



université
Lumière
LYON 2

UNIVERSITÉ JEAN MOULIN

LIFRANUM

eric

Peut-on représenter des auteurs dans un espace latent ?

Recherches de similarité dans la production textuelle
à l'aide de l'Intelligence Artificielle

Julien Velcin, laboratoire ERIC
Université Lumière Lyon 2

Colloque LIFRANUM, 24-25 octobre 2024

image généré par : <https://www.freepik.com>

Plan de la présentation

- La révolution du TAL
- Encoder le sens à l'aide de vecteurs
- Représenter les proximités entre auteurs
- Conclusion et pistes

2

La révolution du TAL

Julien Velcin, laboratoire ERIC

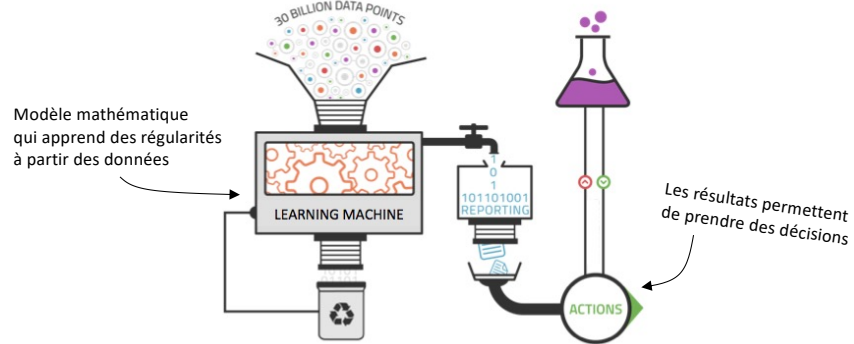
Colloque LIFRANUM, 24-25 octobre 2024

D'Enigma à ChatGPT

- Le Traitement Automatique des Langues (TAL) est un défi dès les premiers travaux en IA. Il permet de :
 - décomposer un texte en ses **constituants**
 - identifier (découvrir) le **sens** des mots et des expressions
 - découvrir des **motifs** (*patterns*) pour classer les textes ou générer du texte
- Quelques applications :
 - **chercher** de l'information dans les BD et le Web (moteurs de recherche)
 - **traduire** automatiquement des textes, les **comparer**
 - **classer** des textes en fonction de leur thématique, des opinions véhiculées...
 - **résumer** un document, **dialoguer** pour répondre à des questions...

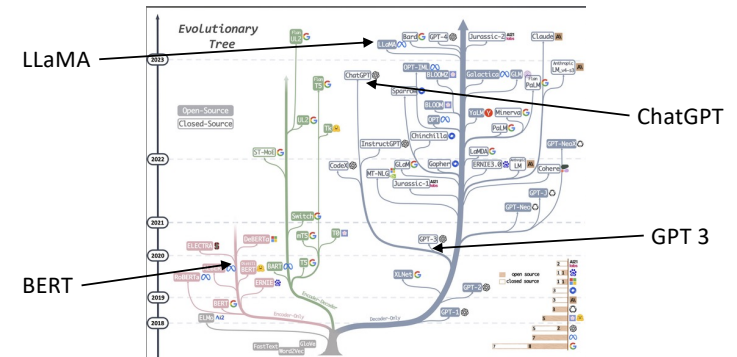
4

Des règles à l'apprentissage automatique



5

Succès des grands modèles de langue (LLMs)



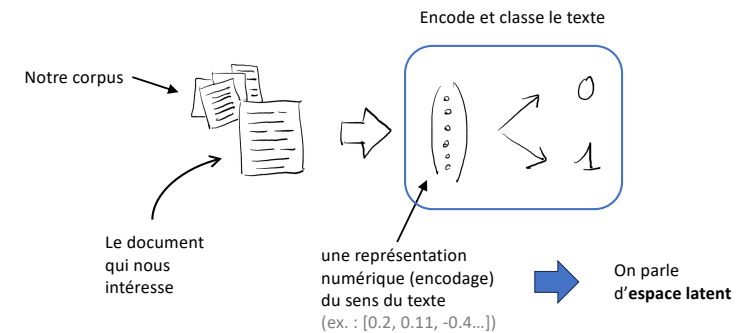
Note : ce graphique est en constante évolution

6

Encoder le sens des textes

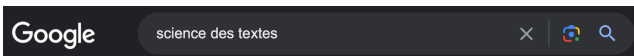
Julien Velcin, laboratoire ERIC
Colloque LIFRANUM, 24-25 octobre 2024

Encodeurs et représentations latentes



8

Recherche d'information

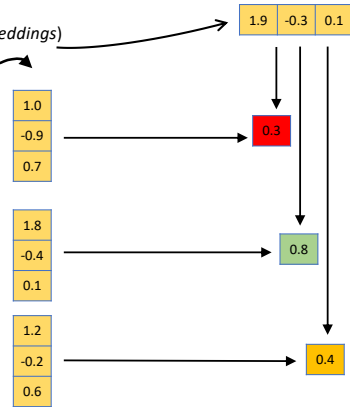


Gerflint
 https://gerflint.fr › Europe12 › olmo_cazevielle
 Des textes scientifiques et techniques aux scénarios ...
 by F Olmo-Cazevielle · Cited by 2 — scientifique et technique française. Chaque texte, écrit ou oral, devrait à travers les activités proposées déclencher des formes particulières de... 18 pages

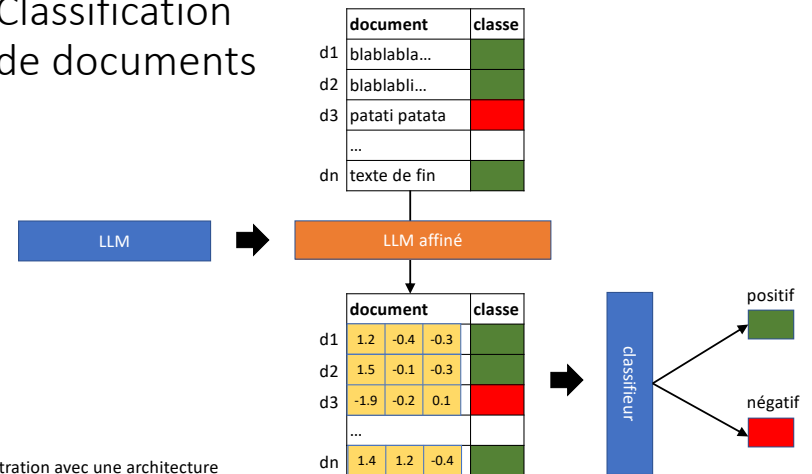
Quora
 https://fr.quora.com › Quelle-sont-les-sciences-qui-sintér...
 Quelle sont les sciences qui s'intéressent à l'étude du texte
 La linguistique est une science qui étudie le langage, comment il fonctionne, comment il est acquis et comment il est utilisé pour communiquer. C'est la ...
 5 answers · Top answer: Tout dépend sous quel angle on l'étudie: philologie, linguistique, st...

wikiHow
 https://fr.wikihow.com › écrire-une-...
 Comment écrire une analyse littéraire au Cégep
 L'analyse du texte littéraire : Un récit ou une pièce de théâtre : « Qui sont les personnages ?
 · Un poème : « Quelles sont les rimes et comment sont-elles ...
 L'analyse du texte littéraire · Planifier la rédaction · La rédaction

vecteurs (embeddings)



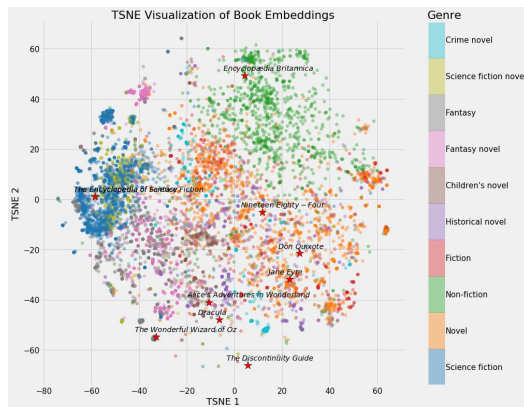
Classification de documents



Ici, illustration avec une architecture encodeur / classifieur (ex. : BERT)

Plongement de textes et représentations

- La plupart des applications nécessitent de « plonger » (embed) les textes dans des espaces vectoriels. Ces vecteurs sont des **représentations** qui visent à capturer la sémantique des textes.



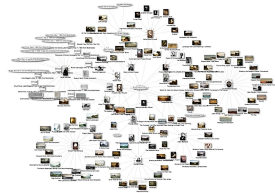
<https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526>

Représenter les proximités entre auteurs

Julien Velcin, laboratoire ERIC
 Colloque LIFRANUM, 24-25 octobre 2024

Représenter les proximités entre auteurs

- Deux exemples illustratifs :
 - [An ocean of books](#)
 - [Hudson River School artists](#) (explorer le [graphe sémantique](#))



- Ces méthodes emploient généralement la structure des données (par ex. les liens entre les pages Wikipedia)
- Comment faire en se basant sur le *contenu* textuel ?

13

Mesurer le style littéraire (Terreau et al., 2021)

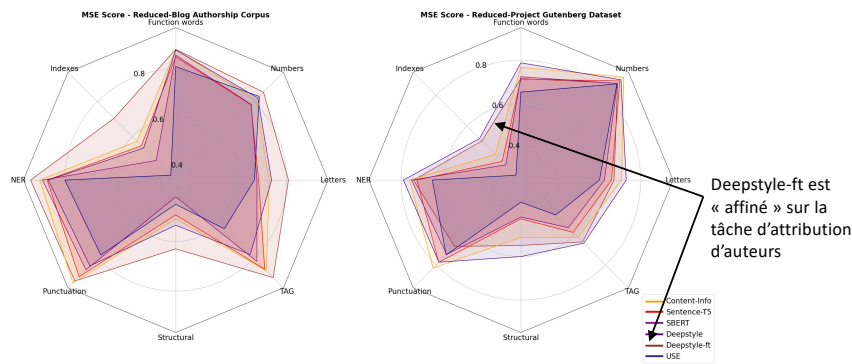
- Suivant la littérature sur le sujet, nous nous basons sur **303 descripteurs stylistiques** :

Catégories	Exemples	Nombre de marqueurs
Lettres	Fréquences de lettre	26
Nombre	Fréquences de nombre	11
Structuel	Longueur moyenne des mots, Hapax Legomena, ...	9
Ponctuation	Fréquences des signes de ponctuation	36
Mots outils	Fréquences des mots outils (does, once, doing, ...)	153
Tag	Fréquences des POS-tag	43
Ner	Fréquences des entités nommées	18
Index	Index de lisibilité et de complexité	7

- On va évaluer à quel point les représentations apprises par les modèles *capturent* ces différentes mesures

14

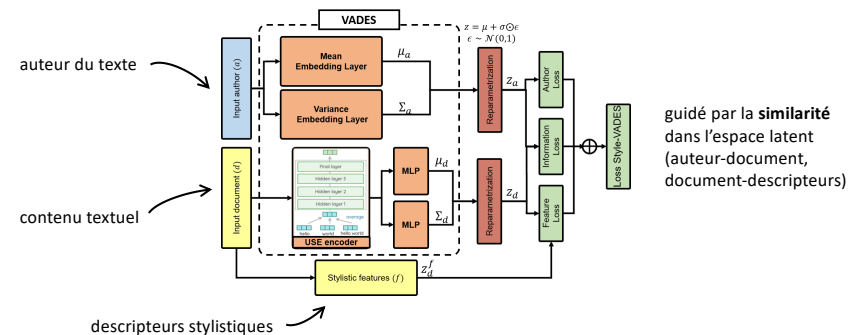
Comparaison des modèles



15

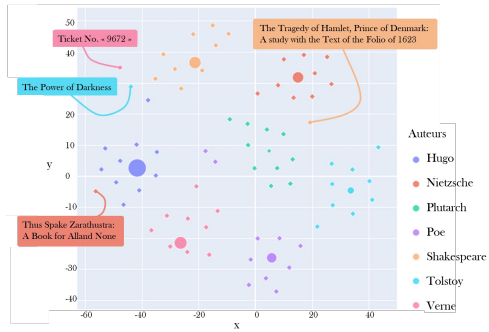
VADES : modèle de représentation des auteurs

(Terreau et al., arXiv 2024)



16

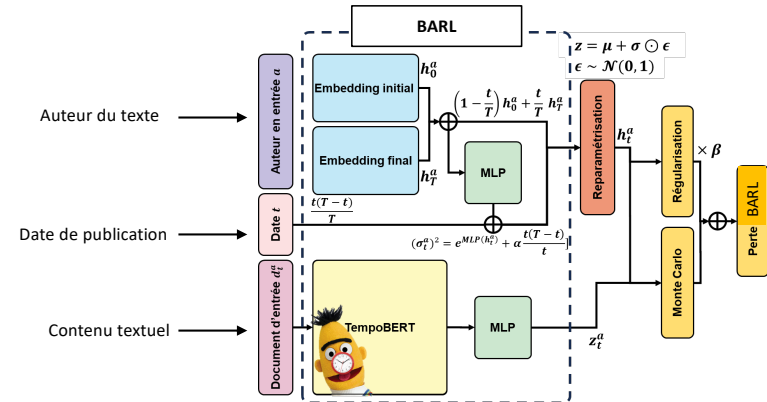
Application à l'analyse du style littéraire



Ici, il s'agit d'un extrait de données tirées du [Projet Gutenberg](#)

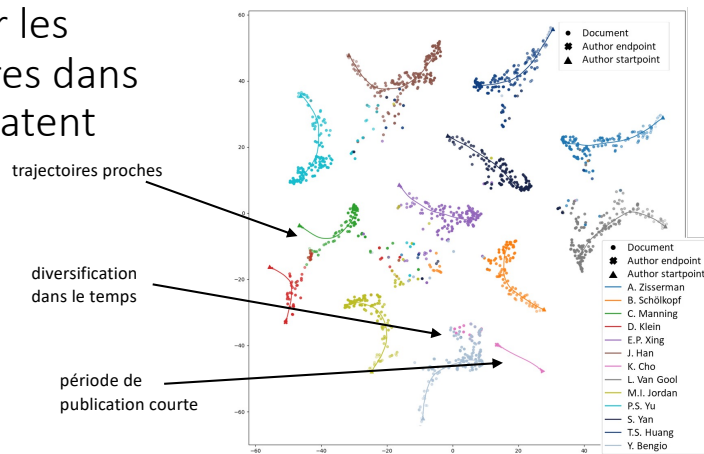
17

BARL : modèle pour apprendre des représentations temporelles (Terreau & Velcin, 2024)



18

Visualiser les trajectoires dans l'espace latent



Il s'agit ici de données issues de bases de données bibliographiques (*Semantic Scholar*).

19

Conclusion et perspectives

Julien Velcin, laboratoire ERIC

Colloque LIFRANUM, 24-25 octobre 2024

Conclusion

- Difficulté de trouver des problématiques de recherche en informatique (ou mathématiques) *directement liées* aux besoins immédiats en LLSHS
- Problème très intéressant et difficile, dommage qu'on n'ait pas réussi à travailler sur les données du projet...
- Toujours un chaînon manquant entre les développeurs de modèles et d'algorithmes et les chercheurs/utilisateurs en SHS
- Néanmoins, plusieurs contributions au domaine du TAL !
- Des échanges toujours enrichissants avec les partenaires LLSHS

21

Des pistes ?

- La question d'encoder le *style* n'est toujours pas résolue
- De nombreuses contributions pourraient être mises à disposition des chercheurs en LLSHS :
 - mesure / visualisation du style à l'aide des descripteurs stylistiques
 - calcul des proximités entre auteurs
 - système de requête et de visualisation basé sur ces nouveaux descripteurs
- Remise en cause du principe d'encodeur avec les mégas modèles de langue (LLMs) à base de décodeurs seuls (*decoder only*)

22

Références

- Terreau E., A. Gourru, J. Velcin: Writing Style Author Embedding Evaluation. Workshop Evaluation and Comparison of NLP Systems, co-situé avec EMNLP 2021
- Terreau E., A. Gourru, J. Velcin: Capturing Style in Author and Document Representation, <https://arxiv.org/abs/2407.13358>, 2024
- Terreau E. & Velcin J.: Building Brownian Bridges to Learn Dynamic Author Representations from Texts. Proceedings of International Symposium on Intelligent Data Analysis (IDA), Dublin, Avril 2024.

23