Rapport de projet TAL – 2020-2021

Sujet : modèle Word2vec pour la construction de vecteurs de mots

Romane Gallienne, Cécile Guitel, Romy Rabemihanta

Introduction

Pour ce projet, nous nous proposons d’implémenter un modèle d’apprentissage pour obtenir des plongements lexicaux, aussi appelés word embeddings grâce à la libraire pytorch et à un réseau neuronal. Notre objectif est donc de construire des vecteurs de mots, autrement dit un espace vectoriel de mots en utilisant la méthode de Skipgram Negative Sampling proposée par word2vec.

Données

Pour mener ce projet, nous avons besoin de données, en particulier d’un large corpus de texte écrit. Nous utilisons donc un corpus libre de droits issu du journal "L'Est républicain", où chaque archive contient environ 40 millions de mots (26 fichiers de 100 000 phrases chacun, environ 1,6 million de mots chacun). Les fichiers sont au format une phrase par ligne, les phrases sont segmentées en tokens et les mots sont taggés par leur catégorie syntaxique.

Pour évaluer la qualité des plongements lexicaux obtenus de manière intrinsèque, nous avons également besoin de données de similarité entre des mots évaluée par des humains. Puisque nous travaillons en français, nous utilisons le jeu de paires de mots RG65 disponible au format HTM : https://www.site.uottawa.ca/~mjoub063/wordsims.htm.

Méthode

Partie extraction

Cette partie s’effectue dans le fichier init.py.

Partie apprentissage

Partie évaluation

Nos difficultés

Améliorations

Références