Exercice 2

Car Features

Includes features such as make, model, year, and engine type

Dans cet exercice, nous allons générer une intégration de mots avec **Word2Vec** à l'aide d'un exemple concret. L'ensemble des données que nous allons utiliser pour ce didacticiel provient de l'ensemble des données Kaggle.

Cet ensemble de données « **Véhicules** » comprend des fonctionnalités telles que **la marque**, **le modèle, l'année**, **le moteur** et d'autres propriétés de la voiture. Nous utiliserons ces fonctionnalités pour générer les incorporations des mots pour chaque modèle de marque, puis comparer les similitudes entre les différents modèles de marque.

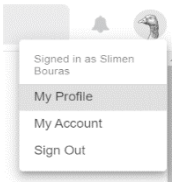
	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg	Popularity
)and output; double click to hide output															
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	28	19	3916
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertible	28	18	3916

A. Importer l'ensemble des données

Pour importer les données, il faut suivre instructions suivantes :

a) Aller à https://www.kaggle.com/

b) Aller à **My profile** (NP : **Si vous** n'**avez pas** encore de **compte**, **vous** devez en **créer** un)

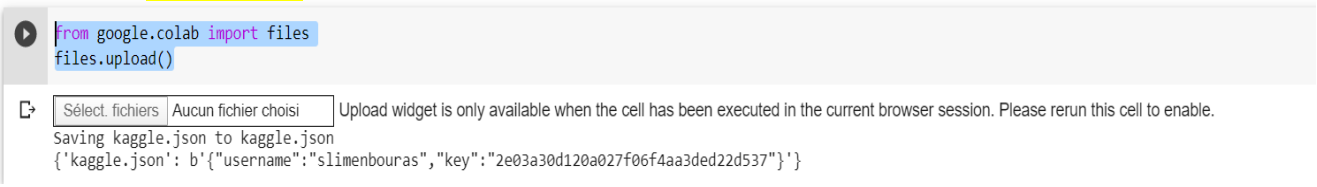


c) Aller à « API » et cliquer sur « Create New API Token» pour avoir votre fichier json

d) Aller à google colab et créer un nouveau Notebook

e) Pour importer le dataset dans notre Notebook, commencer par exécuter cette commande puis cliquer « Select fichiers » et importer votre fichier json

```
from google.colab import files
files.upload()
```



f) Créer le dossier dans votre Notebook

```
!mkdir -p ~/.kaggle
!cpkaggle.json ~/.kaggle/
!chmod 600 ~/.kaggle/kaggle.json
!ls ~/.kaggle
```

g) Exécuter votre fichier json

```
!ls -l ~/.kaggle
!cat ~/.kaggle/kaggle.json
```

h) Installer Kaggle dans google colab

```
!pip install -q kaggle
!pip install -q kaggle-cli
```

i) Importer le fichier avec votre **Username** et votre propre **KEY**

```
import os
os.environ['KAGGLE_USERNAME'] = "slimenbouras" # username from the json file
os.environ['KAGGLE_KEY'] = "2e03a30d120a027f06f4aa3ded22d537" # key from the json file
!kaggle datasets download -d CooperUnion/cardataset # api copied from Kaggle
```

j) Décompresser votre Dataset en CSV

```
if not os.path.exists('./cardataset/'):

!unzip cardataset.zip
```

Il faut avoir un résultat comme celui-ci

B. Incorporation des mots

Le word2vec exige un format de « liste des listes » pour la formation où chaque document est contenu dans une liste et chaque liste contient des listes de jetons de ce document. Dans un premier temps, nous devons générer un format de « liste des listes » pour l'apprentissage de l'incorporation de mots de modèle de fabrication. Pour être plus précis, chaque modèle de marque est contenu dans une liste et chaque liste contient des listes de fonctionnalités de ce modèle de marque.

1. Créer une nouvelle colonne « MakeModel » où elle combine la colonne « Make » et la colonne « Model »
2. Générer un format de « liste des listes » pour chaque modèle de marque avec les caractéristiques suivantes : Engine Fuel Type,Transmission Type, Driven_Wheels, MarketCategory, Vehicle Size, Vehicle Style, Maker_Model.

a) Sélectionner les entités du jeu de données d'origine pour former une nouvelle trame de données
b) Pour chaque ligne, combiner toutes les colonnes en une seule colonne
c) Créer la liste des formats de liste du corpus personnalisé pour la modélisation gensim
d) Afficher les 2 premiers exemples de cette liste

3. Entraîner le modèle word2vec avec notre propre corpus personnalisé avec min_count=1
4. Chercher le vecteur similaire de la voiture « 'Toyota Camry' »
5. Chercher la similarité entre la voiture **'Porsche 718 Cayman'** et la voiture **'Nissan Van'**
6. Chercher les Top5 voitures similaires à **Mercedes-Benz SLK-Class**