Markov Classifier

Introduction

This is a research in the password recovery methods and its main goal is to create an AI model that will optimize brute forcing passwords. Unfortunately it is not quite there yet, so for now it classifies passwords created from Markov Chains. The main idea of the project is actually pretty simple, one weak/really fast generative model and one classification model, which is balanced for performance and speed, might work more efficient compared to only one generative model. There is a little bit of previous research on that topic for example the project PASSGAN link: <https://github.com/d4ichi/PassGAN> uses a neural network as a generative model and a classifier to train the generative model, which is not quite the same approach.

My approach

For now, for a generative model I am using Markov chains trained on this wordlist <https://drive.google.com/file/d/106eHCTga2_s8nOV8u5Z78VMTuqdBFltf/view?usp=sharing>

And the idea for training the classifier is feeding it real passwords and generated ones, then giving it generated ones for prediction and those which are classified as real, are much more likely to be used in the real world. The function for generating Markov chains passwords:

Картина, която съдържа текст, екранна снимка, документ, Шрифт

Описанието е генерирано автоматично

Training the Model:  
To train the model with good hyperparameters I used hyperopt and a custom metric. To get as possible as it gets to real world passwords I used a real wordlist to define the true positives. The wordlist can be found on this link: <https://www.weakpass.com/> and the wordlist is called hashmob.net.medium.found.txt the model is trained with 400000 real passwords with length equal or above 8 from rockyou.txt and 400000 generated ones. The wordlist used to generate passwords is a mixture of rockyou.txt and xató-top-1million and the link for it is in the previous page. The data I trained the model with can be found in: <https://drive.google.com/file/d/1wBCVdsaQqDCb8jkC1k6t9Zv7Y_2Ic5Fp/view?usp=sharing>

And the pipeline and parameters for training the model are these: Картина, която съдържа текст, екранна снимка, документ

Описанието е генерирано автоматично

Custom Features

Картина, която съдържа текст, Шрифт, екранна снимка

Описанието е генерирано автоматично

This is the function that generates custom features from the passwords and returns a sparse matrix to work with hashing vectorizer. I used Hashing Vectorizer, with bigrams because for such classifier speed will be of great significance. And the number of ngrams were chosen by hyperopt.

Conclusion

This is the foundation for creatign a model that will aim to enhance brute forcing techniques in password recovery. The code classifier in here can be very easily implemented with a better generative model and be the core of an AI that creates wordlist trained from an already existing one. The versions of the dependencies used: Картина, която съдържа текст, екранна снимка, Шрифт

Описанието е генерирано автоматично