DATA SCIENCE 2021: Analyse des application de Google play store

LARBI Melissa [1] BENSAFIA Chems Eddine [2]

Introduction

Ce projet a été réalisé dans le cadre d'UE Data Science à Sorbonne université. Dans cet article, on résume et différentes tâches ainsi que les résultats obtenus.

Le projet consiste à analyser un dataset contenant des informations sur les applications Google Play Store, extraire des problématiques et proposer des solutions à l'aide des algorithme d'apprentissage supervisé et non supervisé.

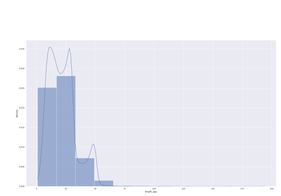
Problématique 1: Analyse du dataset Google Play Store

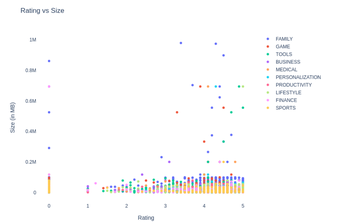
Afin de mieux comprendre les données qu'on a dans le dataset, la première tâche qu'on s'est fixé était d'explorer au maximum le dataset.

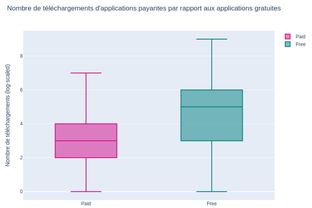
Les tâches réalisées dans cette partie sont :

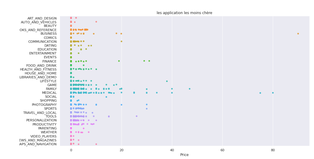
1. Le téléchargement des données.
2. Data cleaning
3. Data Visualisation
4. Featurs engenering

Voici quelques visualisations









Informations pertinentes

- Les utilisateurs préfèrent **payer** pour des applications **légères**
- La plupart des applications **les mieux notées** ont une **taille** optimale comprise **entre** 2 et 40 Mo
- La plupart des applications **les mieux notées** ont un **prix** optimal compris **entre** $1 et 30$
- Les applications **médicales et familiales** sont les plus **chères** et vont jusqu'à 80 Dolars
- Les utilisateurs ont tendance à textbftélécharger davantage une application donnée si elle a été textbfévaluée par un grand nombre de personnes.
- Les applications **gratuites** sont beaucoup plus **téléchargées** que les payantes
- Les utilisateurs sont plus **sévères** lorsqu'ils **évaluent** des applications **gratuites**

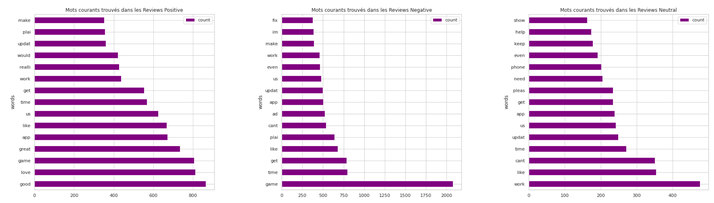
Problématique 2 : Classification selon les reviews

Dans ce problème supervisé on cherche à classer les reviews selon est-ce qu'elles sont positves, negatives ou neutre. Au premir lieu on opte pour une classification binaire (positf VS negatif) et puis pour une classificaiton triple(positif VS negatif VS neutre).

Pour ce fait on a procéder ainsi:

1. Transformer les reviews en listes des mots
2. Supprimer la ponctuation des mots.
3. Supprimer les terminaisons.
4. Supprimer les Stop Words (the, a ...)
5. Calculer les occurences des mots restant pour former un vocabulaire pour entraîner les modèles.
6. Entraîner les modèles et puis l'évaluer

Un aperçu sur le vocabulaire



Les résultas de la classification

Classification binaire										
	CountVecctorier					TFIDFVecctorier				
Classifier	Accuracy	Precision	Rappel	F1	Time	Accuracy	Precision	Rappel	F1	Time
Random	50%	0.50	0.58	0.54	0.02s	52%	0.51	0.61	0.56	0.28s
KNN 5	74%	0.69	0.85	0.76	1.52s	71%	0.66	0.84	0.74	2.48s
Perceptron	87%	0.89	0.83	0.86	0.37s	88%	0.89	0.87	0.88	0.54s
Perceptron Biais	90%	0.91	0.88	0.89	0.71s	87%	0.91	0.81	0.86	1.72s
Adaline	87%	0.89	0.83	0.86	0.44s	87%	0.49	1.00	0.66	1.16s
Perceptron Kernel	90%	0.90	0.89	0.89	0.53s	88%	0.88	0.91	0.89	0.69s

Table 1. Le résulat de la classification binaire

Multi classification				
	CountVecctorier		TFIDFVecctorier	
Classifier	Accuracy	Time	Accuracy	Time
Random	34%	1.47s	34%	0.33s
KNN 5	60%	11.66s	61%	15.32s
Perceptron	80%	1.28s	81%	1.51s
Perceptron Biais	85%	2.74s	82%	2.76s
Adaline	79%	1.70s	79%	12.29s

Table 2. Le résulat de la classification triple

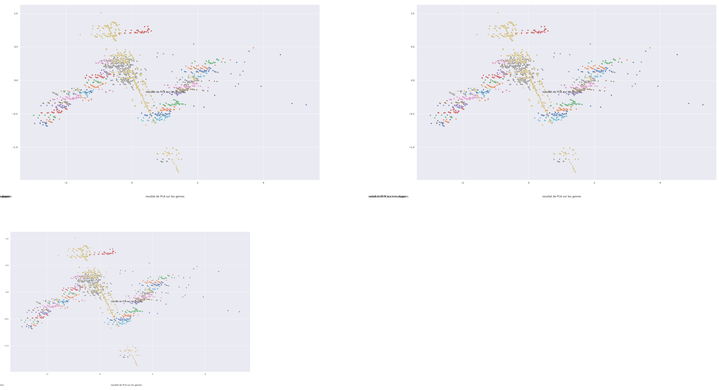
Problématique 03: Clustering

Le but de cette partie était de trouver les groupes dans notre dataset qui regroupe les caractéristiques des applications de Google Play Store (clustering).

Les tâches réalisées:

1. Récupérer le traitement effectué sur les données dans la partie : Analyse du dataset Google Play Store.
2. Appliquer l'algorithme PCA pour réduire la dimension.
3. Visualiser les données après réduction de la dimension.
4. Appliquer l'algorithme Kmean
5. Visualiser le résultat du clustering.

Visualisation des données après PCA



Resulat du clustering