

RAPPORT DE PROJET :

MODELE DE

REGRESSION

LOGISTIQUE

GOODREADS BOOKS

4IABD1

ANNA DIAW

NADEJDA DOROSENCO

ADEM AIT IDIR



OBJECTIF DE L'ETUDE ET DESCRIPTION DU DATASET :

Dans le cadre du projet, nous avons créé un modèle de régression logistique pour prédire la note d'un certain nombre de livres à partir des commentaires des lecteurs fournis dans le dataset de base et ainsi fournir une classification pour chaque livre.

Il nous a été fourni 3 datasets, un pour les tests, un autre pour le training et le dernier qui représente le fichier de soumission.

Le fichier destiné au training est un tableau de 11 colonnes que sont les suivantes :

user_id	book_id	review_id	review_text	date_added	date_updated	read_at	started_at	n_votes	n_comments
d1c1f97f891c392b1105959b56e	7092507	5c4df7e70e9b438c761f07a4620ccb7c	** spoiler alert ** \n This is definitely one ...	Sat Nov 10 06:06:13 -0800 2012	Sun Nov 11 05:38:36 -0800 2012	Sun Nov 11 05:38:36 -0800 2012	Sat Nov 10 00:00:00 -0800 2012	1	0
d1c1f97f891c392b1105959b56e	5576654	8aeaaf13213eeb16ad879a2a2591bbe5	** spoiler alert ** \n "You are what you drink..."	Fri Nov 09 21:55:16 -0800 2012	Sat Nov 10 05:41:49 -0800 2012	Sat Nov 10 05:41:49 -0800 2012	Fri Nov 09 00:00:00 -0800 2012	1	0
d1c1f97f891c392b1105959b56e	15754052	dce649b733c153ba5363a0413cac988f	Roar is one of my favorite characters in Under...	Fri Nov 09 00:25:50 -0800 2012	Sat Nov 10 06:14:10 -0800 2012	Sat Nov 10 06:14:10 -0800 2012	Fri Nov 09 00:00:00 -0800 2012	0	0
d1c1f97f891c392b1105959b56e	17020	8a46df0bb997269d6834f9437a4b0a77	** spoiler alert ** \n If you feel like travel...	Thu Nov 01 00:28:39 -0700 2012	Sat Nov 03 11:35:22 -0700 2012	Sat Nov 03 11:35:22 -0700 2012	Thu Nov 01 00:00:00 -0700 2012	0	0
d1c1f97f891c392b1105959b56e	12551082	d11d3091e22f1cf3cb865598de197599	3.5 stars \n I read and enjoyed the first two ...	Thu Oct 18 00:57:00 -0700 2012	Mon Apr 01 23:00:51 -0700 2013	Sat Mar 30 00:00:00 -0700 2013	Fri Mar 29 00:00:00 -0700 2013	0	0

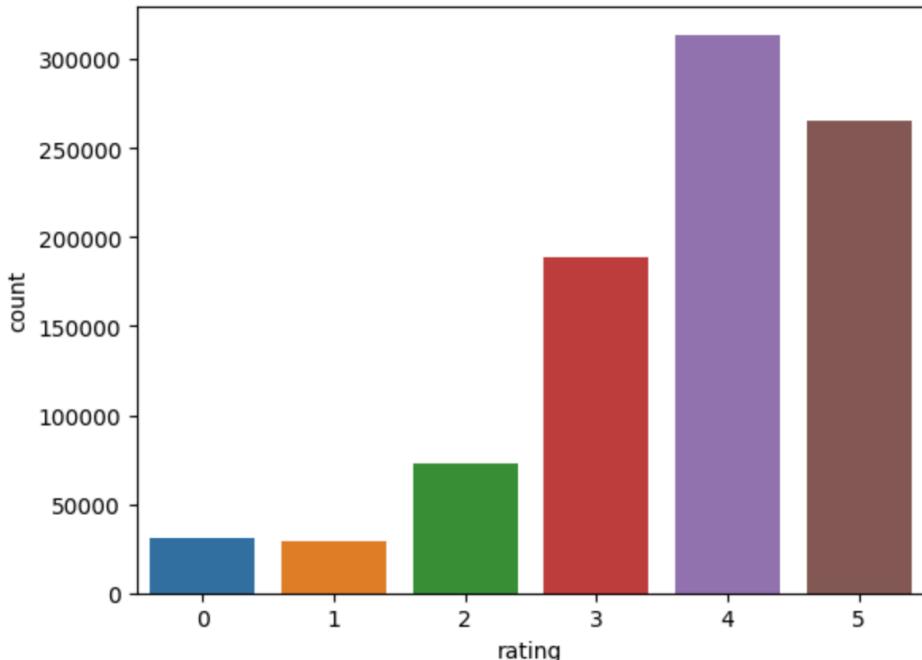
Le fichier de test est un fichier csv contenant 10 colonnes qui sont exactement les mêmes que celles de training excepté la colonne correspondant à la note (rating) que nous devons justement prédire dans le cadre du training.

Data columns (total 11 columns):			
#	Column	Non-Null Count	Dtype
0	user_id	900000	non-null object
1	book_id	900000	non-null int64
2	review_id	900000	non-null object
3	rating	900000	non-null int64
4	review_text	900000	non-null object
5	date_added	900000	non-null object
6	date_updated	900000	non-null object
7	read_at	808234	non-null object
8	started_at	625703	non-null object
9	n_votes	900000	non-null int64
10	n_comments	900000	non-null int64

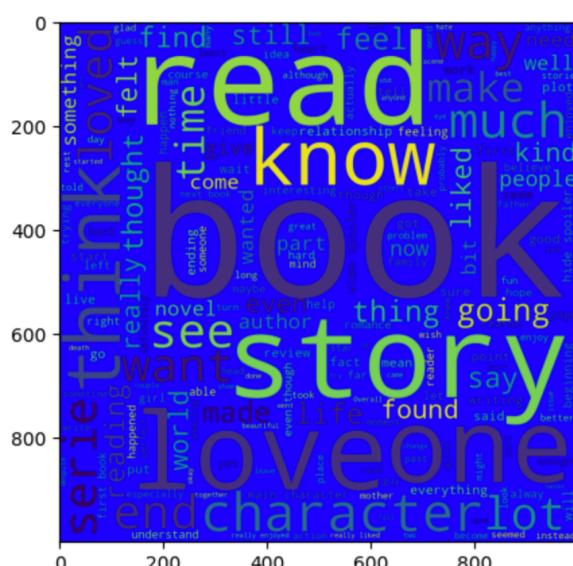
METHODE DE TRAITEMENT :

Nous avons débuté par analyser nos données.

Nous avons recueillis le nombre de commentaire (review) pour chaque note (rating) et avons visualisé les résultats par un diagramme de barre.

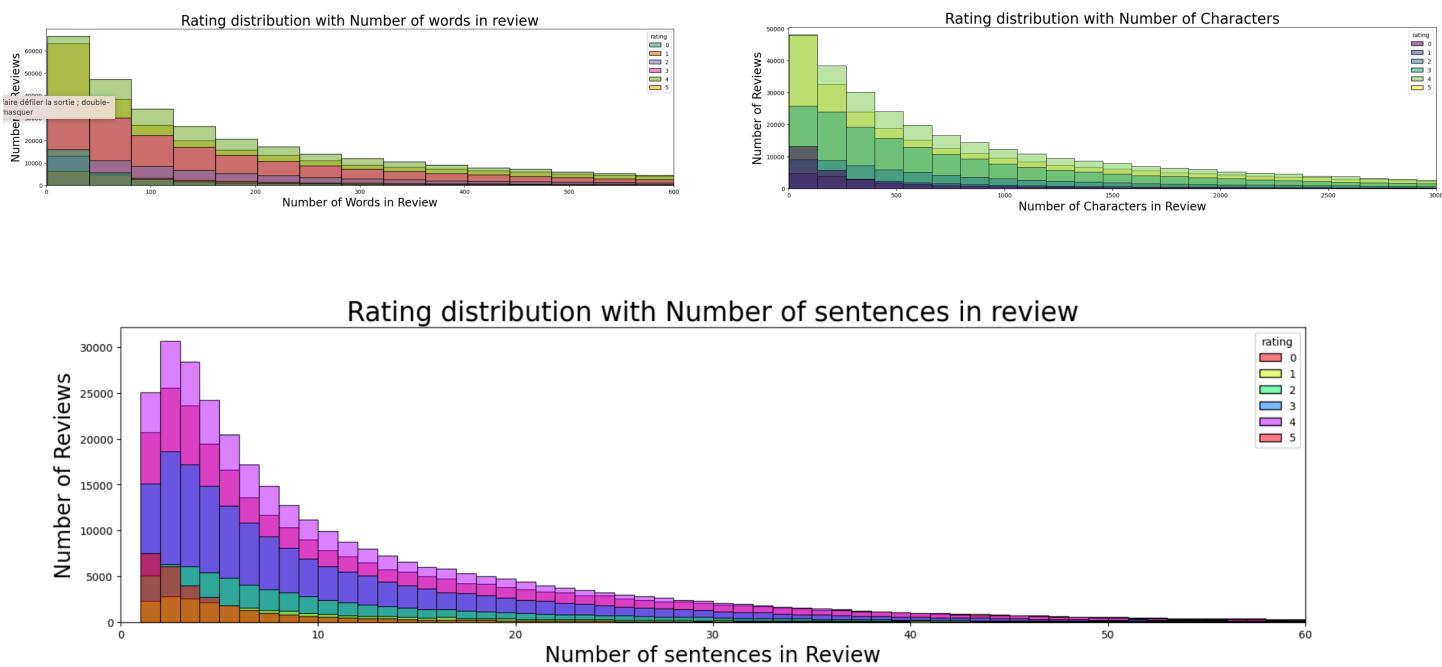


Nous avons ensuite décidé d'aller plus loin en comptant le nombre de caractères, mots et de phrases dans chaque commentaire et l'afficher dans un nuage de mots pour déterminer les mots qui ressortent le plus souvent.



Nous avons ensuite classé les commentaires en fonction du nombre de caractère par commentaire en distinguant, par couleurs, la note du commentaire en question.

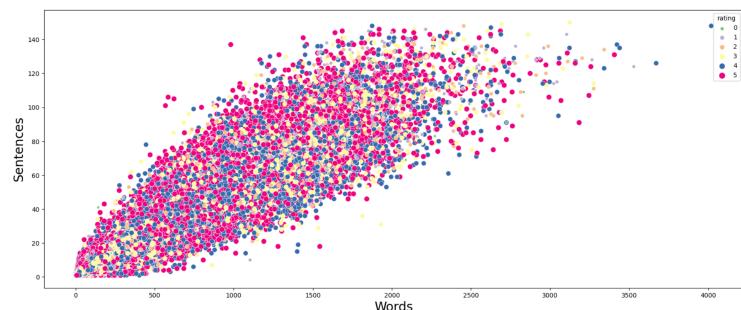
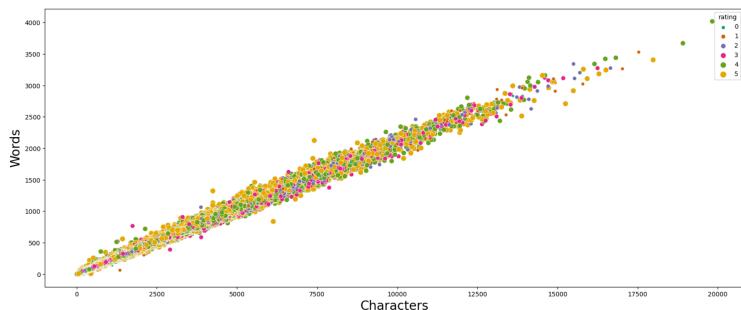
Si on prend l'exemple du morceau entouré sur la figure ci dessous, cela correspond au nombre de commentaires (qui se situe entre 26000 et 47500 environ) qui ont été noté à 4 et dont le nombre de caractères est entre 0 et 133 environ.



Nous avons suivi la même logique pour déterminer le nombre de commentaires par rapport au nombre de mots puis de phrases par commentaire toujours en distinguant la note par les couleurs.

Les 3 diagrammes suivent à peu près la même dynamique qui est la suivante : le nombre de commentaires diminue lorsque le nombre de caractères, de mots et de phrases augmente. Les utilisateurs émettent en général des commentaires courts.

Nous avons rajouté également deux autres graphes correspondant au nombre de caractères par mot et au nombre de mots par phrases et nous avons représenté les notes par des points de couleur. Nous pouvons voir clairement que la note ne dépend pas du nombre de caractères par mot ou du nombre de mots par phrase.



Suite à cette analyse, nous avons décidé de calculer la fréquence des mots en utilisant la fonction **TfidfVectorizer** qui attribue à chaque mot une valeur numérique afin de faciliter la compréhension à la machine lors du training (grâce à l'utilisation des méthodes **.fit** et **.transform**).

Nous avons créé un modèle de régression logistique et avons itéré sur 50000 features.

EXECUTION DU CODE :

Pour exécuter le code, il faut :

1_ Ouvrir un terminal

2_ Se placer dans le répertoire contenant les datasets ainsi que le fichier source du code

3_ Lancer jupyter par la commande : jupyter notebook (on suppose que celui-ci est correctement installé sur le poste)

4_ Modifier les chemin et mettre celui qui mène vers les datasets sur votre poste

```
sub = pd.read_csv("/Users/Nadya/Desktop/ESGI/S1/Deep Learning/Concours/goodreads_sample_submission.csv")
#reading the 2 data files ( The Training set and the testing set)
dftr = pd.read_csv("/Users/Nadya/Desktop/ESGI/S1/Deep Learning/Concours/goodreads_train.csv")
dfte = pd.read_csv("/Users/Nadya/Desktop/ESGI/S1/Deep Learning/Concours/goodreads_test.csv")
```

5_ Run le fichier en entier et le résultat sera transcrit dans le fichier **submission_file.csv**

CLASSEMENT KAGGLE :

≡ kaggle

+ Create

Home

Competitions

Datasets

Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Goodreads Books Revl...

NLP: Spacy Basics

Logistic regression trai...

Pyspark Tutorial

Predict Closed Questio...

View Active Events

Search

Overview Data Code Discussion Leaderboard Rules Team Submissions Submit Predictions ...

Rank	User	Profile	Score	Submissions	Time
56	Nimenides		0.53741	11	1mo
57	Idontknowwhatimdoing		0.53707	1	17d
58	Zeyad Abdelreheem		0.53668	1	6mo
59	Heera Lal		0.53630	1	2mo
60	didou1		0.53297	1	2h
61	Ignacio Velasco #2		0.53239	4	4mo
62	FMinarini		0.53114	1	3mo
63	Nadejda Dorosenco		0.52866	4	1s
Your Best Entry! Your most recent submission scored 0.52866, which is the same as your previous score. Keep trying!					
64	Syed Asim Ali Shah		0.52859	2	7d
65	Tevfik Can Özay		0.52650	2	2mo
66	Black31		0.52407	1	1mo
67	RaphP6		0.51436	1	1d