Probabilistic Context-Free Grammars Based Honeyword Generation

Probabilistic Context-Free Grammars For Passwords

The base structure of passwords is defined as the sequence of letter, digits and symbols. For example, password *(967abcde can be see as $S_2D_3L_5$, Where S_2 represents 2 consecutive symbols, following by D_3 3 consecutive digits and 5 consecutive letters. Therefore, a context-free grammar $S \rightarrow S_2D_3L_5$

With multiple passwords, we can learn probabilistic context-free grammars to fit the training data. For example, we have 1234abcd and qwer as training passwords. The corresponding probabilistic context-free grammar is

$$S \rightarrow D_4L_4 \qquad P = 0.5$$

$$S \rightarrow L_4 \qquad P = 0.5$$

$$L_4 \rightarrow abcd \qquad P = 0.5$$

$$L_4 \rightarrow qwer \qquad P = 0.5$$

$$D_4 \rightarrow 1234 \qquad P = 1.0$$

Honeyword Generation

1. Learns password PCFG model from RockYou passwords and input password files.
2. For each password, genrates 10% × n tough nut honeyword and 90% × n PCFG generated honeyword.