Traiter des verbatims et questions ouvertes à l'aide d'un modèle d'IA local

Thomas Vroylandt[*]
Victoire Chatain[x]
Emmanuel Herbepin[%]

Résumé

Il est fréquent de disposer lors d'enquêtes ou de dispositifs citoyens de données issues de champs libres ou de verbatims. L'analyse classique de ces données[1] passe par le compte des mots et des unités de textes, puis la classification de ces mots à l'aide de dictionnaires. Des packages comme {*tidytext*}[2] existent pour faciliter ces analyses. Enfin, une approche supervisée à l'aide de BERT[3] donne en général de bons résultats, mais demande un travail préalable, parfois important.

De façon complémentaire, il est possible d'utiliser un modèle de langage (LLM) pour analyser de façon non supervisée ou peu supervisée ces verbatims, en identifier les objets principaux ou de les classer dans des catégories. Face à l'impact environnemental important, au coût financier non nul et aux risques de perte de la confidentialité des données, les modèles d'intelligence artificielle (IA) hébergés en local, peuvent être un bon compromis entre puissance et qualité.

Les packages {mall}[4] et {ollamar}[5] sur lequel il repose, proposent des fonctions plus faciles d'accès pour réaliser ces tâches d'analyse. Ils peuvent être complétés par {ellmer}[6] pour les cas plus avancés. Un certain nombre de précautions, à la fois dans la préparation des données, les instructions (*prompts*) aux modèles ou la vérification des résultats, doivent être mis en place pour garantir un résultat de qualité.

Mots-clefs: Text mining – IA – Enquêtes

Développement

Cette communication se fonde sur l'analyse de jeux de données disponibles sur data.gouv.fr contenant des verbatims issus de dispositifs citoyens de contributions, à la fois de façon classique et à l'aide du package {mall} (et dans une moindre mesure du package {ellmer}, qui est plus polyvalent). Ce package se fonde sur l'utilisation d'Ollama, qui permet d'héberger en local des modèles de langages de petite taille, et du package {ollamar} qui l'interface.

L'objectif est d'illustrer l'intérêt, le fonctionnement et les limites du recours à des modèles d'IA, en particulier local, pour l'analyse de données textuelles dans des champs libres, qui demandent souvent un traitement manuel lourd ou sont laissés de côté.

Sont notamment abordés :

- Le fonctionnement du package {mall} et son utilité
- La préparation des données
- Des précautions dans la rédaction des instructions

[*] Kantiles, thomas(at)kantiles.com
[x] Kantiles, victoire(at)kantiles.com
[%] Kantiles, emmanuel(at)kantiles.com

- Les vérifications à réaliser dans l'analyse des résultats
- Les autres options possibles (et notamment l'usage de BERT)

Références

1. Lebart L., Salem A., *Statistique textuelle*, 1994
2. Silge J., Robinson D., *Text Mining with R*, 2017 : https://www.tidytextmining.com/
3. Delvin J., Chang M., Kenton L., Toutanova K., "BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv*, 2018 : https://arxiv.org/abs/1810.04805
4. {mall} package : https://mlverse.github.io/mall/
5. Lin H., Safi T., {ollamar} package : https://hauselin.github.io/ollama-r/
6. Wickham H., Cheng J., Jacobs A., {ellmer} package : https://ellmer.tidyverse.org/