REsearch and methodology in Data Science

Projet - Pro.totype.duction

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

2023-2024

Objectifs

- Réaliser un projet de machine learning moderne
- Construction d'un dataset
- Analyse de résultats obtenus avec des modèles simples
- Mise en production
- Rédaction d'un rapport technique

Contexte

Suite aux récents changements de politique de plusieurs réseaux sociaux (Twitter et Reddit pour ne pas les nommer) concernant leur API, il est devenu difficile d'extraire des données de ces sites pour constituer des datasets d'entraînement. Cependant du fait de ces mêmes changements de politique, de plus en plus d'utilisateurs migrent vers des plateformes reposant sur les logiciels libres et des protocoles basés sur la fédération.

On trouve par exemple:

- Mastodon pour le micro-blogging,
- Pixelfed pour le partage de photos,
- Lemmy pour l'aggrégation de liens,
- Peertube pour le partage de vidéos.

Ce ne sont que des exemples, voir aussi:

- https://en.wikipedia.org/wiki/Fediverse
- https://github.com/BasixKOR/awesome-activitypub

Comme il s'agit d'un protocole fédéré, il y a plusieurs serveurs indépendants mais qui peuvent interagir entre eux.

Ces plateformes constituent un réseau unifié baptisé le Fediverse et reposent sur un protocole commun baptisé ActivityPub, ce qui permet d'interagir d'une plateforme à l'autre, par exemple de réagir à une photo postée sur un serveur Pixelfed à partir d'un compte sur un serveur Mastodon. Une conversation peut donc se répartir sur plusieurs serveurs et à travers plusieurs plateformes.

Il y a donc plein de types de données disponibles sur le Fediverse et chaque groupe se concentrera sur un aspect particulier. Par exemple:

- NER,
- systèmes de dialogue,
- Image et texte,
- vidéos.

Rapports

Rendu 1

Attendus du rapport:

- Description des objectifs du dataset
- Description des sources de données et aperçu des techniques utilisées
- Analyse descriptive des données

Remarques:

- **Pas de code**
- Chaque partie du rapport doit apporter quelque chose au lecteur
- Les choix faits ne sont pas définitifs, ils pourront être mis à jour au fur et à mesure du projet.
- Le rapport est à amender et à compléter en fonction des retours.

Rendu 2

Attendus du rapport:

- Tout ce qui précède
- Choix de modèles simples (pour un entraînement rapide)
- Analyse des résultats
- Pistes à explorer

Remarques:

- Les remarques précédentes restent valable.
- Pas besoin d'une étude expérimentale très large, l'objectif reste avant tout l'analyse et la présentation des résultats

Rendu final

Attendus du rapport:

- Tout ce qui précède
- Conclusion technique sur les performances du système, dans l'optique d'une utilisation réelle