PROJET : Natural Language Processing

Analyse des sentiments

Nicolas Defoort

Sommaire :

Introduction :

Suite aux cours de natural language processing, j'ai réalisé un projet mettant en application les notions appris ce semestre.

Le but du projet est de prédire le sentiment évoqué dans une phrase en utilisant une architecture de deep learning utilisant les RNN. Après un entraînement de notre modèle, l'objectif est de créer un pipeline contenant en entrée une phrase et de récupérer en sortie la connotation de celle-ci (positive ou négative).

Jeu de données :

Afin d'entrainer le modèle, j'ai utilisé le jeu de données disponible sur Sentiment140 dataset with 1.6 million tweets | Kaggle.

Ce jeu de données contient 1 600 000 tweets venant de l'API twitter. Les tweets ont été évalués afin de détecter le sentiment de la personne l'ayant écrit (0 = négatif, 4 = positif).

Le jeu de données contient 6 colonnes :

- Target : contiens le sentiment de la personne l'ayant écrit
- ID : id du tweet
- Date : date à laquelle le tweet à été publié
- Flag : contiens la valeur NO_QUERY
- User : pseudo de la personne ayant posté le tweet
- Text : Le contenu du tweet

Dans ce projet, j'utiliserai uniquement les composants « target » et « text » du jeu de données puisque les autres catégories ne comportent aucune information pouvant être utilisé pour analyser les sentiments. Ainsi, je réduirai mon dataframe contenant ces données.

Dataframe utilisée :

	feeling	text
0	0	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	is upset that he can't update his Facebook by ...
2	0	@Kenichan I dived many times for the ball. Man...
3	0	my whole body feels itchy and like its on fire
4	0	@nationwideclass no, it's not behaving at all....
...	...	...
1599995	4	Just woke up. Having no school is the best fee...
1599996	4	TheWDB.com - Very cool to hear old Walt interv...
1599997	4	Are you ready for your MoJo Makeover? Ask me f...
1599998	4	Happy 38th Birthday to my boo of alll time!!! ...
1599999	4	happy #charitytuesday @theNSPCC @SparksCharity...

Détails de la solution :

Afin de pouvoir créer le modèle, on doit préalablement traiter les données, j'ai utilisé la classe tokenizer de keras qui permet de mettre sous forme de vecteur d'integer un corpus de texte. Par default, toute la ponctuation est supprimée et les mots sont séparées en fonction des espaces dans la phrase. Cette séquence est ensuite séparée en liste de tokens puis vectorisé.

On récupère donc ces données vectorisées ainsi que les sentiments de l'utilisateur qui eux aussi sont vectorisés pour ensuite créer le modèle.

Le modèle est produit via un réseau de neurones récurrent (RNN) qui est un réseau de neurones spécialisé dans le traitement des séquences de valeurs ou séries temporelles.

Résultats :

Pour tester le modèle effectué, j'utilise l'API yelp qui fournit un json comportant plusieurs tweets.

J'ai pu appliquer mon modèle sur le jeu de données de l'API Yelp contenant des commentaires, les résultats semblent concorder avec le contenu du json.

Ces résultats peuvent toutefois être inexacts dus à la faible quantité de tweets utilisés pour effectuer mon modèle. J'ai utilisé 2000 tweets pour réaliser le modèle, si le modèle avait été réalisé avec les 1 600 000, la précision aurais été beaucoup plus amélioré.

Améliorations possibles :

La solution proposée peut être améliorée de différentes façons, il est toujours possible d'améliorer le réseau de neurones afin d'obtenir une meilleure précision. De même, il est aussi possible d'entraîner le modèle avec une plus grande quantité de tweets afin d'obtenir le moins d'erreur possible au niveau de la prédiction.

Il est aussi possible par la suite de récupérer les pseudos des utilisateurs les plus positifs et pessimistes sur leurs tweets afin de connaître la morale de la population sur un mois ou une année en fonction de leur âge, région, statut social...

Conclusion :

Ce projet m'a permis de mieux comprendre la création de modèle et l'architecture du deep learning en utilisant les RNN que je n'avais pas totalement compris au premier semestre. Même si je suis satisfait de ce projet, j'aurais aimé rajouter des graphiques représentant la répartition de la taille des tweets en fonction du sentiment de l'utilisateur ou récupérer les pseudos des utilisateurs ayant mis le plus d'avis positif/négatifs que je n'ai malheureusement pas pu rajouter par manque de temps.