Recherche en IA

Rattrapage

Sujet : Liens entre TAL et Trademark

DEBET Stanislas
GOBBO Fanny

Sommaire

Introduction

Le domaine de l'IA est en constante évolution, et cette affirmation est encore plus vraie ces temps-ci avec l'émergence de plus en plus fréquente de nouveaux modèles. La matière de Recherche en IA est un excellent moyen de nous initier au concept de veille technologique car il est nécessaire pour toute personne travaillant dans le domaine de l'informatique de rester à jour sur les technologies actuelles et encore plus vis-à-vis de l'intelligence artificielle.

Parmi les sujets proposés pour cette année, nous avons choisi le lien entre TAL et Trademark.

Chaque année, un nombre monumental de marques est déposé, dont de plus en plus de marques frauduleuses. Afin de vérifier si une marque déposée est frauduleuse, le processus passe actuellement par une recherche manuelle qui a 2 principaux défauts : le procédé est long et il est facilement possible de rater des marques similaires dans la masse immense de marques déjà déposées. L'objectif de notre recherche est de re-mettre en application des propositions de procédés permettant de solutionner complètement ou partiellement le processus de validation de dépôt de marque par rapport à ces marques frauduleuses. Lors de la première étape de cette recherche, nous devions trouver des articles proposant des solutions à ce problème.

Dans un premier temps, nous commencerons par une remise en contexte des articles choisis pour la première partie de cette matière tels que nous les avions présentés lors du premier oral, en entrant plus en détail sur l'article choisi pour la seconde partie. Ensuite, nous expliquerons la démarche réalisée pour l'implémentation, suite à quoi nous présenterons les résultats obtenus. Enfin, nous conclurons avec un retour critique sur notre démarche et sur ce qui a ou n'a pas fonctionné avec cette-dernière.

Récapitulatif des recherches de articles

La recherche de papiers a été effectuée avec les outils proposés lors de la présentation de la matière : ArXiv, Google Scholar, worldtrademarkreview, dblp.
Trois papiers ont été retenus lors de notre première présentation :

- TradeMarker - Artificial Intelligence Based Trademarks Similarity Search Engine, *Juillet 2019, Idan Mosseri, Matan Rusanovsky, and Gal Oren*
- Trademark approval system using semantic similarity computation, *Avril 2018, Vijay Karvnde, Sagar S. Sanghavi*
- Text Similarity Estimation Based on Word Embeddings and Matrix Norms for Targeted Marketing *Juin 2019, Tim vor der Bruck, Marc Pouly*

Les trois articles proposaient des méthodes d'implémentation, cependant ils n'étaient pas forcément reproductibles et/ou adaptés à notre situation. Le plus proche était le TradeMarker, mais il proposait une approche plus globale sur le traitement des marques, comprenant aussi le logo et ne montrait la démarche complète de recherche de similarités ni les résultats obtenus via la méthode de comparaison textuelle.

Finalement, après d'autres recherches, nous avons trouvé un papier qui nous paraissait traiter avec plus de justesse le sujet que nous avions choisi. Il s'agissait de mettre en place un algorithme pour recueillir des données de marque (nom, description), et les donner à un modèle BERT pour en obtenir un vecteur facilement comparable au moyen de la similarité cosinus. L'objectif de reproduire une étude est donc atteignable dans ce cas.

Le papier que nous avons choisi est le suivant :

- A Novel Patent Similarity Measurement Methodology : Semantic Distance and Technological Distance, *Mars 2023, Yongmin Yooa, Cheonkam Jeongb, Sanguk Gimc, Junwon Leed, Zachary Schimkee, Deaho Seo*

L’algorithme

Afin d’implémenter le modèle décrit dans l’article choisi précédemment, nous avons tout d’abord cherché un dataset qui permettrai d’appliquer l’implémentation. Nous en avons trouvé un de 4Mo avec des marques françaises (Source : data.inpi.fr).

A partir du dataset, nous avons commencé par une phase de nettoyage (*reshape_data.py)* pour obtenir :

- N° de la marque
- Marque
- Type de la marque
- Date de dépôt/enregistrement
- Produits et services

Suite à cela, nous avons réalisé une phase d’échantillonnage (*test_sampling.py*) en retirant des données du dataset pour créer des données qui devraient techniquement être "validées" comme étant des "nouvelles" données. Nous avons aussi copié une partie des données pour en modifier légèrement le contenu afin d’avoir des fausses données qui ne devraient techniquement pas être validées.

Chaque marque est ensuite traitée par le modèle de BERT pour construire un vecteur représentatif du nom et des produits et services de la marque. (*model.py*).

Afin d'optimiser les temps de calcul et éviter d’avoir à tout recalculer à chaque fois, les embeddings sont sauvegardés dans un fichier csv contenant :

- Nom de la marque
- Embedding (vecteur de 384 réels)

Le calcul des embedding est effectué en concaténant le nom de la marque à la description textuelle de celle-ci, puis en calculant le vecteur d’embedding de toute cette chaîne de caractère avec le modèle de BERT *all-MiniLM-L6-v2* (celui du papier de recherche).

Pour vérifier si une nouvelle marque à déposer n’est pas frauduleuse, on extrait son vecteur d’embeddings (*main.py* : *check()*) qu’on compare à ceux de la table. Si une valeur de similarité cosinus obtenue est supérieure à un seuil déterminé par l’utilisateur, alors l’algorithme retourne la/les marque/s de la table concernées pour un examen manuel.

Résultats obtenus

Bien qu'il nous a été possible d'obtenir des résultats à partir de notre dataset, il ne nous a pas été possible de comparer les résultats avec ceux obtenus dans le papier original car nous n'avions pas le même dataset. Ainsi, nous avons donc reproduit *des* résultats à partir de l'article, mais ce ne sont pas *les* résultats exacts de l'article.

En ce qui concerne la méthode d'obtention de résultats, il faut noter que le calcul d'un vecteur d'embeddings et sa la comparaison de tous les autres vecteurs de la table est long : environ 2 secondes sur une machine sans GPU.

En matière de résultats purs, voici ce que nous avons :

- On observe que les marques contenant un mot similaire voire identique à celui d'autres marques obtiennent une similarité cosinus élevée (ex : **chateau d'anice** et **chateau de l'hurbe**, similarité de **0.84** bien que la description ne soit pas la même).
- Pour les marques qui ont été extraites de la base de données et qui sont donc censées être validées comme nouvelles, aucune similarité n'a été trouvée (voir fin du fichier *results.csv*).
- Pour les marques copiées de la base et falsifiées, qui doivent donc être similaires à au moins une autre marque, une similarité a été trouvée avec au minimum une marque de la base.

Si on tente de comparer ces résultats aux résultats obtenus dans l'article d'origine, nos résultats sont clairement supérieurs à ceux donnés dans le tableau 3 du papier exploité. Les causes peuvent être multiples :

- Les données que nous avons utilisées ne sont pas les mêmes,
- lLs données ne sont pas suffisamment variées pour faire diverger les résultats efficacement,
- Notre méthode de construction de chaîne de caractères à donner à BERT n'est pas la même due à la structure du tableau de données.
- La méthode de falsification des marques repose sur des changements de mots par leurs synonymes, cette méthode est assez bateau et est loin de la réalité de la falsification de marques.

Conclusion

Au cours de cet exercice, nous avons pu avoir un aperçu du travail de veille, essentiel à notre futur métier d'ingénieur. Nous avons pu re-mettre en application une méthode présentée dans un article scientifique. L'exercice de reproduction du procédé a été très intéressant à faire car il nous permettait d'apprendre par nous même à reproduire des modèles présentés dans de tels articles. Au premier semestre, lors de l'oral de présentation, nous avions présenté notre reproduction du code sans nous attarder sur les résultats. Nous étions donc passés à côté de l'exercice demandé, à savoir : la reproduction des résultats (et non du procédé seulement). Dans ce rapport, nous vous avons présenté nos résultats un peu plus en détail malgré la différence entre notre dataset et celui utilisé pour l'article. Nos résultats semblent tout de même cohérents au vu des données que nous avons recueillies et de notre méthode de falsification : celle-ci a été développée succinctement et n'est donc pas aboutie pour imiter de vrais fraudeurs. De plus le format et le contenu des données ne sont pas ceux du papier de recherche, nous avons donc construit une base qui nous semblait cohérente. Ainsi les résultats ne sont pas représentatifs du papier mais nous ont permis tout de même de mettre en place le procédé détaillé.