Data Mining Application Play Store

Canete Valentin, Perrin Thibaud, Tisserand Jean-philippe, Bruyat Julian, Dailler Arnaud

**Introduction :**

Le but de ce rapport est d’étudier les caractéristiques des applications du Play Store pour connaître les types d’applications qui réussissent le mieux sur le Play Store et de faire une estimation des revenues des applications de ces différents types pour comparer leurs rendements. Les résultats seront vérifiés par une méthode de clustering (K-mean/ DBScan), L’étude sera donc portée sur les corrélations entre les différentes caractéristiques d’une application. L’utilisation d’une base de données de commentaire de certaines applications permettra de comparer l’avis d’une personne sur une application et sa réussite sur le Play store.

Mots clés: Data mining, Clustering, Play-Store, Correlation

**Données :**

Les données sont récupérées du fichier Google-Play Store-Full.csv. Ce fichier contient 267 053 applications et leurs caractéristiques (App Name : nom de l’application, Catégories, Reviews : nombres de notes, Installs : nombre de téléchargement, Size : taille de l’application en Mo, Price : prix en dollars, Content Rating : public visé, Last Updated : date dernière Maj, Minimum version : version minimum, Latest version : dernière version) . Nous avons décidé de ne pas utiliser les versions pour l’analyse.

Traitement des données : Avant de commencer toute analyse il a fallu mettre en place une phase de traitement des données : Rassembler certaines catégories, Convertir type de donnée, Supprimer les missing value (workflow 1 : Annotation “Traitement des données”)

Beaucoup d’applications portent la mention “Varies with device”. Pour gagner un nombre d’applications suffisantes, nous avons décidés de ne pas convertir cette colonne. A la fin du traitement des données il reste 267 051 Applications.

Pour la répartition du nombre d’applications par catégories on obtient

Une image contenant capture d’écran

Description générée automatiquement

Nous avons rassemblé tous les types de jeu au sein d’une même catégorie “GAME”, le détail de la répartition des applications par type de jeu est le suivant :

Une image contenant capture d’écran

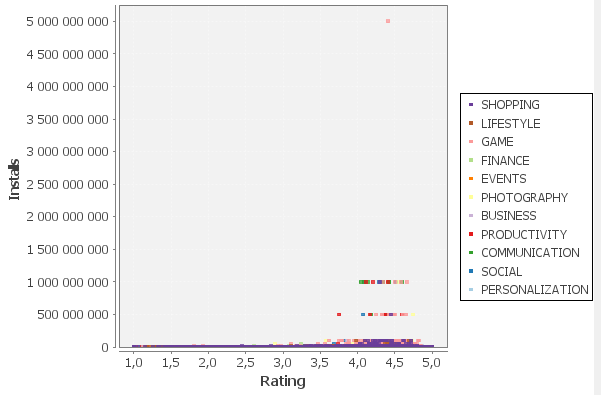
Description générée automatiquement

**Analyse des caractéristiques :**

Le nombres de téléchargements et la note d’une application sont deux paramètres importants pour indiquer la réussite sur l’App Store, dans un premier temps nous avons cherché des corrélations entre les paramètres souhaités (note, nombre de téléchargement) et les paramètres influent (taille, note, prix, public visé, catégorie) afin d’extraire les paramètres qui peuvent faire d’une application, une bonne application.

Nous présentons un scatter plot (coloré) pour montrer la répartition des applications en fonction de leurs groupes :

Etude des catégories sur le nombre de téléchargements :



Comme on peut le voir, les catégories n’ont pas réellement d’influence sur les paramètres de réussite d’une application dans le Play store, cependant on peut commencer à voir certaines tendance (jeu plus souvent téléchargé, etc..) mais on peut également voir qu’une application qui n’a pas une note inférieure à 3 ne dépassera pas les 100 000 000 de téléchargements.

Etude des poids sur le nombre de téléchargement