Passwords (2020-01-14)

BOURGEOIS Gabriel, HABERT Thomas

26 avril 2022

Résumé

L'objectif de ce projet est d'étudier le jeu de données *"Password"* en appliquant les méthodes étudiées en cours. Dans ce compte-rendu de semi-projet nous présenterons nos premières étapes de recherche ainsi que les premiers résultats et observations obtenus.

Notre jeu de données est constitué des 500 mots de passe les plus utilisés, tirés de l'article "Information is Beautiful".

1 Introduction

La base de donnée contient une unique table contenant 500 individus et 9 variables. Les variables sont les suivantes : *rank* (le classement du rang par ordre d'utilisation), *password* (le mot de passe), *category* (catégorie des mots de passe en fonction du thème de ceux-ci), *value* (le temps nécessaire au crack "en ligne"), *time_unit* (l'unité de temps de la colonne précédente), *offline_crack_sec* (le temps en seconde pour craquer hors-ligne), *rank_alt* (un autre classement par popularité), *strength* (la robustesse d'un mot de passe obtenue en fonction de son appartenance à des classes spécifiées dans la documentation) et enfin la *font_size*.

Au vu de ces éléments, la finalité de notre projet pourrait être de comprendre et d'expliquer les différents mécanismes qui influencent la robustesse d'un mot de passe ainsi que le temps nécessaire à son craquage. On pourra également étudier l'impact des catégories de password dans notre analyse.

2 Ouverture et nettoyage des données

Les données présentent dans ce fichier sont globalement assez propres. Nous nous sommes débarassés des quelques valeurs NA pour au final bien obtenir 500 individus. Nous avons choisis de ne pas prendre la variable rank comme index car celle-ci peut éventuellement nous être utile dans la visuation des données.

Les données sont exclusivement soit de type float64 soit de type category. Nous avons définies les types des variables catégorielles "category" et "time_unit" en reprenant les valeurs du tableau de manière non-ordonnées.

Enfin nous avons ajoutées 5 colonnes qui pourrait nous donner des informations utiles pour l'analyse des données :

- *length* : la longueur du mot de passe ;
- *nb_alpha* : le nombre de caractères alphabétiques dans le mot de passe ;
- *nb_num* : le nombre de caractères numériques dans le mot de passe ;
- *ratio_melange* : le ratio obtenu en divisant nb_alpha par nb_num ;
- *log_offline_crack_sec* : la variable obtenue en appliquant log à la colonne *offline_crack_sec* afin d'obtenir des valeurs plus digestes.

3 Première exploration simple

1. L'influence des catégories

En traçant la distribution des catégories, on observe que ceux contenants des noms sont largement les plus nombreux. Les mots de passes à propos de nourriture ou les mots de passe grossiers sont quant à eux très peu choisis. (cf Annexe figure 3)

Concernant la longueur, les mots de passe se regroupent autours de 6 à 7 caractères peu importe la catégorie, à l'exception des mots de passes constitués uniquement de lettres et des mots de passes dits "fluffy" (=mignon) qui en moyenne sont un peu plus courts (5 à 6 caractères). (cf Annexe figure 4)

Les mots de passe à caractères "nerdy-pop" sont en moyenne largement plus robustes que les autres catégories dont les robustesses sont assez homogènes. On remarque que les mots de passe à propos de nourriture

sont les moins bien classés en terme de robustesse. (cf Annexe figure 5)

A première vue, on ne remarque pas de corrélations particulières entre la distribution des catégories, la longueur des mots de passe des catégories et la robustesse des mots de passe des catégories.

2. L'influence de la longueur du mot de passe sur le temps de crackage (cf. Annexe figure 6)

Le temps du crackage augmente linéairement avec la longueur du mot de passe. Les valeurs suivant la courbe de tendance.

Ainsi plus les mots de passe sont longs plus ils demandent du temps afin d'être craqués.

3. Recherche de corrélations entre les variables (cf. Annexe figure 7)

Nous avons ensuite réalisé un heatmap des différentes variables.

On observe tout d'abord que les mots de passes constitués uniquement de lettres et les mots de passe constitué uniquement de chiffres sont très fortement décorellés. Cela signifie que parmis les 500 premiers mots de passes, beaucoup d'entre eux ne mélangent pas ces différents types de symboles.

Comme vu précédemment, la longueur des mots de passe est correlée au temps de crackage hors ligne.

Le temps de crackage est correlée positivement avec nb_alpha mais négativement avec nb_num : on en déduit que plus un mot de passe est composé de chiffres plus il se craque facilement, et plus un mot de passe est composé de lettres plus il se craque difficilement. Les chiffres semblent plus simples à cracker que les lettres.

4 Analyse en composantes principales

Nous avons enfin réalisé une ACP sur les valeurs afin d'obtenir une meilleur visualisation des données. On représente l'inertie expliquée en fonction des axes (cf. figure 1)

On remarque que l'axe 1 est largement premier en terme d'inertie. Celà peut provoquer un "effet taille" qui nous empêche de bien visualiser l'ensemble des données. Pour cela nous avons essayé de réaliser une APC à l'aide de grandeurs relatives, mais celà a eu pour résultat une encore plus grande inégalité d'inertie, donc nous sommes revenues à nos grandeurs absolues.

Ci-dessous nous avons représenté le premier plan factoriel de notre ACP (cf. figure 2).

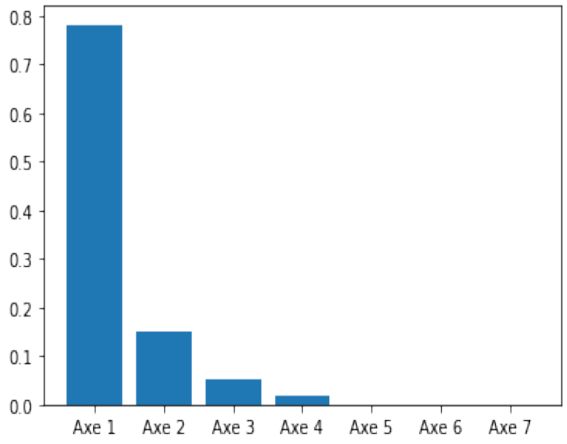

FIGURE 1 – Inertie expliquée en fonction des axes.

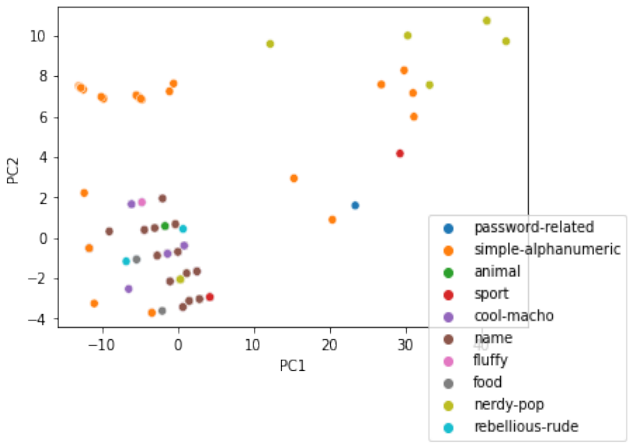

FIGURE 2 – ACP : Premier plan fractionnel.

On n'observe pas de groupe de même catégorie. Pour obtenir des résultats plus satisfaisant on peut colorer les différents points selon une variable différente.

5 Conclusion

Pour conclure, cette première études du jeu de données nous a permis d'effectuer un premier traitement pour nettoyer et compléter notre table. Nous avons également pu commencer à mettre en valeur quelques relations de corrélations entre les différentes variables.

Pour poursuivre notre étude, nous pourrions réaliser un clustering afin d'identifier par exemple des éventuelles similarités entre les catégories.

6 Annexe

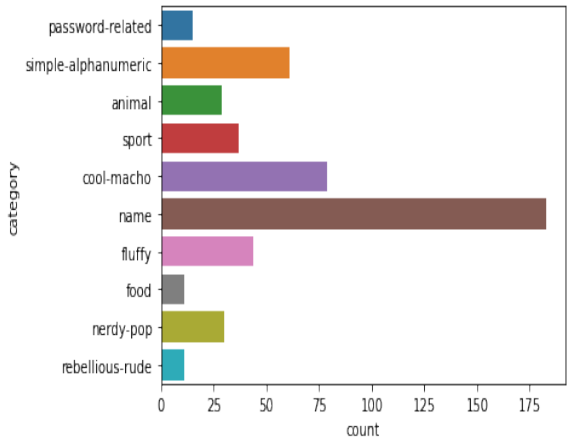

FIGURE 3 – Distribution des catégories.

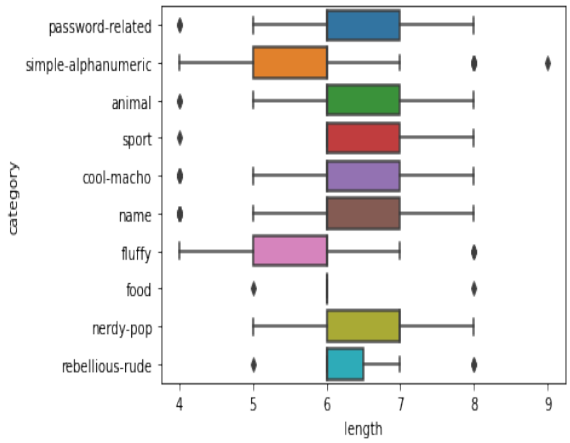

FIGURE 4 – L'influence des catégories sur la longueur des passwords.

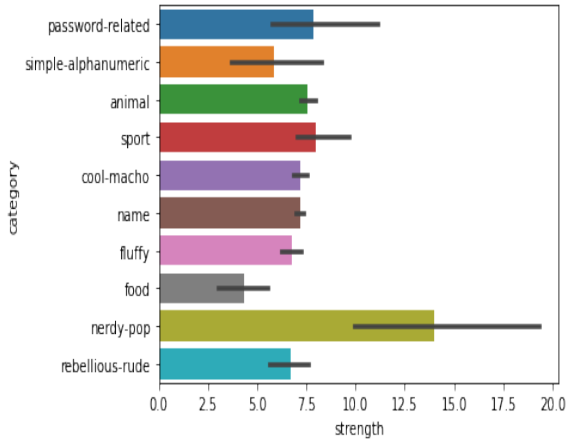

FIGURE 5 – L'influence des catégories sur la robustesse des passwords.

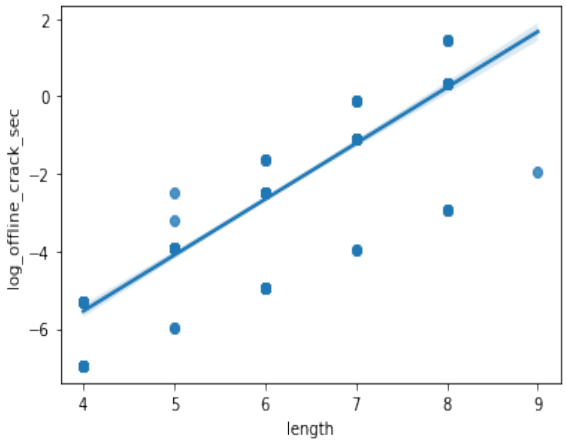

FIGURE 6 – Distribution des catégories.

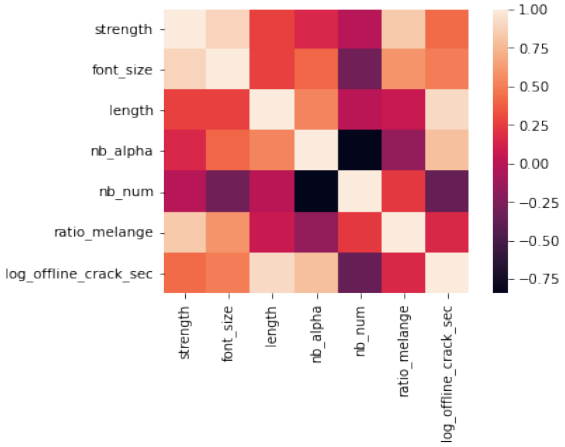

FIGURE 7 – Heatamp des corrélations entre variables