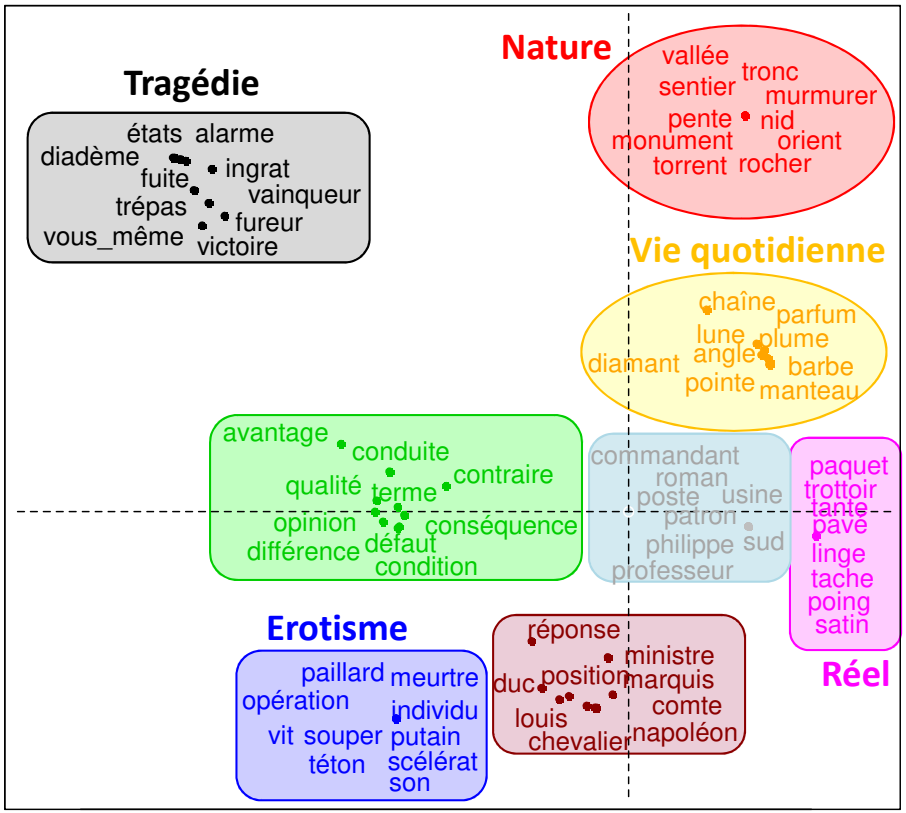
Caractérisation par les mots

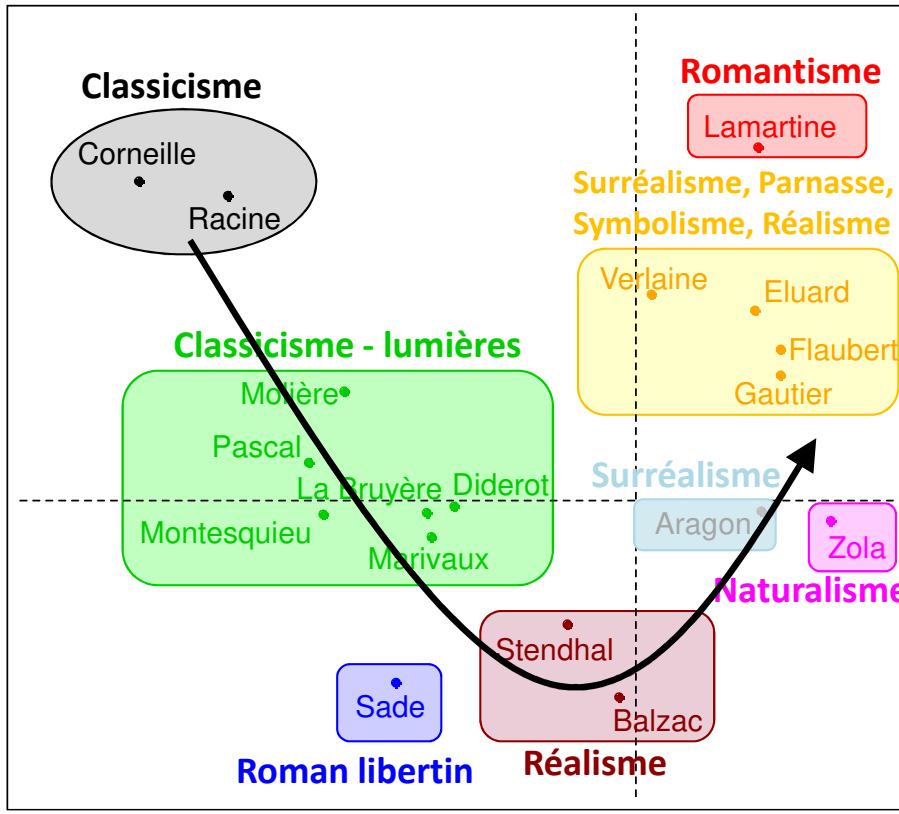




- Le naturalisme est la suite logique du réalisme : le naturalisme montre le milieu où vit le protagoniste pour expliquer son comportement de façon "scientifique"
- Évolution du vocabulaire selon les courants littéraires

Caractérisation par les mots





- Le naturalisme est la suite logique du réalisme : le naturalisme montre le milieu où vit le protagoniste pour expliquer son comportement de façon "scientifique"
- Évolution du vocabulaire selon les courants littéraires

Analyse de données textuelles

Autre approche : l'analyse sémantique latente

Coder la matrice documents-mots par tf-idf (term frequency and inverse document frequency), i.e. multiplier la fréquence d'un mot par l'opposé du logarithme de la proportion de documents contenant ce mot (commande `bind_tf_idf` dans `tidytext`).

Son nom est célébré par le bocage **qui** frémit, et par le ruisseau **qui** murmure, les vents l'emportent jusqu'à l'arc céleste, l'arc de grâce et de consolation que sa main tendit dans les nuages.	À peine distinguait-on deux buts à l'extrémité de la carrière : des chênes ombrageaient l'un, autour de l'autre des palmiers se dessinaient dans l'éclat du soir.	Ah ! le beau temps de mes travaux poétiques ! les beaux jours que j'ai passés près de toi ! Les premiers, inépuisables de joie, de paix et de liberté ; les derniers, empreints d'une mélancolie **qui** eut bien aussi ses charmes.

TF(t) = Nb d'apparition de *qui* / Nb total de termes (dans le document) = 2 / 38

$$\mathrm{idf}_1 = \log \frac{\text{nb documents}}{\text{nb docs contenant } \textit{qui}} = \log \frac{3}{2}$$ (« qui » absent du 2e document)

Ainsi $\mathrm{tfidf}_1 = \frac{2}{38} \cdot \log \frac{3}{2} \approx 0.0092$

Pour les autres documents : $\mathrm{tfidf}_2 = 0 \cdot \log \frac{3}{2} = 0$ et $\mathrm{tfidf}_3 = 1/40 \cdot \log \frac{3}{2} \approx 0.0044$

Le premier document apparaît ainsi comme « le plus pertinent ».

Projection dans un espace de dimension inférieure par une décomposition en valeurs singulières de la matrice tf-idf

$\Rightarrow$ Analyse sémantique latente