The 20 newsgroups

Pereg Hergoualc’h

Le dataset

Le dataset est constitué de 11 314 articles de journaux rangés dans 20 catégories.

Pour rendre la prédiction plus difficile, j'ai supprimé les headers, footers et quotes au moment de l'import.

Voilà en exemple, le premier article de la base.

```
I was wondering if anyone out there could enlighten me on this car I saw
the other day. It was a 2-door sports car, looked to be from the late 60s/
early 70s. It was called a Bricklin. The doors were really small. In addition,
the front bumper was separate from the rest of the body. This is
all I know. If anyone can tellme a model name, engine specs, years
of production, where this car is made, history, or whatever info you
have on this funky looking car, please e-mail.
```

Le Tokenizer

Afin d’encoder les articles pour les entrer dans le réseau de neurones j’ai utilisé la class Tokenizer() de Keras.

Cela permet de transformer du texte en matrice. J’ai fait le choix d’utiliser une taille de 15000. Chaque valeur de X sera alors un vecteur de 15000 par 1.

Le model

J’ai utilisé un réseau de neurones de plusieurs couches afin de prédire la catégorie de l’article. Il s’agit de couches Dense et Dropout à la suite

Nous avons une input size de 15000 et une sortie de 20, ce qui correspond au 20 classes.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 512)	7680512
activation (Activation)	(None, 512)	0
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 512)	262656
activation_1 (Activation)	(None, 512)	0
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 20)	10260

Total params: 7,953,428
Trainable params: 7,953,428
Non-trainable params: 0

Résultats

Le modèle obtient une accuracy de 0.6714.

Nous pouvons observer la matrice de confusion de y_test et y_pred.
La diagonale est bien visible ce qui démontre que la plupart des articles ont été bien classés.