Projet Interpromo 2021

Groupe 8 : Innovation

Similarités entre articles

Cheffe : Flora Estermann
Cheffe adjointe : Katia Belaid
Chef qualité : Tanguy Dirat

Samba Seye
Célya Marcélo
Théo Saccareau
Damien Sonneville
Mélina Audiger
Jordi Mora Fernandez

SOMMAIRE

1. Introduction

Dans le cadre du soutien apporté au groupe G7 Interface Web, nous avons contribué à produire un dataset établissant la liste de correspondances entre articles similaires, associée au score de similarité entre chaque paire d'articles.

Notre travail a notamment reposé sur la représentation des documents sous forme vectorielle, avant d'appliquer des mesures de similarités données afin d'évaluer les distances séparant chacune de nos instances et de pouvoir les comparer.

Dans un premier temps, nous présenterons certaines des méthodes d'embeddings testées, puis nous détaillerons les méthodes que nous avons choisies afin de calculer la similarités entre les articles considérés en suivant.

2. État de l'art

Pour calculer la similarité entre deux articles, nous avons utilisé des vecteurs de mots avec la bibliothèque FastText créée par le laboratoire de recherche sur l'IA de Facebook. Le modèle permet de créer un apprentissage non supervisé ou un algorithme d'apprentissage supervisé pour obtenir des représentations vectorielles des mots. Le choix de cette bibliothèque est expliqué par le choix de la langue possible puisque FastText peut être utilisé pour 157 langues différentes, dont le français. Le FastText permet une représentation de mots dans une dimension de 300.

Avec ces représentations des mots, nous avons pu calculer le vecteur d'un article en sommant ou en prenant la moyenne des vecteurs de mots présents dans le texte. Pour le calcul de la similarité à partir de ces vecteurs d'articles, on a pu utiliser la fonction cosinus qui fait le produit scalaire des deux vecteurs divisé par le produit des normes. Le résultat est compris entre 0 et 1.

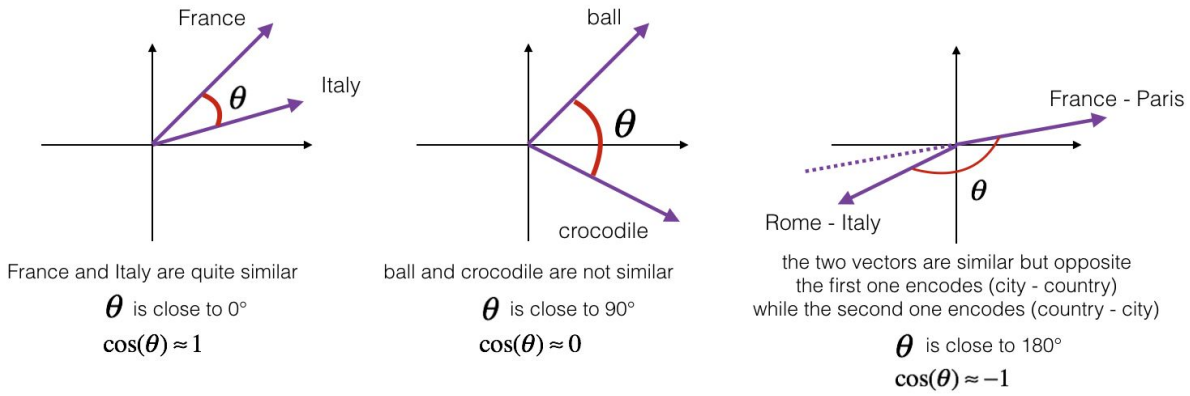

Figure 1. Exemples de calcul de la similarité cosinus.

Une autre méthode que nous avons utilisé est le Word Mover Distance (WMD) où les embeddings de mots sont incorporés dans le calcul de la distance entre deux documents. Avec des embeddings de mots pré-entraînés donnés par FastText, les différences entre les documents peuvent être mesurées avec des significations sémantiques en calculant la distance minimale dont les mots incorporés d'un document ont besoin pour parcourir pour atteindre les mots incorporés d'un autre.

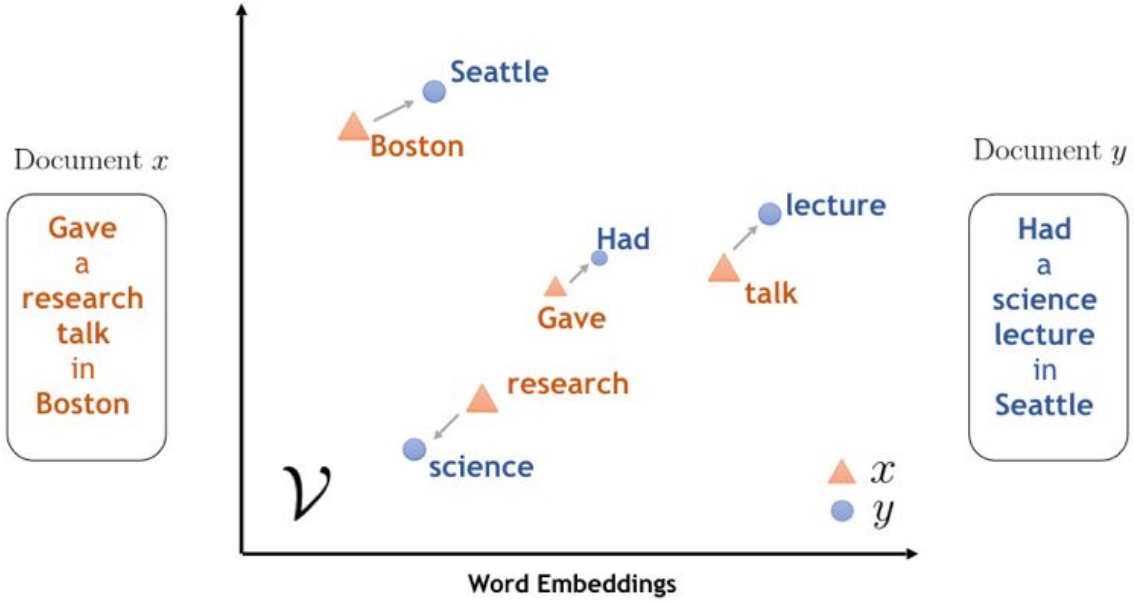

Figure 2. Exemple de calcul de la Word Mover Distance entre deux textes.

3. Résultats

Les deux méthodes nous donnent à peu près les mêmes résultats. Plus la distance entre les articles est faible, plus la similarité entre ces deux articles est élevée. Malheureusement, le WMD a un temps d'exécution plus élevé que l'utilisation de la similarité cosinus. C'est donc la similarité cosinus que nous avons utilisé pour la suite.

Avec le calcul de similarité, on a pu enregistrer pour chaque article, les 10 articles les plus similaires avec le score de similarité correspondant. Nous avons enregistré les résultats dans un dataFrame pour le groupe G7.

4. Discussion

A terme, le groupe G7 Interface Web ira itérer sur la table de similarités entre articles créée, qui aura au préalable été intégrée par le groupe G4 dans la base de données, afin d'afficher un aperçu des articles similaires en bas de chaque article consulté et ainsi de permettre à l'utilisateur de parcourir du contenu similaire sans passer par la barre de recherche.