Data Science Avancée

Thomas Falcone Lucas Gonçalves Yannick Mayeur

18/08/2019

Table des matières

TP1: Mise en place d'une classification textuelle

Introduction

Il nous semble intéressant d'analyser les tweets de différents politiciens afin de pouvoir ensuite voir si il est possible de classifier un tweet au bon politicien.

Le code est disponible sur le repository suivant :

https://github.com/ThomasF34/WhoTweetedThis

Outils

Nous avons choisi d'utiliser Python3 comme langage avec la libraire *fasttext*. Cette librairie, permet de facilement faire une classification supervisé à partir d'un ou plusieurs datasets.

Les données

Afin d'effectuer notre analyse, nous avons utilisé des données mises à disposition sur le site Kaggle.

Nous avons choisi d'exploiter les tweets des personnalités suivantes:

— Donald Trump, 45ème président des États-Unis (actuel)
— Barack Obama, 44ème président des États-Unis
— Hillary Clinton, politicienne, candidate à la 45ème élection présidentielle des États-Unis

Le processus de traitement

Formatage des données

Afin d'effectuer un traitement sur les différentes données que nous avons recueillies, nous avons mis en place un processus d'harmonisation des données afin qu'elles soient homogènes et utilisables par la librairie *fasttext*.

Pour ce faire il a fallu:

— Extraire les données des différents CSV.
— Mettre en forme les données en ajoutant sous la forme: "`__label__nom texte`".
— Mettre toutes les données en petite case afin de réduire le nombre de mots.
— Retirer les URL dans les tweets.
— Ecrire les données dans un nouveau fichier 75% de ces données, dont on se servira pour l'apprentissage.

Classification supervisée

Pour le processus de classification nous avons procédé à un certain nombre d'optimisations:

— La réduction du nombre de mot décrite dans la partie précédente: cela permet de réduire le vocabulaire et donc d'augmenter la précision
— La suppression des URL car celles-ci ne nous ont pas parues significatives pour l'analyse.
— Nous avons augmenté le nombre "d'epoch" afin d'augmenter le nombre de fois que chaque donnée est vue. Une augmentation à 25 nous a permis d'observer une bonne augmentation de la performance.

— Nous avons ajusté le “learning rate”, qui représente a quel point le modèle change après chaque donnée. Un “learning rate” de 0,3 nous a permis d’observer les meilleurs performances.

Prédiction

Suite à la mise en place de la classification supervisée nous avons fait la vérification de notre modèle avec 25% des données initiales.

Cela nous a permis d’obtenir les valeurs suivantes:

```
Hillary: 2629
Donald: 7375
Obama: 6851
Read 0M words
Number of words:  17944
Number of labels: 3
Progress: 100.0% words/sec/thread:  570620 lr:  0.000000
loss:  0.109835 ETA:   0h 0m
N       4843
P@1     0.957
R@1     0.957
```

Nous pouvons observer que *fasttext* maximise à la fois le rappel et la précision, et retient donc le point d’équilibre qui est à 95,7% de précision et de rappel. Ce résultat est bien meilleur que sans les optimisations, ou la précision et le rappel était de 92,9%. Celles-ci nous permettent d’obtenir une augmentation significative.

Conclusion

En conclusion la librairie de classification supervisée nous a permis de mettre en place une classification textuelle des tweets des trois personnalités que nous avons choisies.

Nous avons pu voir qu’il est important de préparer au mieux les données afin d’augmenter la performance de l’outil. L’optimisation de chaque élément permet d’obtenir un ensemble plus cohérent et donc plus facile à analyser.