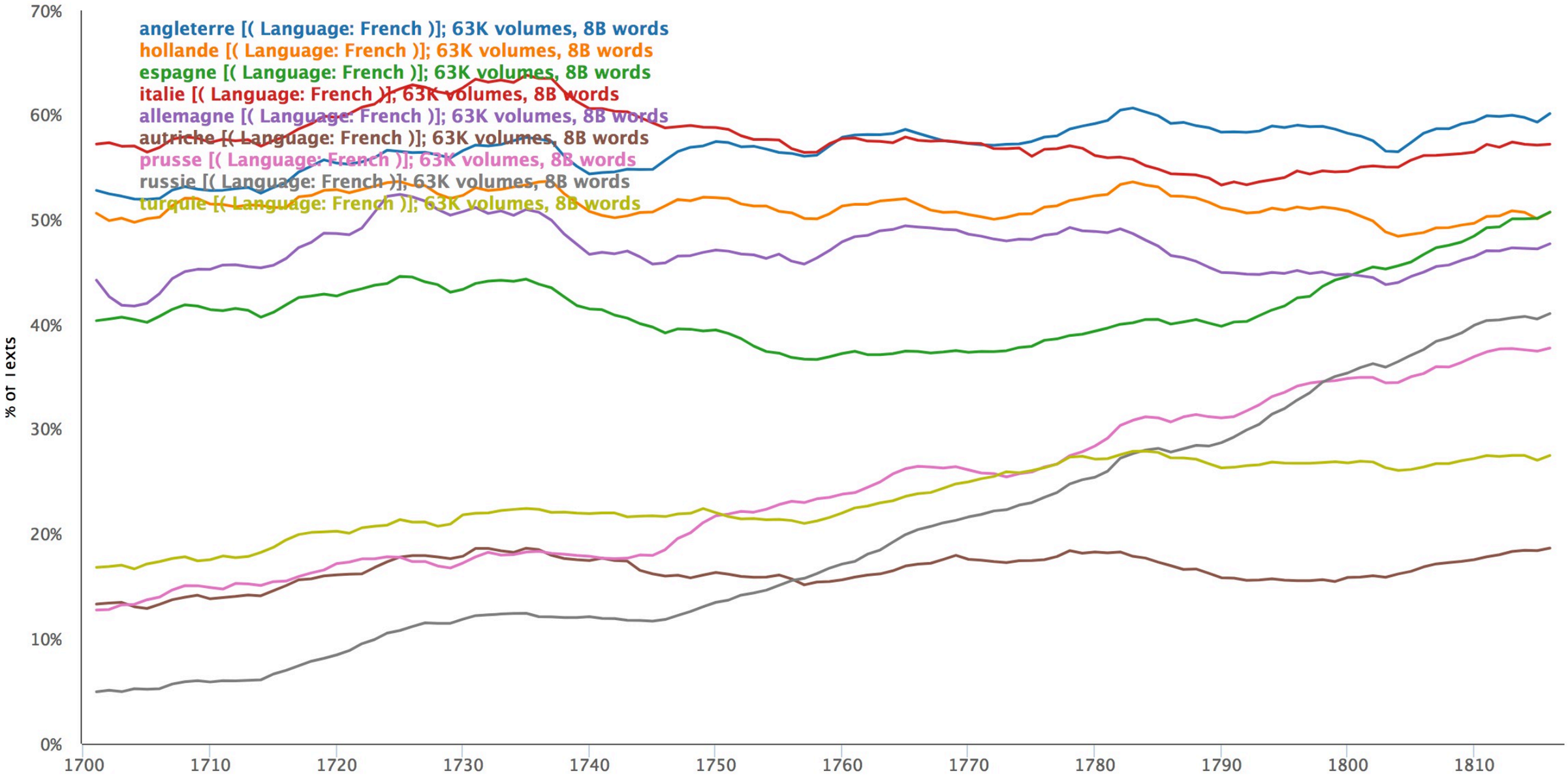
Chercher l’Amérique française dans deux siècles d’imprimé

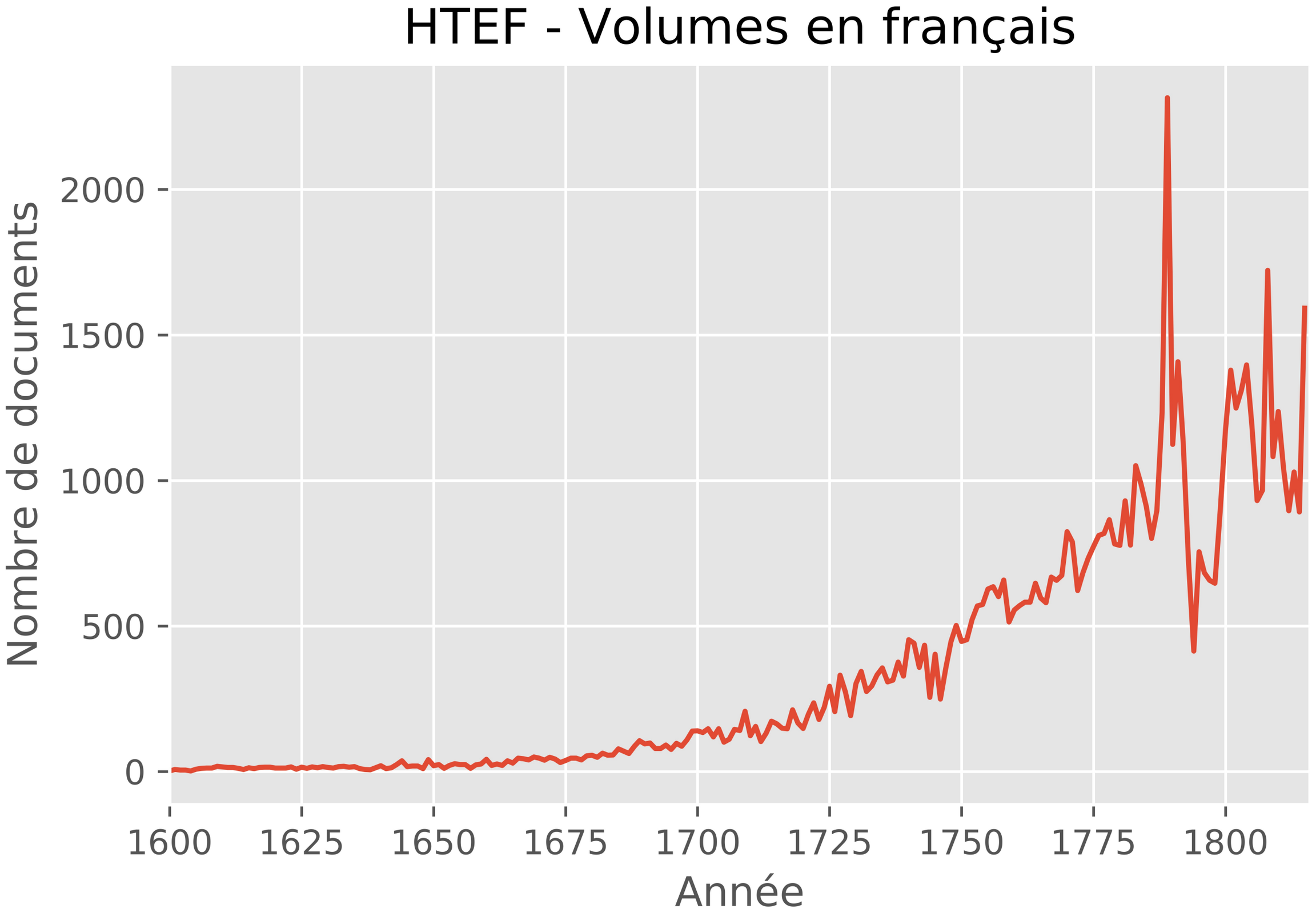
le jeu de données Hathi Trust Extracted Features

François Dominic Laramée
Chercheur postdoctoral, Université d’Ottawa
Institut d’histoire de l’Amérique française, 19 octobre 2019

Hathi Trust: Bookworm



Hathi Trust Extracted Features



Les «sacs de mots» du HTEF

Source: The Programming Historian

				count
page	section	token	pos	
27	body	those	DT	1
		within	IN	1
28	body	a	DT	3
		be	VB	1
		deserted	VBN	1
		faintly	RB	1
		important	JJ	1

HTEF: Qualité des données

- Données d'océrisation
- Enjeux des sources historiques:
  - Qualité de transcription
  - Orthographe irrégulière
  - Changements d'orthographe dans le temps
  - Changements au sens des mots dans le temps

Source: Michael Piotrowski (2012), *Natural Language Processing for Historical texts*

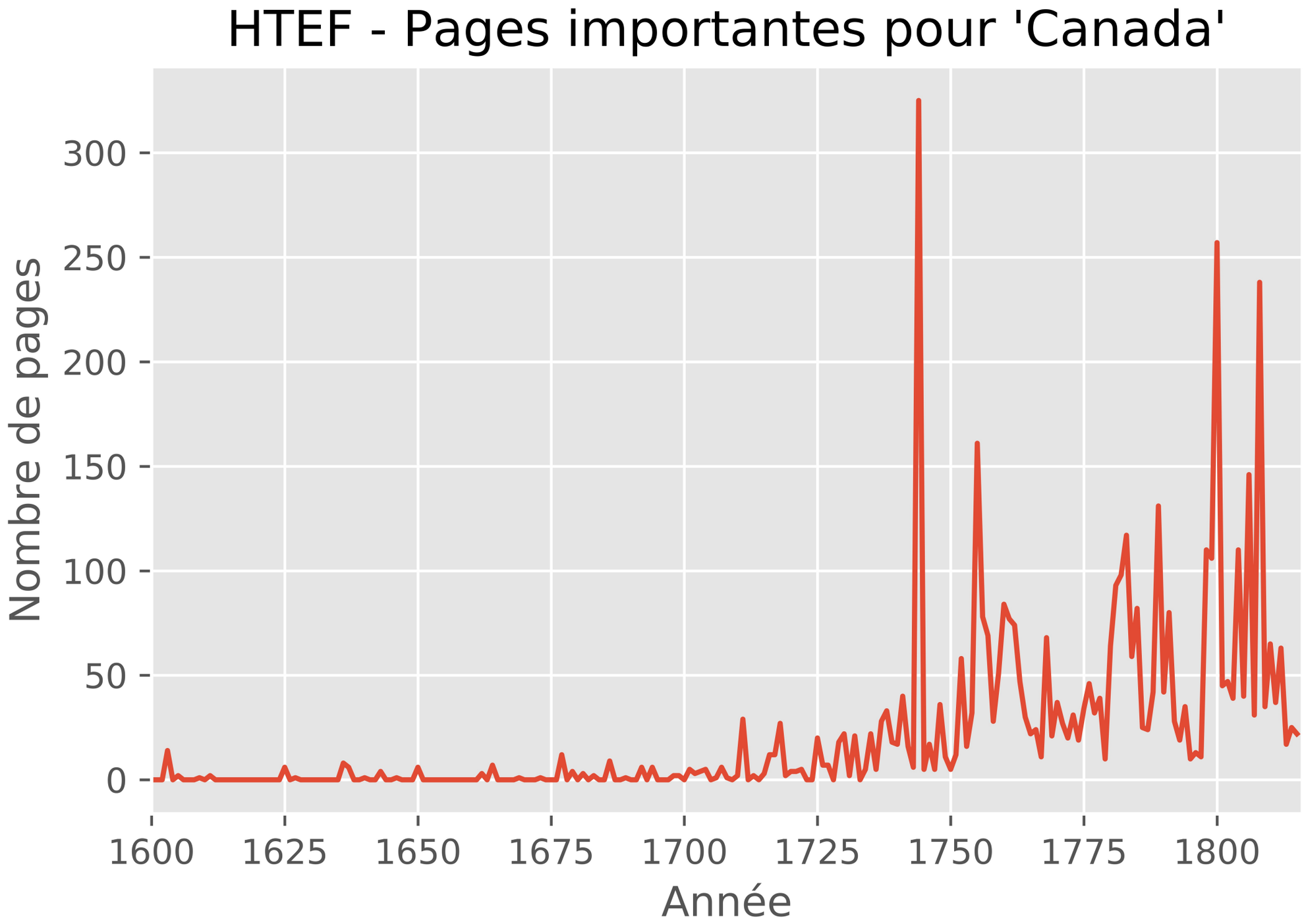
Algorithme de Levenshtein

Forme #1	Forme #2	Distance de Levenshtein	Opérations
Amérique	Amérique	0	—
Amérique	Amrique	1	é effacé
Amérique	Cmévrique	2	A —> C v inséré
Amérique	Musique	3	A effacé é —> u r —> s

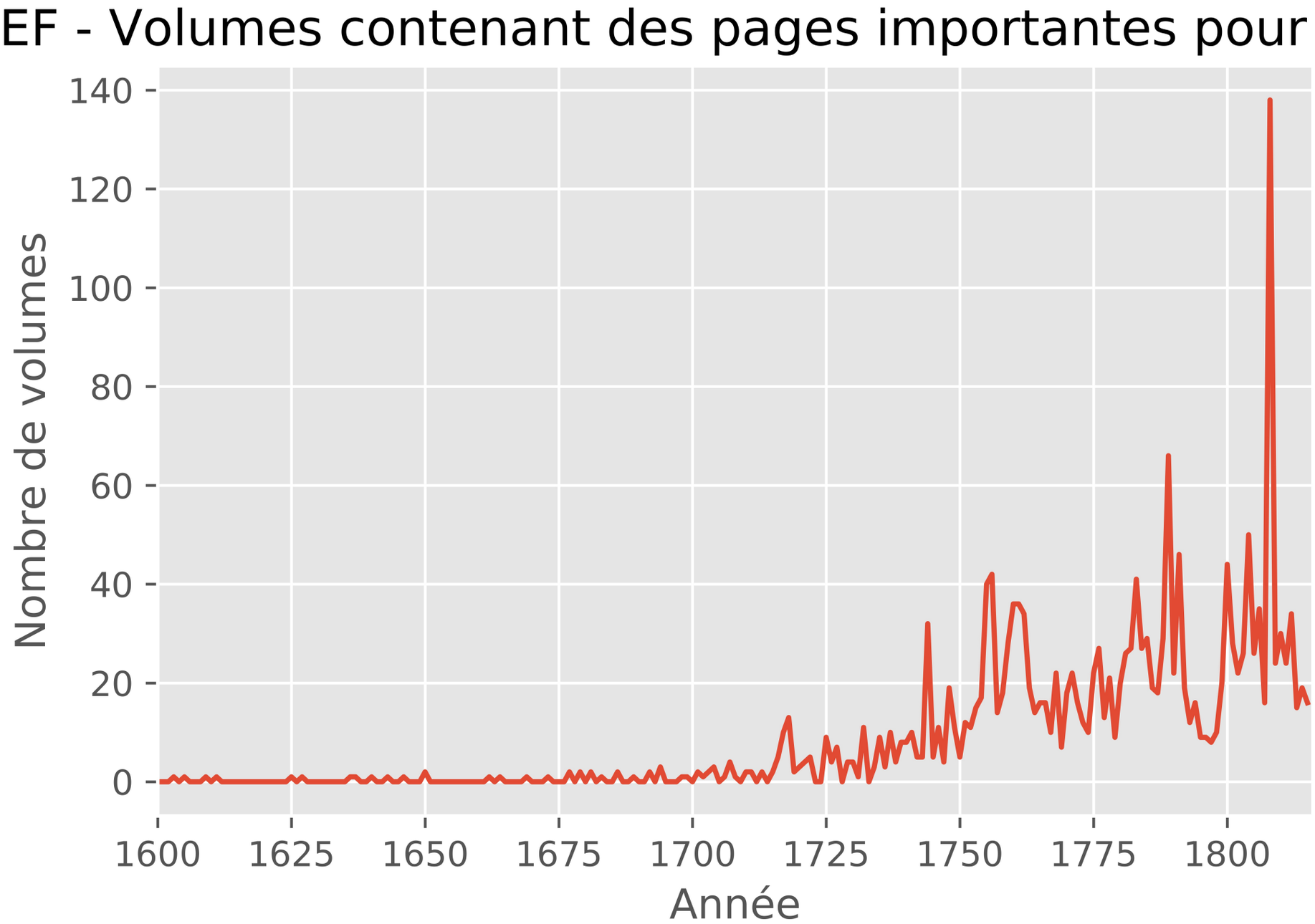
Design des questions de recherche

- Petit nombre de mots-clés
- Processus itératif
- «Preuve de concept» : un seul mot-clé CANADA
  - *Contenu pertinent: pages du HTEF contenant **au moins 3 occurrences***

Le Canada dans HTEF: Pages



Le Canada dans HTEF: Volumes



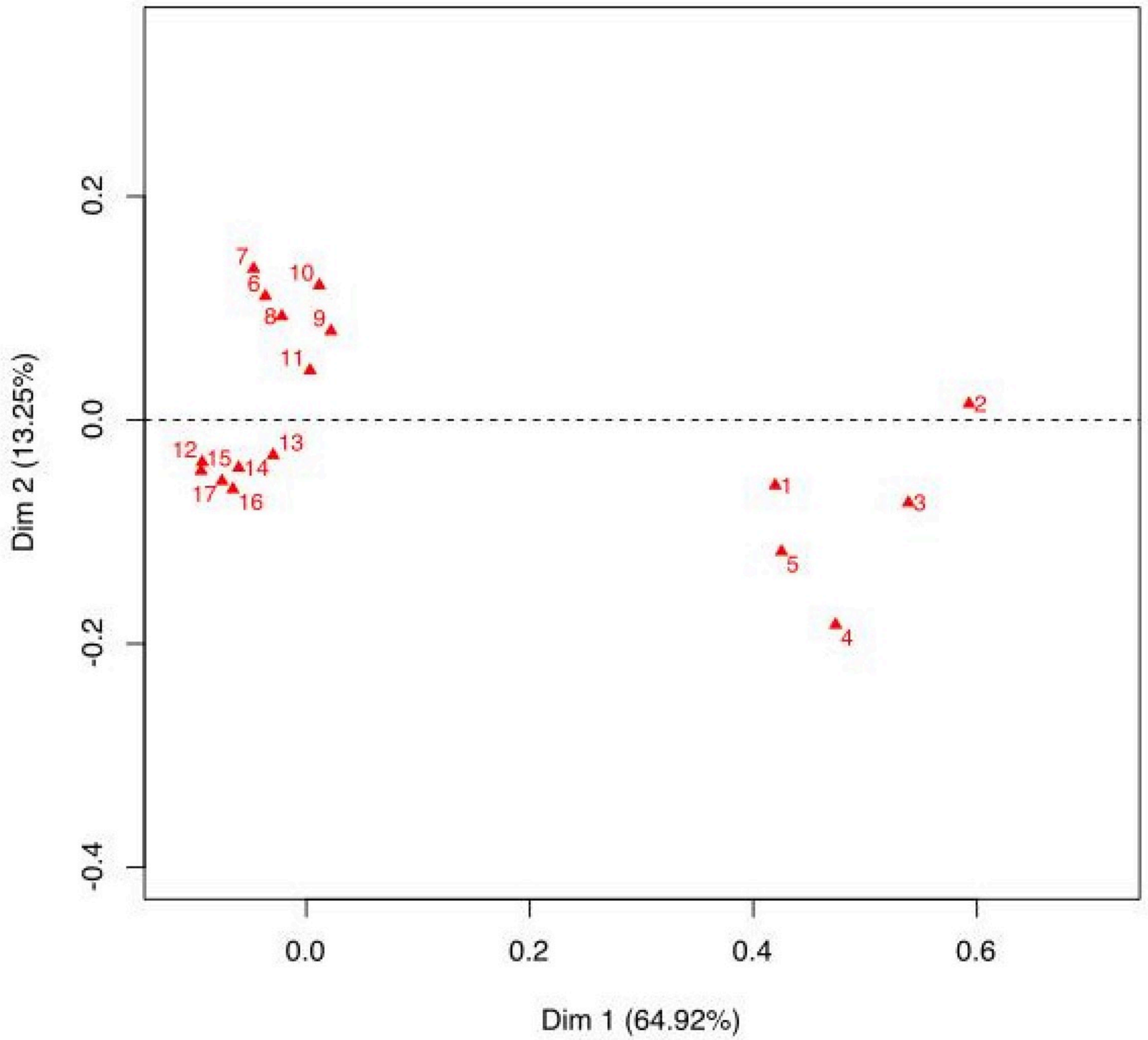
Méthodes: Décomptes simples

roi (121 533)	france (63 738)	ville (62 005)	prince (53 863)	guerre (48 501)
général (45 430)	général (45 430)	général (45 430)	général (45 430)	général (45 430)

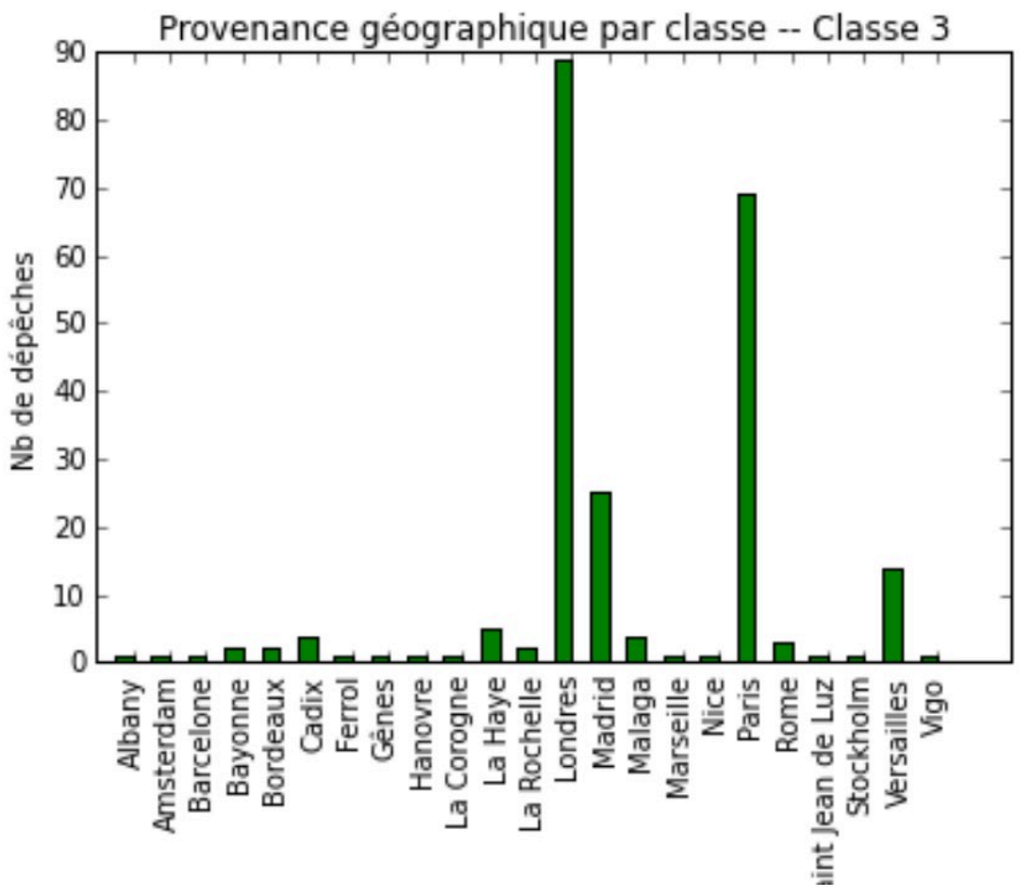
Méthodes: Modélisation thématique

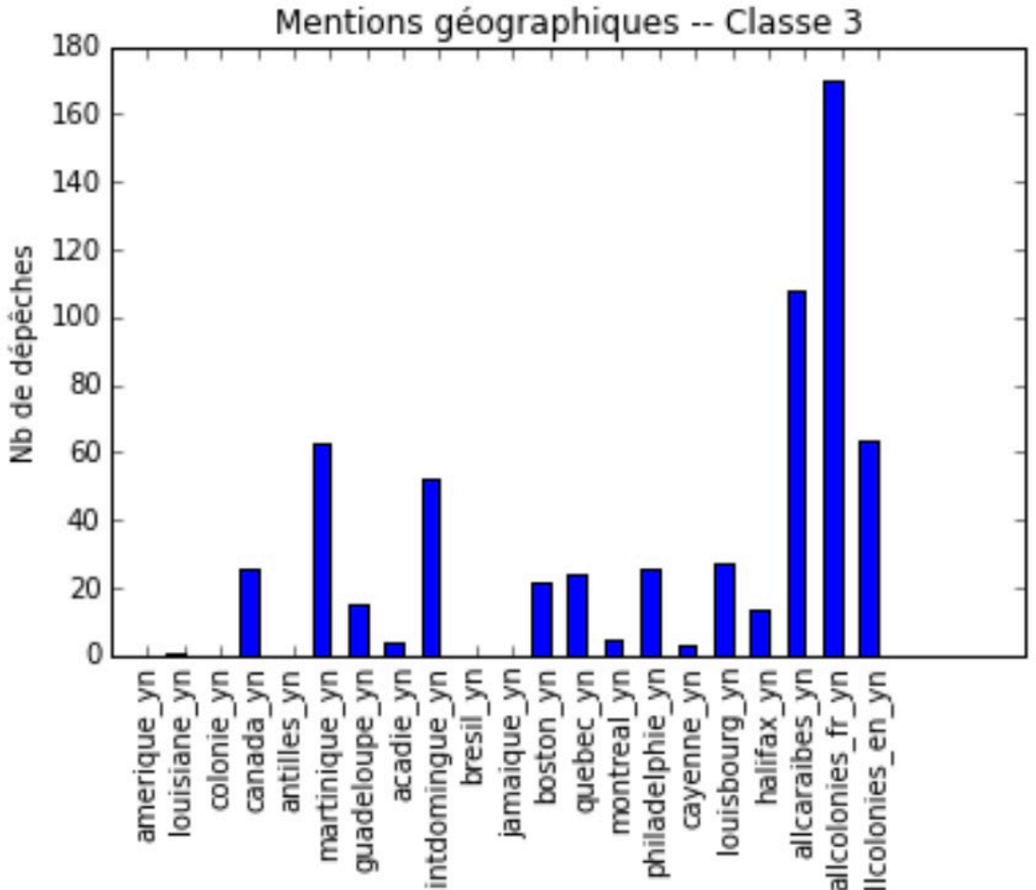
Mots les plus significatifs du thème	Majoritaire
roi troupes guerre pologne général prince point grand cour commerce	1755, 1757, 1769, 1771, 1773, 1775078, 1780, 1785, 1789-82, 1794-95, 1799
ville roi prince grand comte nom point duc troupes pologne	1740-49, 1752, 1756, 1758-60, 1762-63, 1766, 1768, 1770, 1772, 1775, 1786
général roi france ville prince français l’empereur guerre troupes grand	1805, 1809-10
général roi troupes prince guerre france ville l’empereur corps hommes	1806-07
général france roi prince guerre corps ville l’empereur point	1812, 1814
ville france général l’empereur roi prusse prince comte français corps	1813, 1815

Méthodes: Analyse factorielle

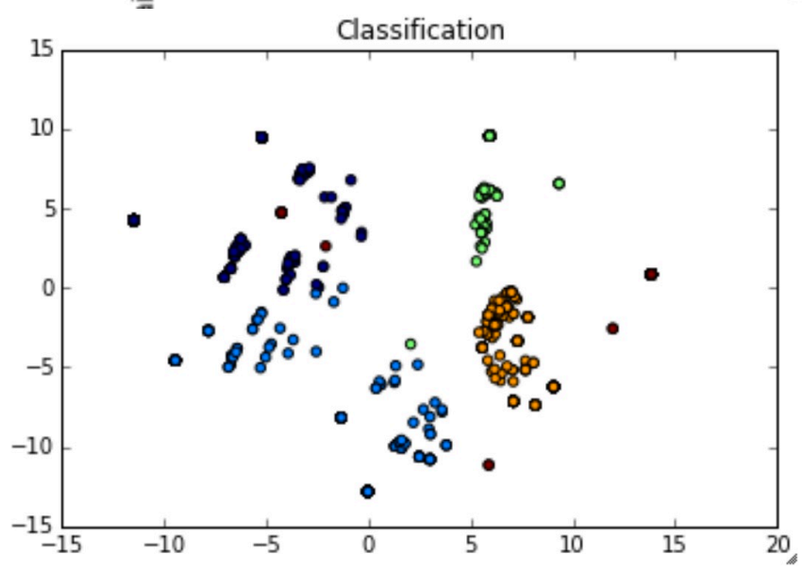


Méthodes: Classification (K-Moyennes)



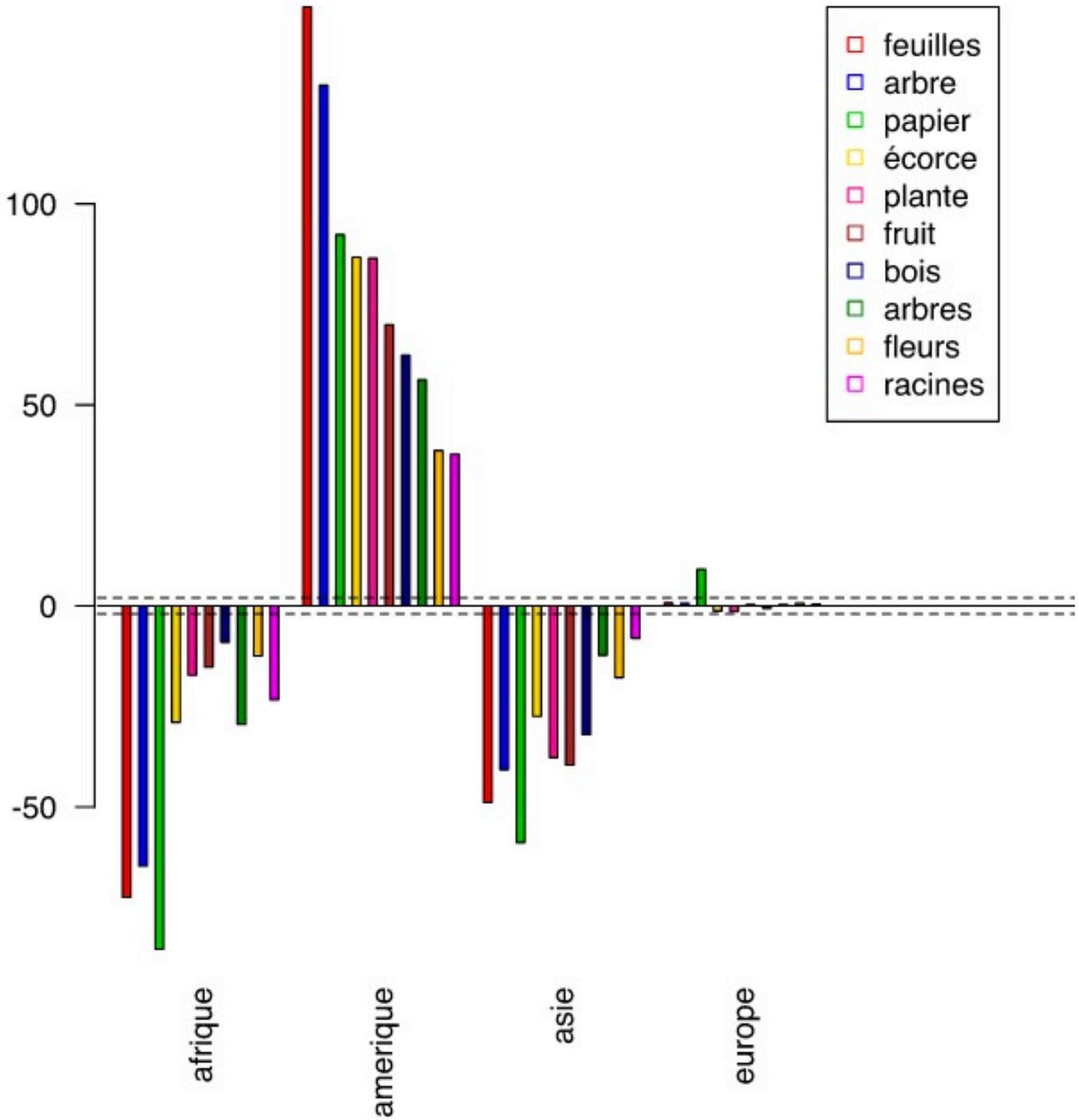


Lieu d'origine



Région

Méthodes: Spécificités lexicales



Méthode indisponible: Concordance

(requiert l'accès au texte complet)

euple de l'Amérique septentrionale, dans le Canada . Il occupe le 309. de long. & le 46
le, située sur les frontieres orientales du Canada , entre Terre-Neuve & la nouvelle An
lan, le loup marin, & autres que fournit le Canada . Voyez Canada. Quant à la pêche de
arin, & autres que fournit le Canada. Voyez Canada . Quant à la pêche de la morue, elle
NS, peuple de l'Amérique septentrionale, au Canada ; ils habitent entre la riviere d'On
a Californie, la Loüisiane, la Virginie, le Canada , Terre-neuve, les îles de Cuba, Sai
TOROQUE * ANDIATOROQUE, (Géog. mod.) lac du Canada ou nouvelle France dans l'Amérique s
vince de l'Amérique septentrionale, près du Canada & de la mer Septentrionale. lat. 41-
'arbrisseau qui nous vient de Virginie & du Canada ; il a la feuille du groseiller, & c
(Géog. mod.) baie de la nouvelle France au Canada propre, près des monts Notre-Dame, &

Perspectives

- «Marché des idées» mesurable dans HTEF
  - Canada: probablement
  - Amérique coloniale en général: oui
    - Caraïbes françaises, Louisiane, Acadie: à démontrer
- Locutions («Nouvelle France») dans un sac de mots?
- **Numérique *pour guider/compléter la lecture seulement***

Merci!

Courriel: flaramee@uottawa.ca
Web: www.francoisdominiclaramee.com
Twitter: @fdlaramee
ORCID: 0000-0001-5542-3754