Plan de la présentation

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

La révolution du TAL

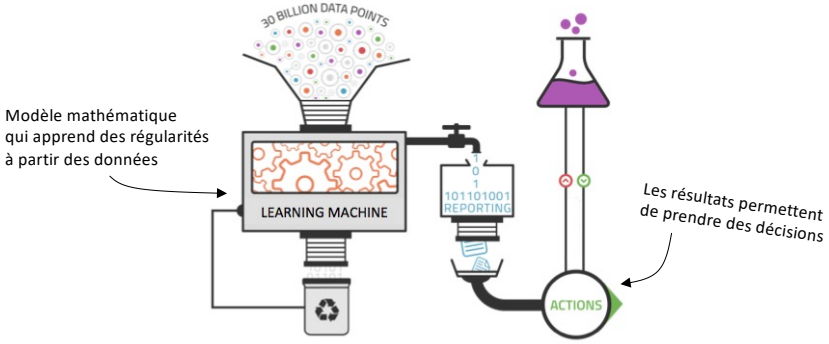
Julien Velcin, laboratoire ERIC

Colloque LIFRANUM, 24-25 octobre 2024

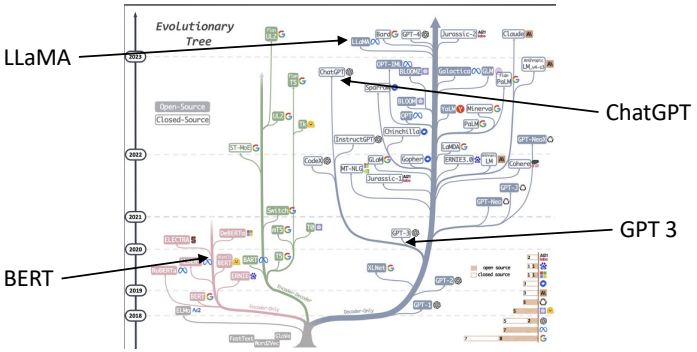
D'Enigma à ChatGPT

- 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…

Des règles à l'apprentissage automatique



Succès des grands modèles de langue (LLMs)



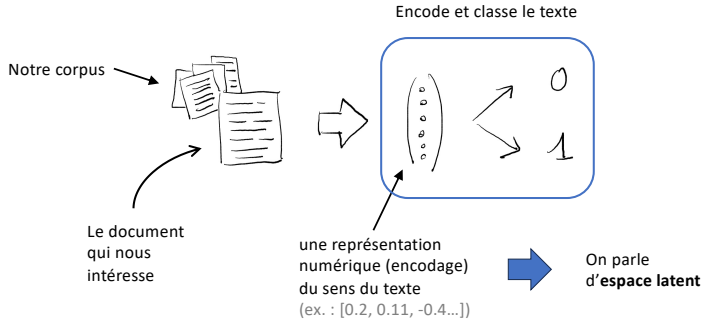
Note : ce graphique est en constante évolution

Encoder le sens des textes

Julien Velcin, laboratoire ERIC

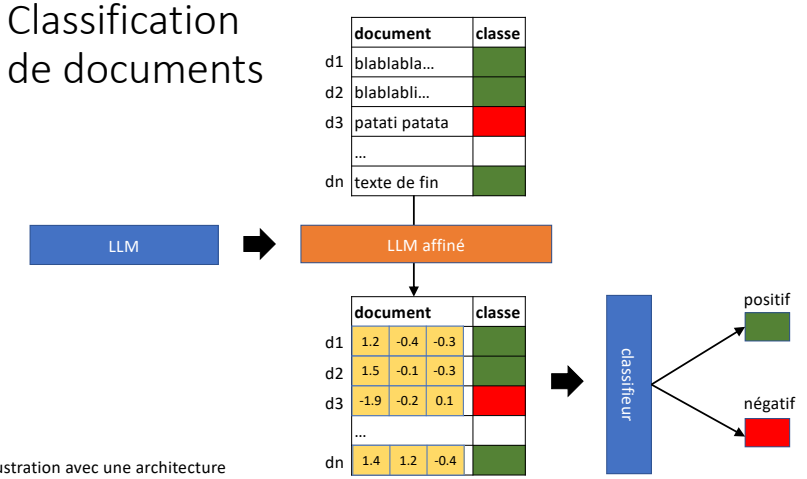
Colloque LIFRANUM, 24-25 octobre 2024

Encodeurs et représentations latentes



Recherche d'information

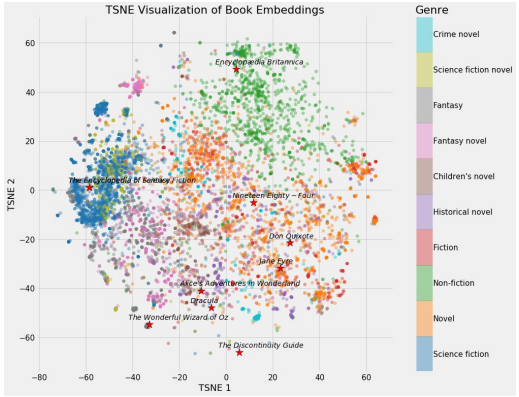
Classification de documents



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

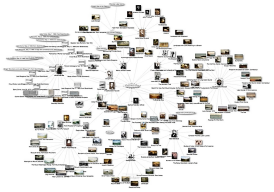
Représenter les proximités entre auteurs

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 ?

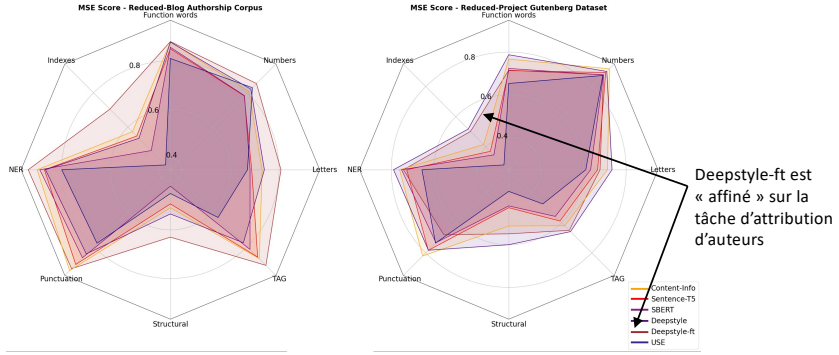
Mesurer le style littéraire (Terreau et al., 2021)

- 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
Structurel	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

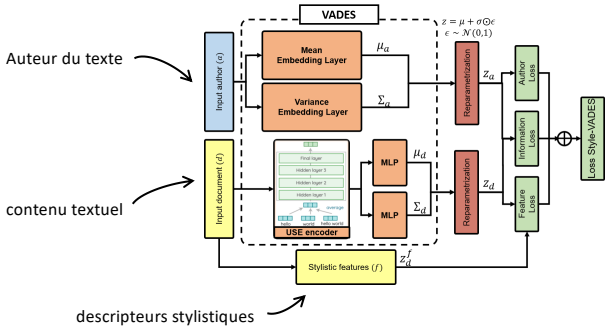
Comparaison des modèles



Deepstyle-ft est « affiné » sur la tâche d'attribution d'auteurs

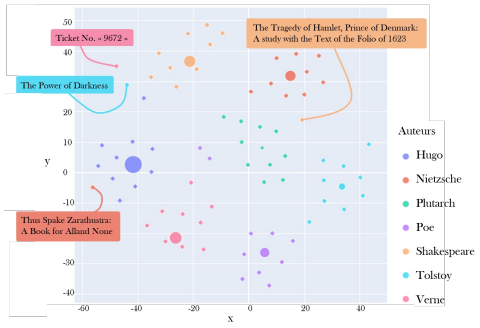
VADES : modèle de représentation des auteurs

(Terreau et al., arXiv 2024)



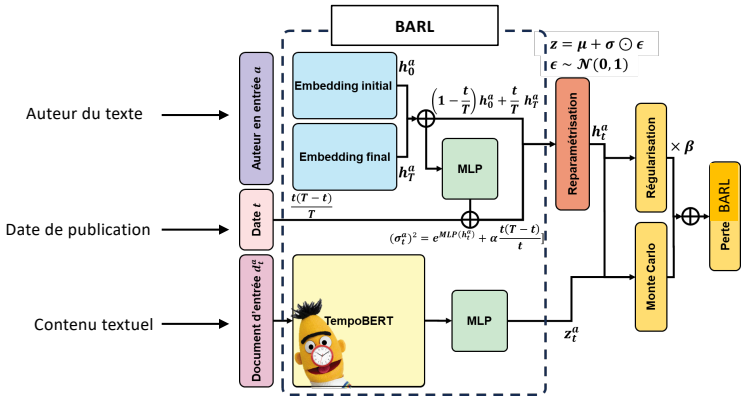
guidé par la **similarité** dans l'espace latent (auteur-document, document-descripteurs)

Application à l'analyse du style littéraire

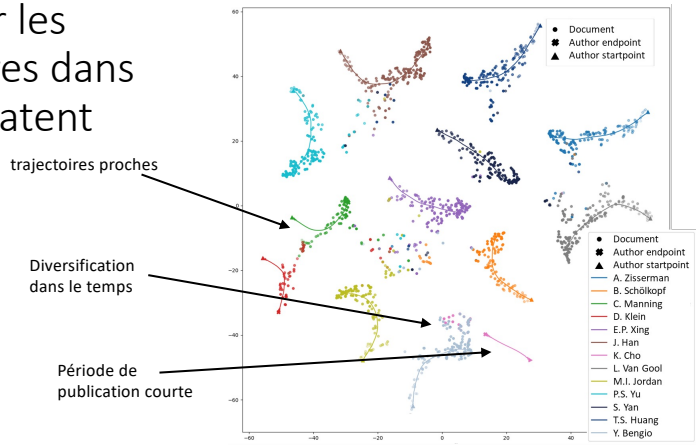


Ici, il s'agit d'un extrait de données sont tirées du [Projet Guthenberg]

BARL : modèle pour apprendre des représentations temporelles (Terreau & Velcin, 2014)



Visualiser les trajectoires dans l'espace latent



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

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 chainon 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

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*)

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.