APPORT DE LA CLASSIFICATION MULTILABELS POUR CARACTÉRISER DES OFFRES D'EMPLOI

CARTON Cérès, ccarton@hellowork.com, HelloWork, Rennes (Oratrice)

BEL-LETOILE Justine, jbel-letoile@hellowork.com, HelloWork, Pessac (Oratrice)

LEGAULT Pierre, plegault@hellowork.com, HelloWork, Pessac

Thématique : Analyse des données

Résumé : La classification de texte, c'est à dire le fait d'attribuer une étiquette à un document, est un des moyens existants en traitement automatique des langues (ou NLP) pour structurer la connaissance à partir de textes bruts. Mais comment associer la bonne étiquette lorsqu'on doit choisir parmi plusieurs milliers, dont certaines sont très proches voire se recoupent ? En s'aidant d'un cas pratique dans le domaine de l'emploi, nous allons étudier comment la classification multilabels peut être mise en œuvre pour mieux catégoriser des documents.

Mots clés : *NLP, classification de texte, classification multilabels, fastText, évaluation*

1. Introduction

HelloWork utilise l'intelligence artificielle pour faciliter la mise en relation entre les candidats et les recruteurs. Avant de publier une annonce, on cherche par exemple à détecter différentes informations : métier, localité, type de contrat, compétences attendues, etc. Pour la détection du métier, il s'agit d'associer à l'annonce un concept issu de notre thésaurus métier, parmi plusieurs milliers possibles.
Si on aborde cette question comme un problème de classification, on se retrouve dans une configuration particulière : un nombre important de classes, dont certaines hiérarchisées, des frontières floues entre les différents concepts rattachés aux classes, ainsi que des datasets fortement déséquilibrés.

2. Méthodologie

En s'inspirant des méthodes utilisées en classification extrême, on peut reformuler cette problématique en classification multiclasses et multilabels : on attribue à un texte une ou plusieurs classe(s), ici un ou plusieurs métier(s). Nous comparons un modèle de classification de texte unilabel et une approche multilabels, et nous allons voir comment le fait d'autoriser plusieurs labels peut améliorer la compréhension automatique d'un texte. Parmi les méthodes possibles, nous proposons notamment un focus sur une approche via fastText.

3. Originalité / perspective

La classification de texte est un sujet courant en NLP (analyse de sentiment, spam filtering, etc.). Ici le grand nombre de classes et les potentielles collisions sur les données disponibles nous poussent à explorer d'autres méthodes que celles habituellement exposées, tant en terme de préparation du dataset que du choix du modèle et de ses paramètres.

De plus, la question de l'évaluation est cruciale : quelles métriques suivre dans un cas multiclasses – multilabels ? Comment pondérer la gravité des erreurs ou juger les prédictions partielles pour avoir une meilleure intuition de la performance de la solution ? De plus, comment tenir compte du déséquilibre des classes ?

Références

- Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification (cite arxiv:1607.01759)
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. & Mikolov, T. (2016). FastText.zip: Compressing text classification models (cite arxiv:1612.03651Comment: Submitted to ICLR 2017)
- *Workshop on eXtreme Classification: Theory and Applications.* ICML 2020. https://icml.cc/Conferences/2020/ScheduleMultitrack?event=5719