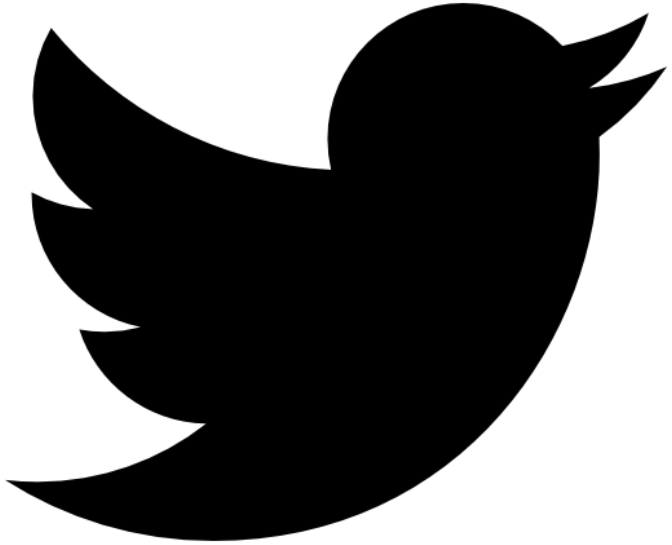
Deep learning et humanités : entre score et application

Thibault Clérice

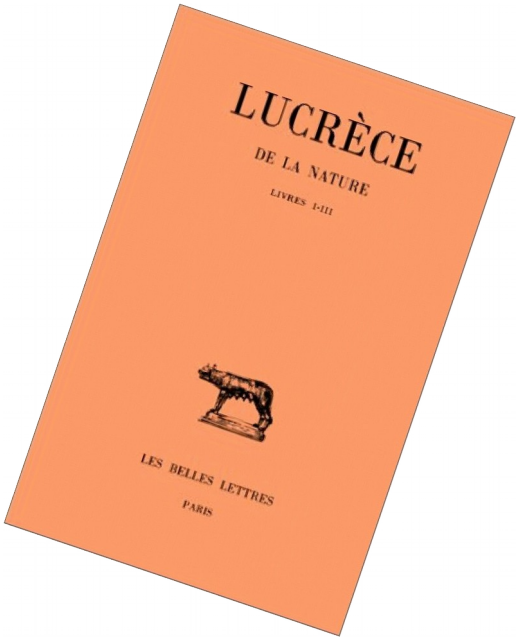
École nationale des Chartes, Centre Jean Mabillon, PSL
Université Lyon 3, Hisoma
Thibault.clerice@chartes.psl.eu
Github/Twitter : @ponteineptique

Introduction

L'apprentissage machine n'est pas pensé (à l'origine) pour les données historiques.



!=



Les modèles non-supervisés

« You shall know a word by the company it keeps »

J. R. Firth (1957)

Glove, Word2Vec, Fasttext

(201x)

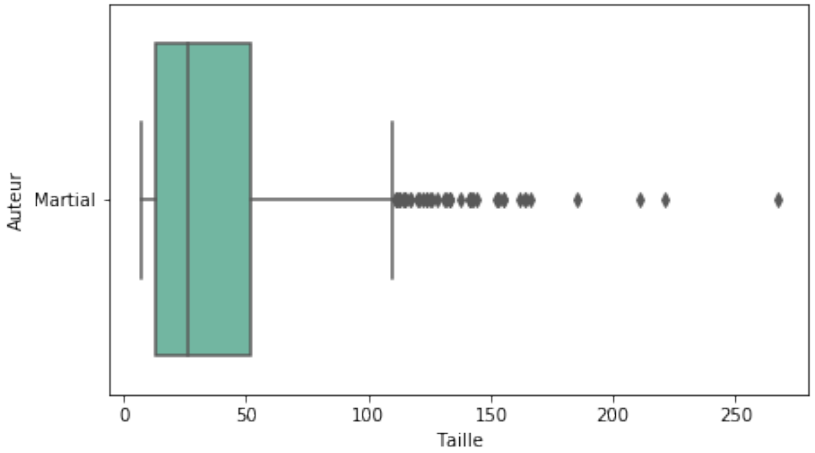
Word Embeddings : la notion de document

foulée du livre I. Au Moyen Âge, les *Épigrammes* ont été censurées; cependant, les moines copistes continuèrent à reproduire ces textes ce qui leur a permis d'atteindre la postérité.

Le *Liber spectaculorum* [modifier | modifier le code]

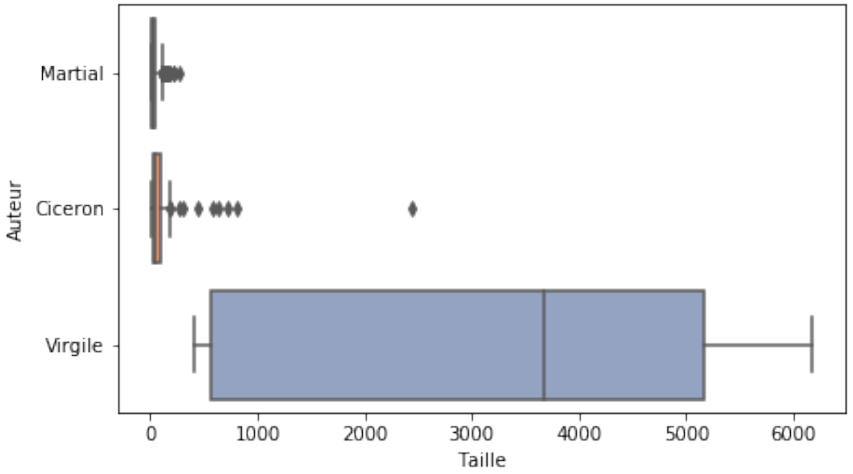
Placé en tête des *Épigrammes*, ce premier recueil de 33 pièces épigrammatiques, offert à Titus lors de l'inauguration du Colisée en 80 – et dénommé aujourd'hui *Liber spectaculorum* – n'est pourtant pas le

Word Embeddings : la notion de document

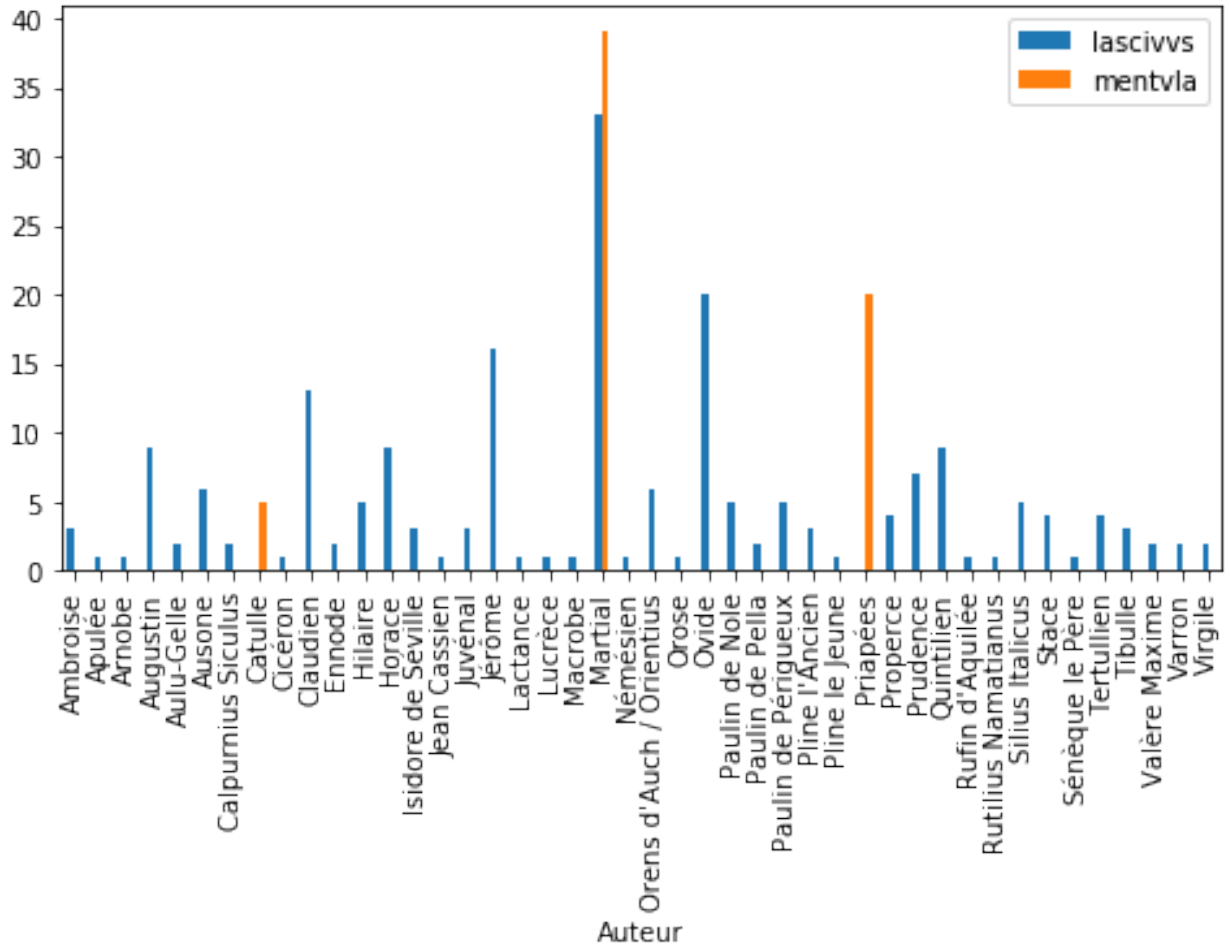


Taille des passages chez Martial

Taille chez Cicéron, Martial et Virgile



Word Embeddings : la notion de document



Répartition des occurrences de *lascivus* et *mentula*

Word Embeddings : la notion de document

T : taux de contamination

C : corpus de n textes

F : taille de la fenêtre

P : nombre de passages

M : nombre de mots

$$T_C = \frac{\sum_{n \in C} 2F(P_n - 1)}{M_C}$$

Word Embeddings : reproductibilité

Antoniak, Maria, et Mimno, David, « Evaluating the Stability of Embedding-based Word Similarities », *Transactions of the Association for Computational Linguistics*, t. 6, 23 février 2018, p.107-119.

Downstream-centered	Corpus-centered
Big corpus	Small corpus, difficult or impossible to expan
Source is not important	Source is the object of study
Only vectors are important	Specific, fine-grained comparisons are important
Embeddings are used in downstream tasks	Embeddings are used to learn about the mental model of word association for the authors of the corpus

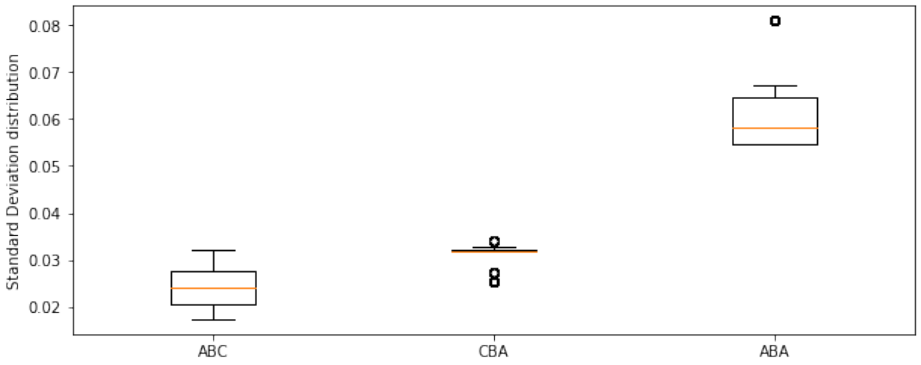
Word Embeddings : reproductibilité

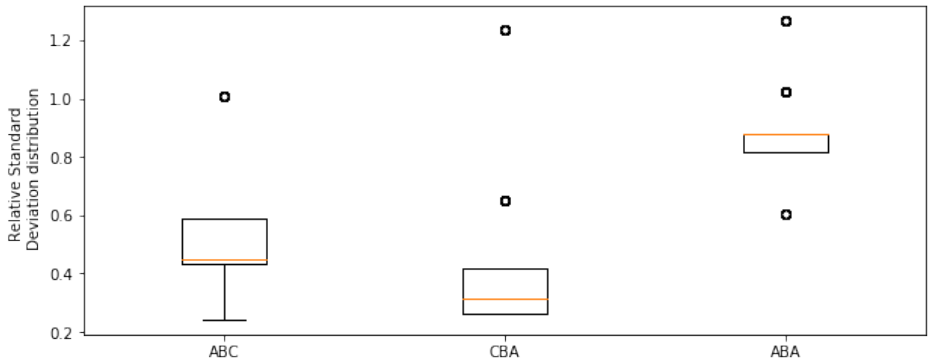
1.50 modèles pour chaque corpus et chaque algo

2.200 topic models sur l'un d'eux, sélection des 20 termes les plus importants

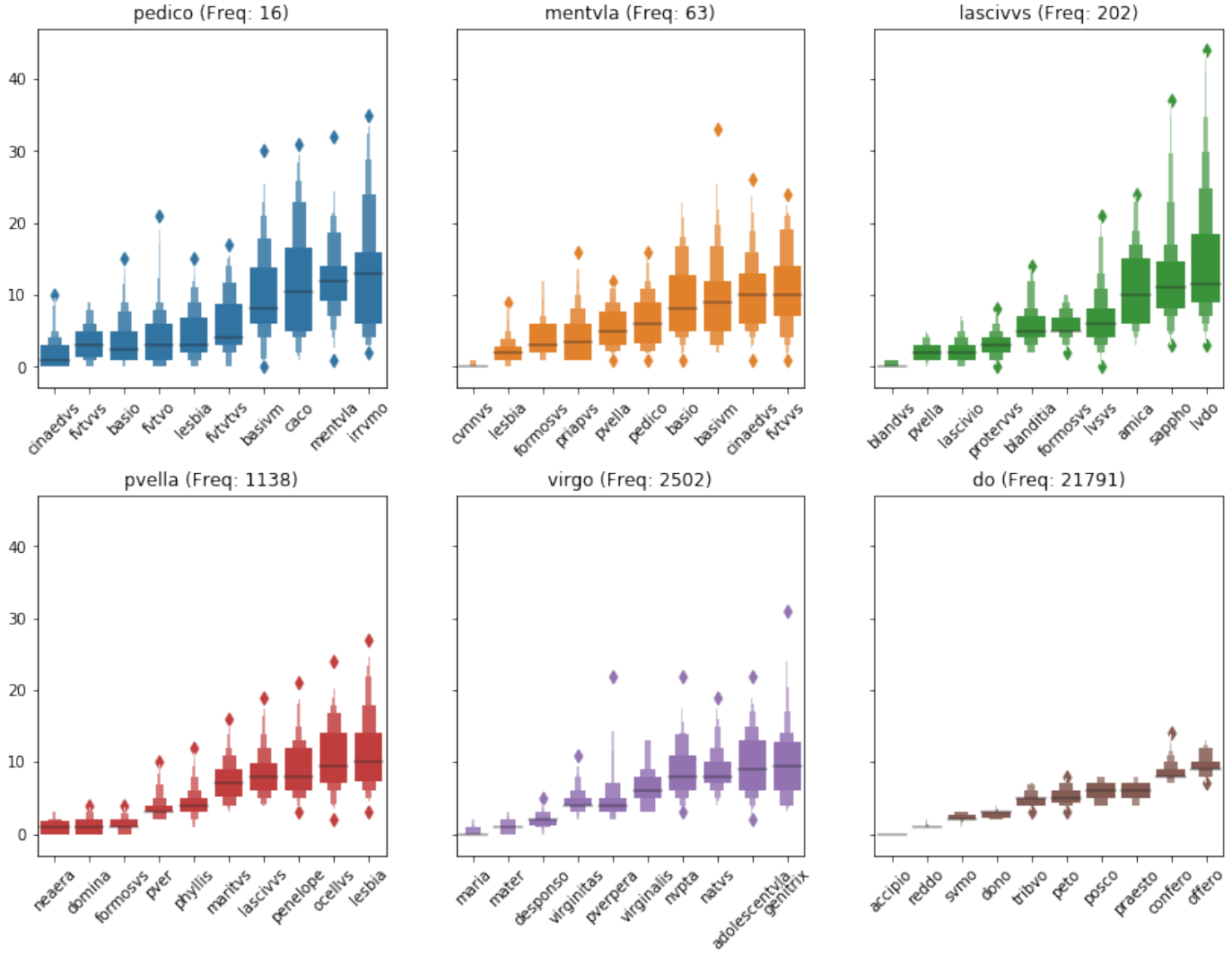
3. Similarité cosinus de chacune des paires de mots de chaque topic model

4.Écart type sur l'ensemble des mots





Word Embeddings : reproductibilité



Les modèles supervisés

« When a measure becomes a target, it ceases to be a good measure. »

Strathern, Marilyn, « 'Improving ratings': audit in the British University system », European Review, n° 3, t.5, juillet 1997, p.305-321.

Les modèles supervisés

Supervisé : Identifier la mesure

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$F\beta = (1+\beta^2 \frac{)\cdot precision \cdot recall}{(\beta \cdot precision)+recall}$$

Supervisé : Quantités et scores

Clérice, Thibault, « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin », https://hal.archives-ouvertes.fr/hal-02154122 (consulté le 24 novembre 2019).

- Lordinateurnesaitpaslirecamaisvoussi
- 10000000001010001001000101000101000101

Supervisés : quantités et score

Accuracy	Recall	Precision	F-Score
0.992	0.990	0.990	0.990

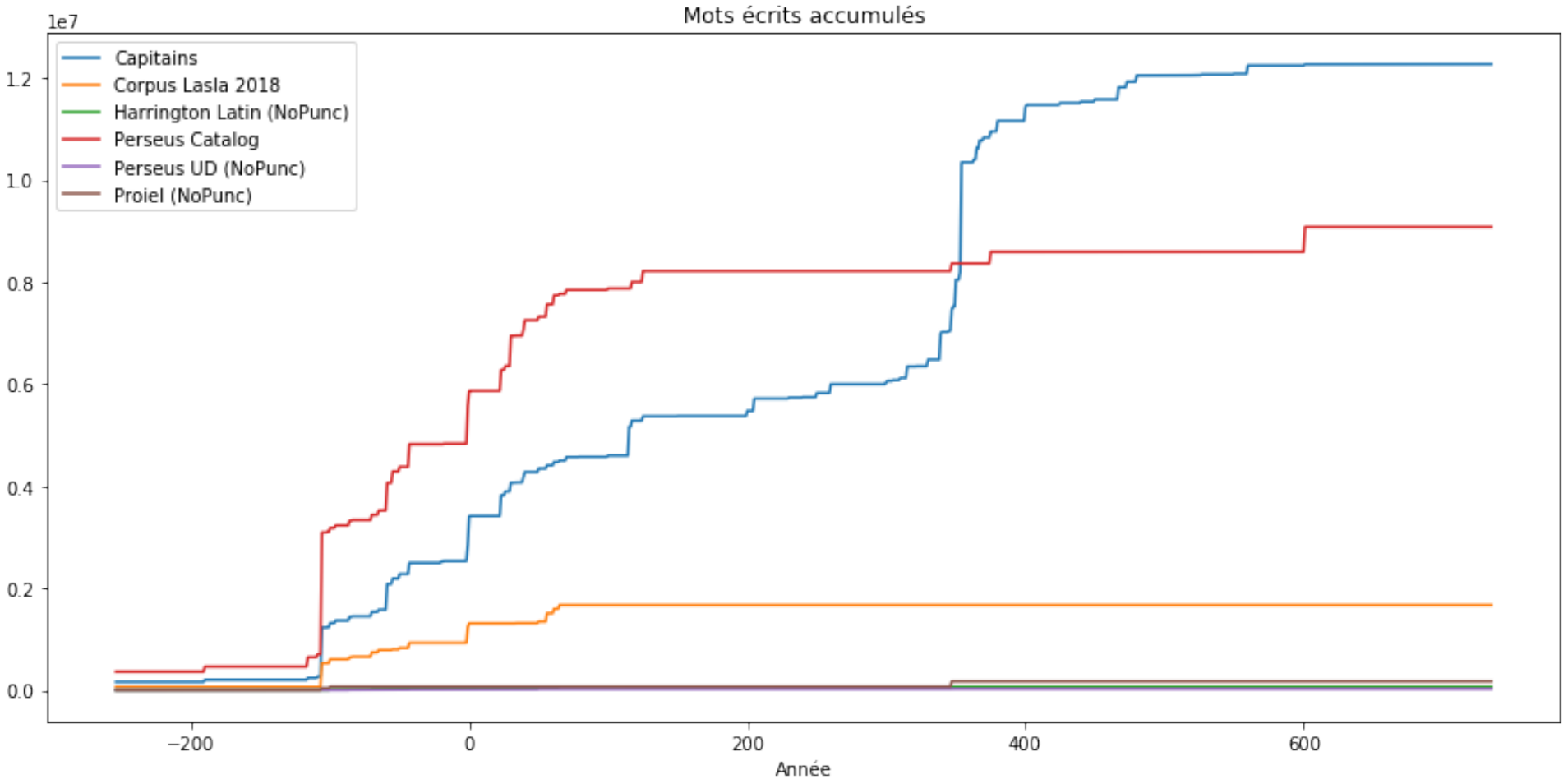
Accuracy	Recall	Precision	F-Score
0.833	0.50	0.42	0.45

Supervisé : représentativité et extensibilité

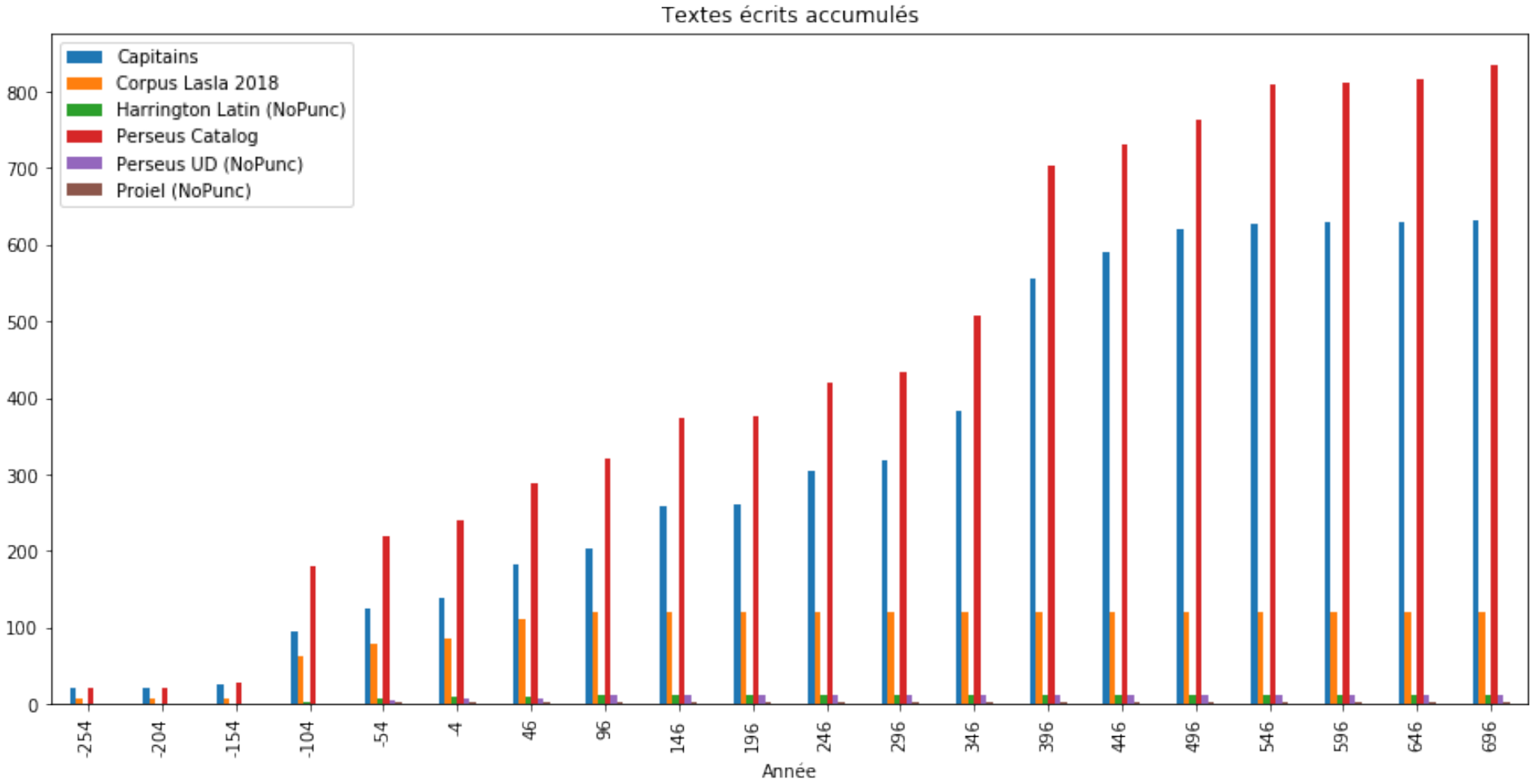
Gallia omnis diuisa in partes tres quarum unam incolunt Belgae aliam Aquitani tertiam qui ipsorum lingua Celtae nostra Galli appellantur

gallia_n omnis diuido in pars tres qui_1 unus incolo_2 belgae_n alius aquitani_n tertius qui_1 ipse lingua celtae_n noster galli_n appello_1

Supervisé : représentativité



Supervisé : représentativité



Supervisé : extensibilité

Model	Accuracy	Precision	Recall	FScore	WB FN	WB FP
Baseline	0.989	0.986	0.984	0.985	4031	3229
CNN	0.991	0.985	0.990	0.987	2137	3860
CNN L	0.991	0.979	0.990	0.985	**2117**	3750
CNN P	**0.993**	**0.990**	**0.991**	**0.990**	2432	**2114**
CNN N	0.991	0.987	0.988	0.988	2756	3312
CNN L N	0.992	0.988	0.989	0.988	2500	3567
LSTM	0.939	0.637	0.918	0.720	21174	18662
GRU	0.933	0.645	0.645	0.910	23706	19427

Table 3: Scores over the test dataset.
For models: N = normalized, L = Lower, P = no position embedding.
In headers, FN = False Negative, FP = False Positive

	Accuracy	Precision	Recall	FScore	WB FN	WB FP
Baseline	0.882	0.893	0.808	0.838	3658	644
CNN P	**0.957**	**0.948**	**0.944**	**0.945**	**854**	**723**

Table 4: Scores over the unknown dataset. FN = False Negative, FP = False Positive

Supervisé : Petit rappel

En lemmatisation, 95 % = 1 erreur tous les 20 mots

En OCR/HTR, 5 % en CER = 1 erreur tous les 20 caractères

Mécaniquement : WER > CER

Supervisé : classes mixtes et classes séparées

En analyse morphologique , on a « Neuve »:

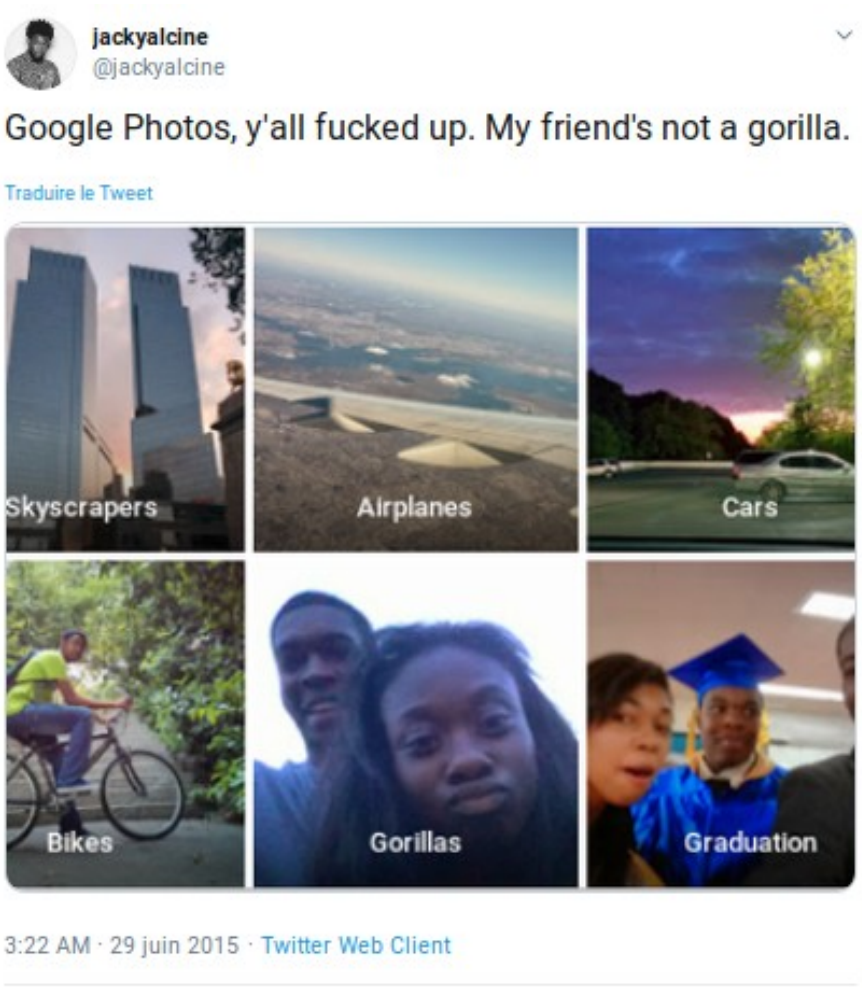
- Féminin (Genre)
- Singulier (Nombre)
- Positif (Degré)
- Ou… Fem|Sing|Pos

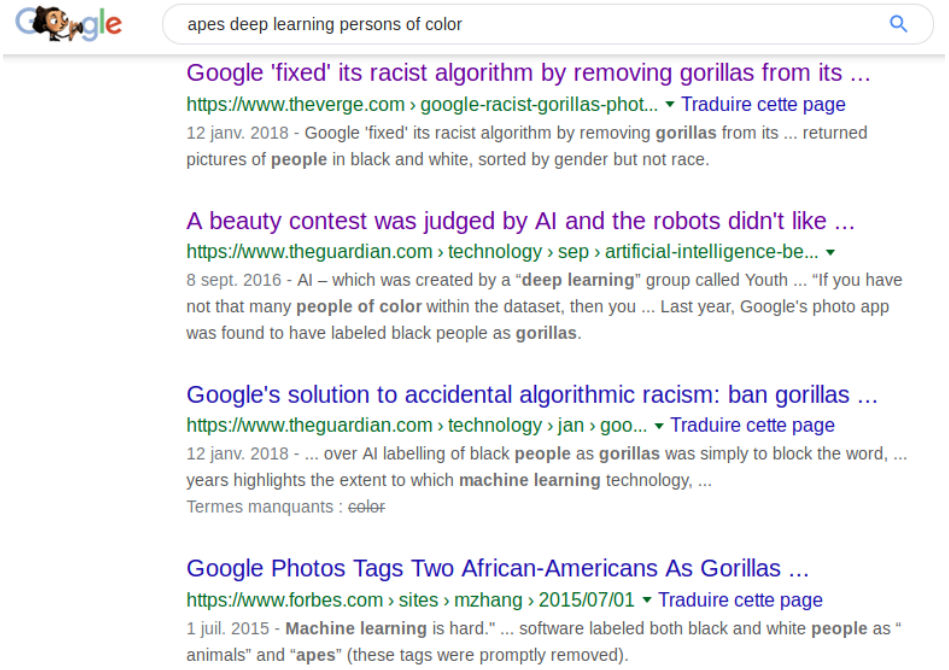
Supervisé : classes mixtes et classes séparées

Classe	Nombre
FémininSingulier	5000
FémininPluriel	7000
MasculinSingulier	5000
Féminin	**12000**
Singulier	**10000**

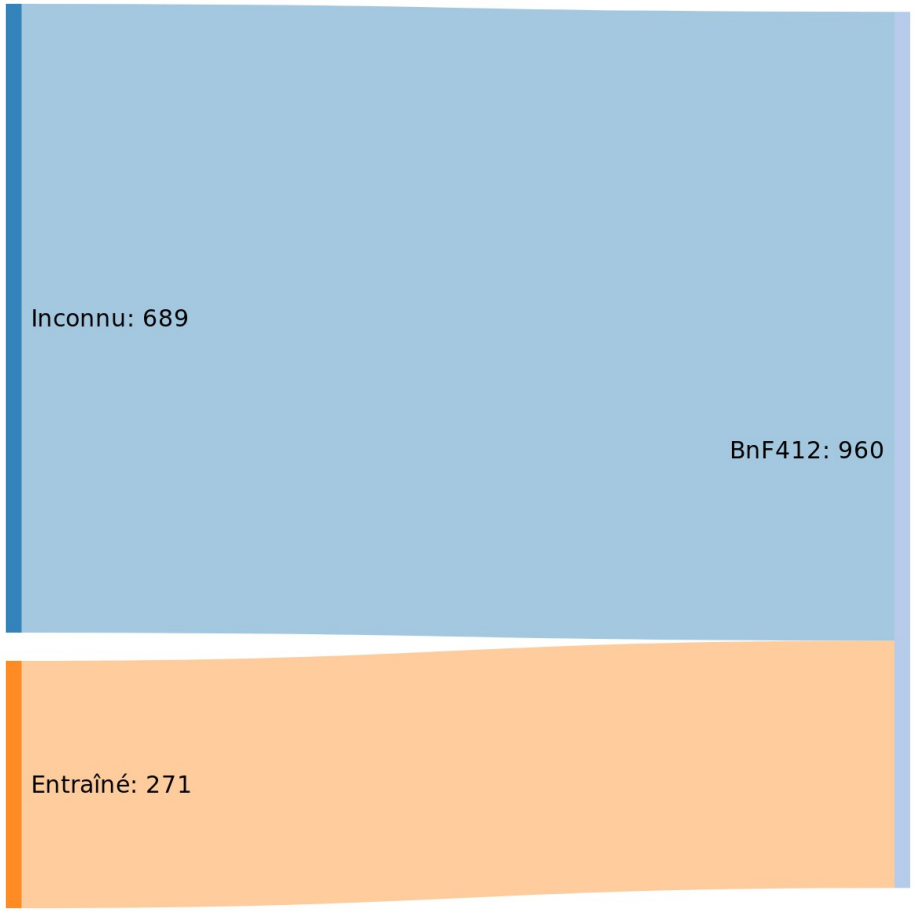
Application des modèles : les risques

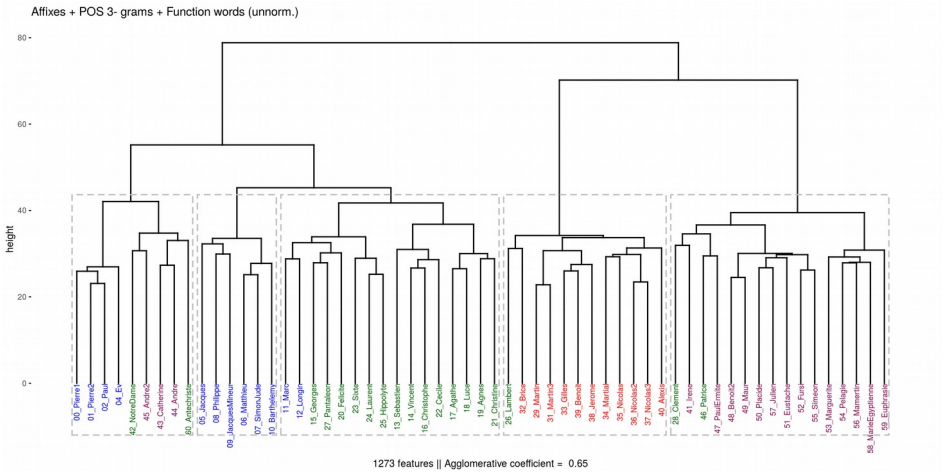
Risques : Prendre en compte le biais de corpus



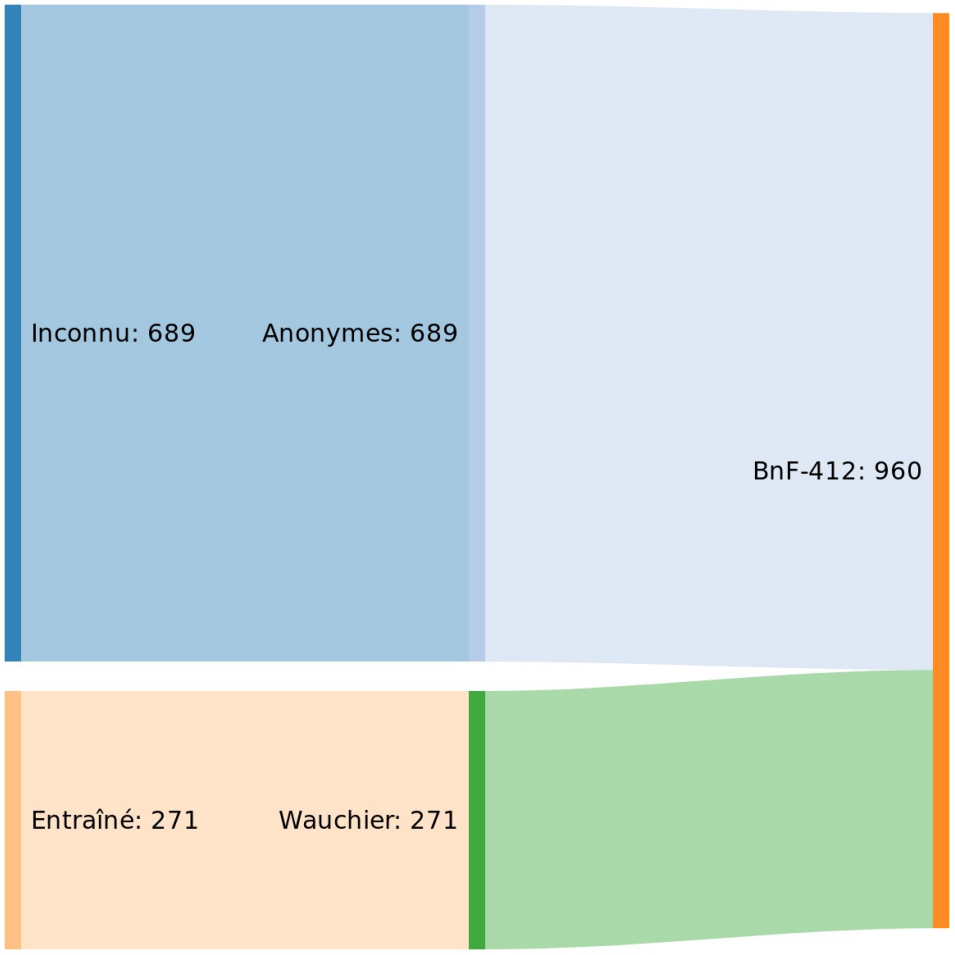


Risques : Prendre en compte le biais de corpus





Risques : Prendre en compte le biais de corpus



Risques : Analyser le bruit de pipeline

- Segmentation de lignes
- Acquisition du texte (3 % CER)
- Découpage des mots (0.99 Fscore)
- Résolution des abréviations (0.95%)
- Lemmatisation + POS Tagging (~ 96 % et 92 % accuracy)

Risques : Analyser le bruit de pipeline

Corpus	All	Function	Moisl	All	Moisl
Martin	33.35	10.69	11.43	**44.8**	**32.28**
Dialogues	29.38	9.77	**9.99**	48.38	35.85
Brice	39.49	12.14	**16.56**	**66.09**	47.51
Gilles	32.24	9.83	11.07	46.38	34.03
Martial	28.26	**7.92**	9.68	50.29	39.09
Nicolas	29.44	9.33	10.02	47.19	35.42
Jerome	**34.13**	**12.59**	14.38	61.92	**52.07**
Benoit	**27.97**	9.64	11.93	52.88	44.09
Alexis	30.19	10.65	11.58	57.71	47.77

$$\Delta_{A,B} = \frac{\sum_{i=1}^{n} |\mathrm{tf}_{¿} - \mathrm{tf}_{¿}|}{\sum_{i=1}^{n} \mathrm{tf}_{¿}}$$

Risque : Format des données

Gallia omnis **divisa** in partes tres , quarum unam incolunt Belgae **,** aliam Aquitani tertiam **,** qui ipsorum lingua Celtae nostra Galli appellantur **.**

gallia omnis **dico** in pars tres **qvi** qvi vnvs incolo belgae **vnde** alivs aqvitani tertivs **vi** qvi ipse lingva celtae noster galli appello **nonvs**

gallia omnis **divido** in pars tres , qvi vnvs incolo belgae **,** alivs aqvitani tertivs **,** qvi ipse lingva celtae noster galli appello **.**

Conclusion

Prendre en compte l'origine de l'algorithme ;

Faire attention à l'applicabilité et la reproductibilité des expériences : les résultats sont-ils stables ?

Revenir toujours au document, source des sciences humaines, en l'approchant de manière critique ;

Remettre en question les scores ;

Évaluer les quantités et la représentativité ;

Évaluer qualitativement, sans s'offusquer d'une erreur tous les 10 mots (et donc d'un très bon taux de 90%) ;

Conserver les données : les algorithmes changent rapidement.

Bibliographie

Maria Antoniak and David Mimno. 2018. Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119, February.

Thibault Clérice. 2019. Evaluating deep learning methods for word segmentation of scripta continua texts in old french and latin.

John Rupert Firth. 1957. *Papers in Linguistics, 1934-1951*. Oxford University Press.

Matthew Munson. 2017. Biblical Semantics: Applying Digital Methods for Semantic Information Extraction to Current Problems in New Testament Studies. *Shaker*.

Ariane Pinche, Jean-Baptiste Camps, and Thibault Clérice. 2019. Stylometry for noisy medieval data: Evaluating paul meyer's hagiographic hypothesis. In *Digital Humanities Conference 2019 – DH2019*. ADHO and Utrecht University.

Alexandra Schofield, Laure Thompson, and David Mimno, 2017. Quantifying the Effects of Text Duplication on Semantic Models, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marilyn Strathern. 1997. Improving ratings: audit in the british university system. *European Review*, 5(3):305–321.