Stylométrie : analyse des caractères typographiques utilisés par un auteur

Olivier Dulcy

1. Introduction
2. Modalités d'action
   - Approche qualitative : Étude de faisabilité
   - Approche quantitative : le test d'indépendance du χ^2 de Pearson
   - Développement des outils en Python
3. Analyse des résultats
   - Peut-on différencier des auteurs via la ponctuation ?
   - Peut-on associer objectivement un texte à son auteur ?
4. Conclusion
5. Annexe

Histoire

- Affaire Roy Van Wyk en 2001
- *Pericles, Prince of Tyre*, Shakespeare, unique auteur ?

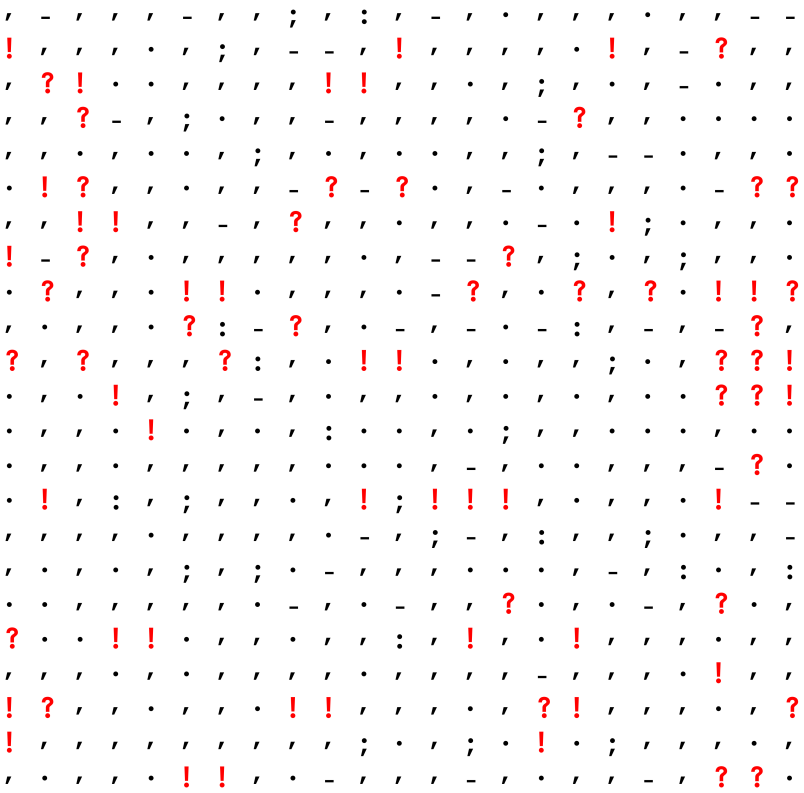
Sujet d'étude - Objectifs

- Choisir la base de données : ebooks libres de droit
- Détermination de la signature de l'auteur par la ponctuation
- Variabilité de cette signature en fonction du nombre de mots
- Création et automatisation de l'analyse des œuvres en Python
- Analyse à l'aide du χ^2

Visuellement : A.Dumas est différent de J.Verne

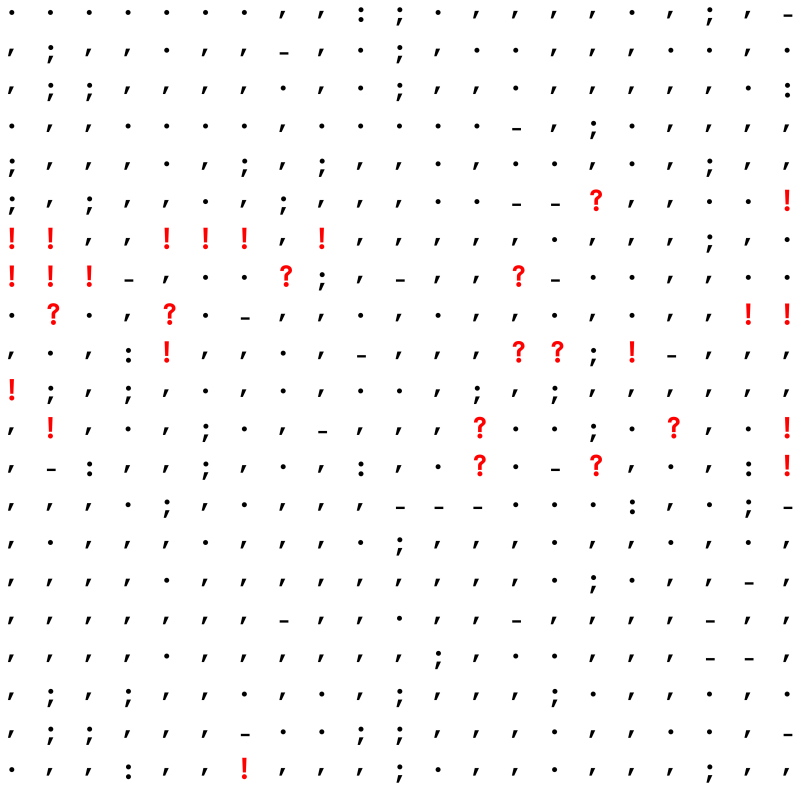
Alexandre Dumas

Le collier de la reine



Jules Verne

Cinq semaines en ballon

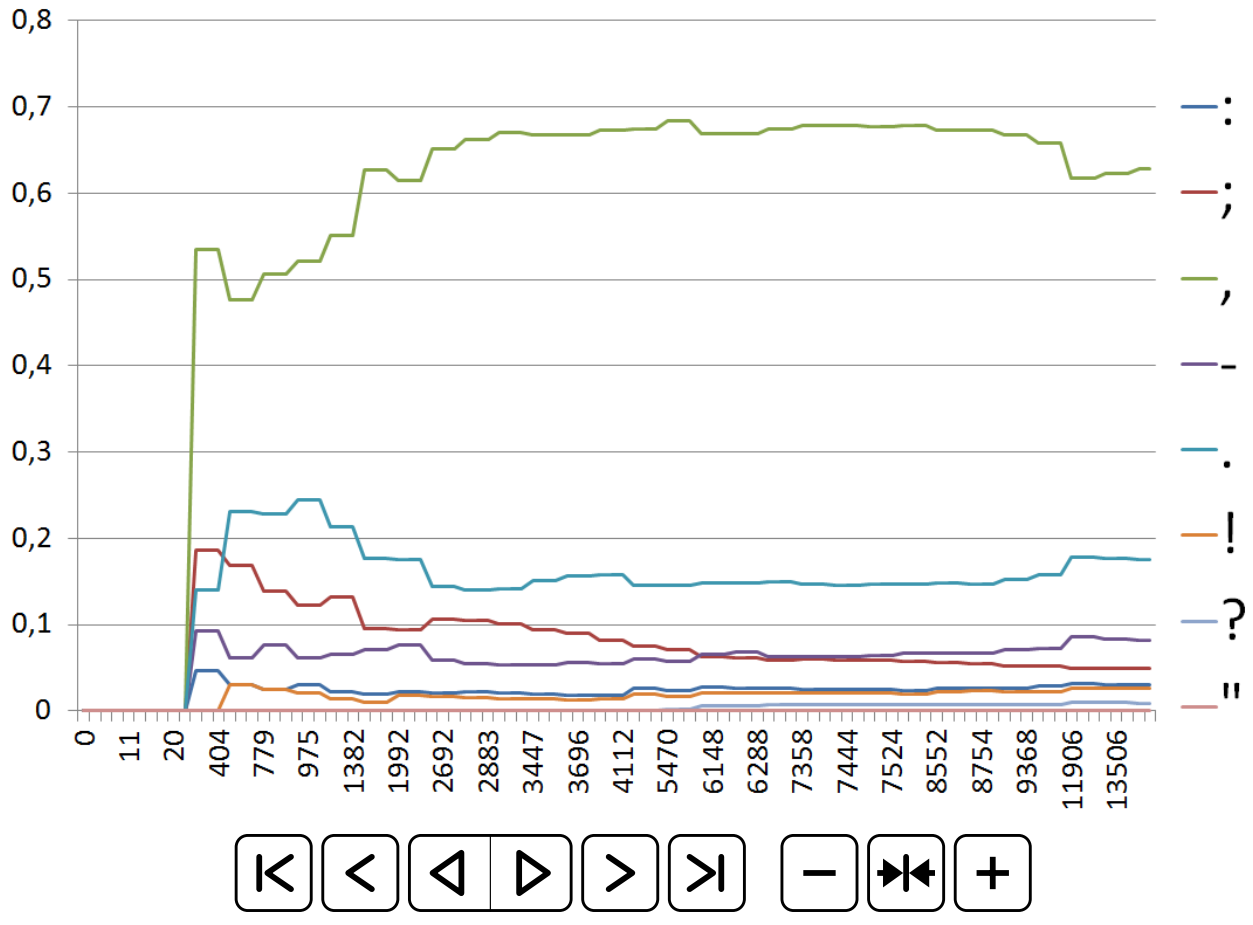


Blanc . , « » ; - : \ () []

Rouge ? !

Existence d'une signature : exemple de Proust

Part de la ponctuation (en %) en fonction du nombre de mots



Stabilité au delà d'un certain nombre de mots.

Introduction au test de Pearson

	:	;	,	-	.	!	?	Total
Alexandre Dumas	70	236	2372	377	999	255	232	4541
Gustave Flaubert	44	176	1618	172	768	66	44	2888
	114	412	3990	549	1767	321	276	7429

Question : Est-ce que la ponctuation est liée à son auteur ?

Test du χ^2 de Pearson

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

- $O_{i,j}$: observation de type i, j
- r : nombre de lignes, c : nombre de colonne (cf tableau ci-avant, étiquettes et total exclus)
- p_i : probabilité d'être de type i
- p_j : probabilité d'avoir le caractère j
- $E_{i,j} =$ Effectif total $\times\ p_i p_j$ la fréquence théorique de type i, j sous l'hypothèse d'indépendance
- Degré de liberté $= (r-1) \times (c-1)$

Loi du χ^2 à n degrés de liberté notée $\chi^2(n)$

$X_1, \ldots, X_n$ variables aléatoires indépendantes, de même loi normale, de moyenne respective μ_i et d'écart-type σ_i. On considère leurs variables centrées réduites $Y_i = \frac{X_i - \mu_i}{\sigma_i}$.

$$\chi^2 = \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

p-value : probabilité, pour un modèle statistique, d'obtenir un résultat supérieur ou égal sous l'hypothèse H_0 à celui fournit par les observations.

Démarche

$$\chi^2 + \text{ d.d.l } + \text{ Hypothèse nulle } H_0 = \text{p-value}$$

Exemple

Hypothèse zéro : Est ce que la ponctuation est indépendante de l'auteur ?
Hypothèse d'indépendance

	:	;	,	-	.	!	?	Total
Alexandre Dumas	0,94%	3,39%	32,83%	4,52%	14,54%	2,64%	2,27%	61,13%
Gustave Flaubert	0,60%	2,16%	20,88%	2,87%	9,25%	1,68%	1,44%	38,87%
	1,53%	5,55%	53,71%	7,39%	23,79%	4,32%	3,72%	100,00%

	:	;	,	-	.	!	?	Total
Alexandre Dumas	69,68	251,84	2438,90	335,58	1080,08	196,21	168,71	4541
Gustave Flaubert	44,32	160,16	1551,10	213,42	686,92	124,79	107,29	2888
	114	412	3990	549	1767	321	276	7429

Table – Application du test de Pearson : calcul des effectifs théoriques

- $\chi^2 = \frac{(70 - 69,68)^2}{69,68} + \ldots + \frac{(44 - 107,29)^2}{107,29} = 142,48$
- Degré de liberté : $(\text{Nbr Colonne} - 1) \times (\text{Nbr ligne} - 1) = 6$
- $p = 2,98 \times 10^{-28} \leqslant 0,05$

Conclusion

On peut rejeter l'hypothèse d'indépendance avec 5% de chances de se tromper, donc la ponctuation dépend de l'auteur

Pour chaque œuvre contenue dans le dossier :

Conversion eBook : epub au format txt

Extraction de la ponctuation des fichiers textes

- Fichier .punct

Recensement de la ponctuation par œuvre

- Fichier .sign

Mise en forme sous Excel

- Feuille par auteur
- Création des graphes
- Calculs des Khi-Deux

Comparaison Dumas \ Autre auteur

Hypothèse zéro : Est ce que la ponctuation est indépendante de l'auteur ?

d.d.l $= 6$	Flaubert	Jules Verne	Maupassant	Montesquieu	Marcel Proust	George Sand	Stendhal	Paul Verlaine	Victor Hugo	Émile Zola
χ^2	142	72	125	120	516	249	207	172	89	164
p	10^{-28}	10^{-13}	10^{-24}	10^{-23}	10^{-108}	10^{-51}	10^{-42}	10^{-34}	10^{-17}	10^{-33}

Le test de Pearson permet de différencier systématiquement les auteurs, en calculant la signature sur 4 œuvres complètes, au seuil de 5%

Tentative d'association d'un texte à son auteur

Entre 2 pages d'une même œuvre

Hypothèse zéro : Existe-t-il un lien entre 2 pages d'une même œuvre de Jules Verne ? (*Voyage au centre de la Terre*)

Auteur	:	;	,	-	.	!	?	Total
J.V Page 1	1	2	78	9	53	23	6	172
J.V Page 2	2	6	112	23	99	38	15	295
	3	8	190	32	152	61	21	467

$\chi^2 = 3,88$ avec 6 degrés de liberté $\Rightarrow p = 0,69 \geqslant 0,05$

On accepte l'hypothèse du lien entre les 2 pages au seuil de 5%

Tentative d'association d'un texte à son auteur

Entre 1 œuvre et la signature d'un auteur

Hypothèse zéro : Existe-t-il un lien entre une œuvre et la signature d'un même auteur ? (ici, Paul Verlaine)

Auteur	:	;	,	-	.	!	?	Total
Mes Hôpitaux	5	4	90	13	48	2	1	163
Signature	55	38	1488	175	323	119	51	2249
	60	42	1578	188	371	121	52	2412

$\chi^2 = 32,96$ avec 6 degrés de liberté $\Rightarrow p = 1 \times 10^{-5} \leqslant 0,05$

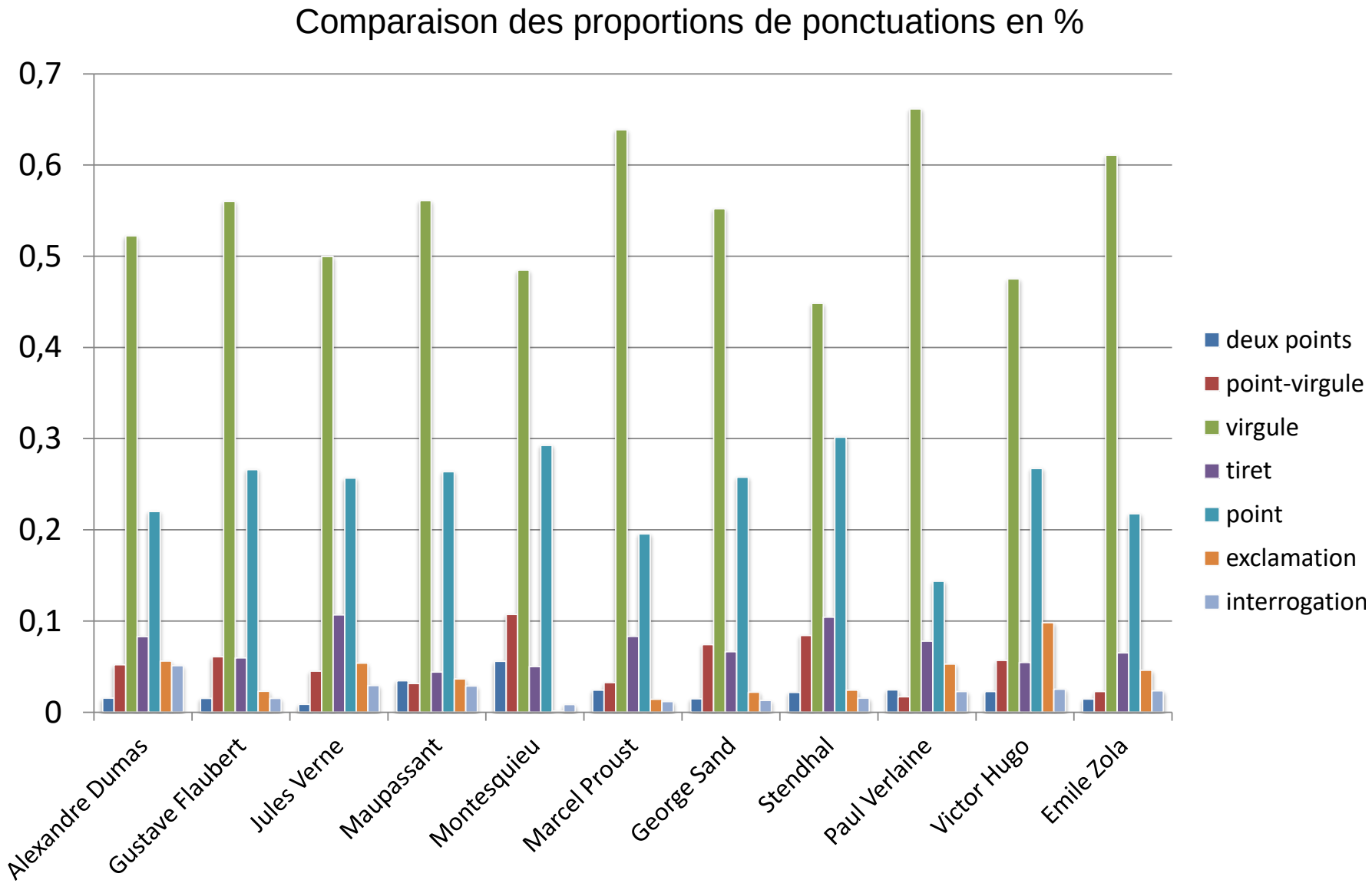
On rejette l'hypothèse d'un lien entre une œuvre et la signature d'un même auteur, au seuil de 5% (Échec d'association)

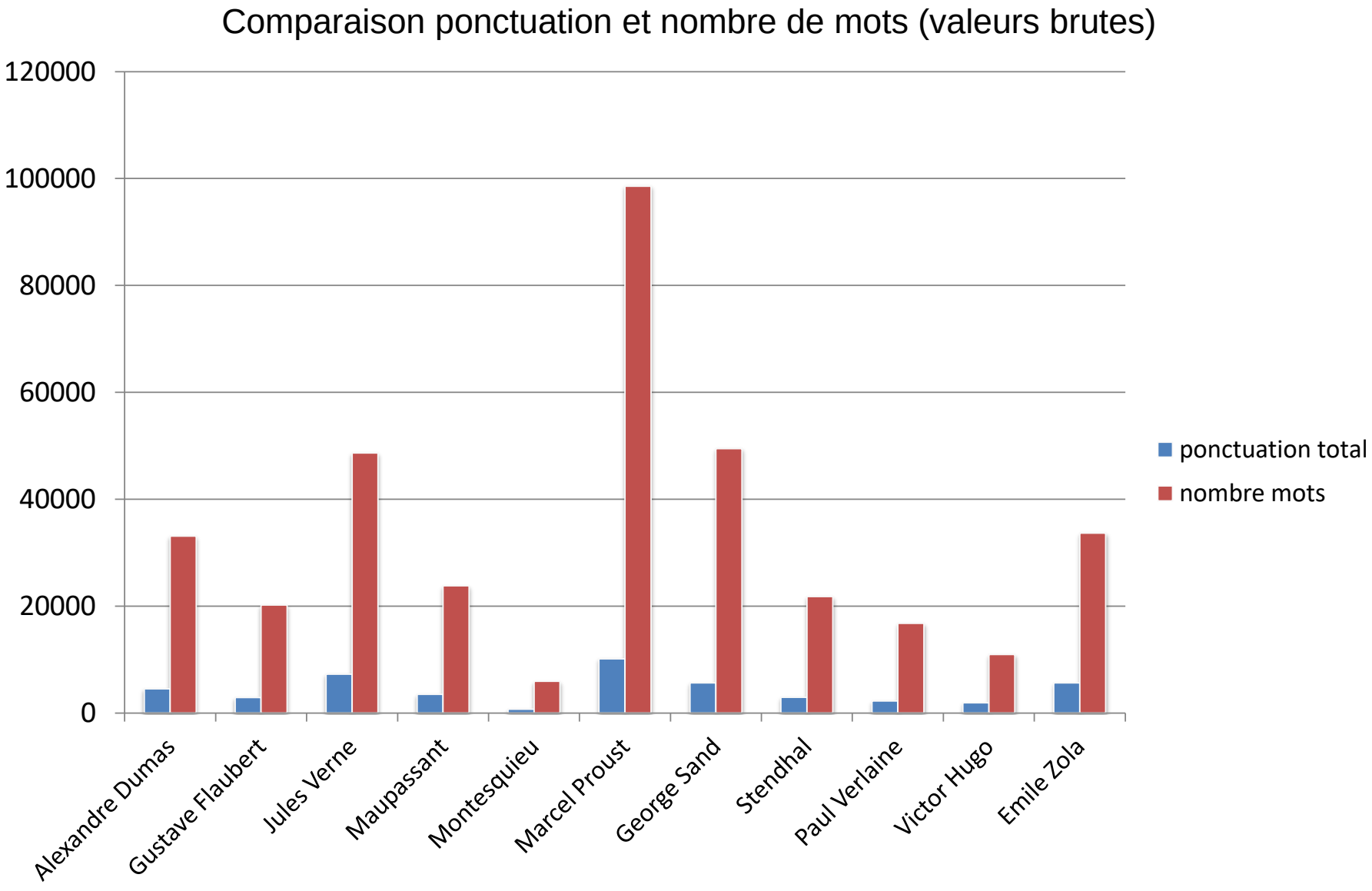
Différentiation des auteurs via la ponctuation : ✔

Corrélation entre 2 pages d'un même auteur : ✔

Association entre une oeuvre et la signature d'un auteur : ✘

- Discussion :
  - Contrainte sur le nombre de mots (exemple de Proust)
  - Temps d'exécution du programme : 13,5 s pour 42 fichiers (1000 à 40000 mots)


Comparaison des proportions de ponctuations en %
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Alexandre Dumas
Gustave Flaubert
Jules Verne
Maupassant
Montesquieu
Marcel Proust
George Sand
Stendhal
Paul Verlaine
Victor Hugo
Emile Zola
deux points
point-virgule
virgule
tiret
point
exclamation
interrogation


Comparaison ponctuation et nombre de mots (valeurs brutes)
120000
100000
80000
60000
40000
20000
0
Alexandre Dumas
Gustave Flaubert
Jules Verne
Maupassant
Montesquieu
Marcel Proust
George Sand
Stendhal
Paul Verlaine
Victor Hugo
Emile Zola
ponctuation total
nombre mots

Définition : Loi à densité

Loi à densité, de densité f, si :

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(t)\mathrm{d}t$$

Définition : Loi normale

Loi normale standard $\mathcal{N}(0, 1)$: densité $f_{0,1}(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

De manière générale, on dit que la variable aléatoire suit une loi normale $\mathcal{N}(m, \sigma^2)$ si elle a pour densité la fonction :

$$f_{m,\sigma}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-m)^2}{2\sigma^2}}$$