Sistema Operacionais: Trabalho Extra SysCall Tracer
Alan Portela Bandeira
02/06/2017

(b) A entrada para esta etapa do trabalho são arquivos texto, um ou mais para cada aplicação alvo, com as sequências de system calls geradas por execuções de cada aplicação. A tarefa deste item consiste nos dois seguintes passos:

(I) Gerar um arquivo texto com todas as subsequências da sequência principal, de tamanho N. Adotaremos N = 10. Isso é feito por um algoritmo simples de janela deslizante, com janela de tamanho 10. Cada posição da janela sobre a sequência original, gera uma subsequência. Para fins de entendimento, note que um arquivo texto de tamanho 100 geraria 91 subsequências de tamanho 10. Na literatura estas subsequências são chamadas n-gram. Obs.: De fato o método requer gerar todas as n-gram de tamanho 1 até N. Faremos isso mais a frente.

(II) Escrever um procedimento que toma como entrada dois arquivos texto de n-grams (subsequências), de duas versões da mesma aplicação, uma sadia e outra contaminada, ou de duas aplicações diferentes, e calcula um escore de diferença entre elas contando quantas subsequências estão presentes na sequência de teste (versão contaminada) que não estão presentes na versão normal. Esse valor pode ser apresentado em percentual dividindo pelo total de subsequências na sequência de teste. Quanto maior esse escore maior é a probabilidade da aplicação de teste está contaminada. Quando esse valor ultrapassa um certo limiar, um alerta de contaminação é gerado.

O arquivo “.gram” contendo todos os grams foi gerado a partir dos dados dos arquivos “.idx” que contem os syscalls e seus id’s. Ele é construído no modelo csv e cada linha contem uma sequência de 10 id’s, representando os syscalls presentes em um ngram.

- Trecho do arquivo ls_syscall.gram

0,1,2,3,4,5,6,3,7,4
1,2,3,4,5,6,3,7,4,5
2,3,4,5,6,3,7,4,5,5
3,4,5,6,3,7,4,5,5,8
4,5,6,3,7,4,5,5,8,5
5,6,3,7,4,5,5,8,5,6
6,3,7,4,5,5,8,5,6,3
3,7,4,5,5,8,5,6,3,7
7,4,5,5,8,5,6,3,7,4
4,5,5,8,5,6,3,7,4,5

O arquivo para teste, “test.gram” foi gerado como uma cópia do arquivo “ls_syscall.gram” tendo 6 de suas linhas substituídas por um ngram do tipo “1,1,1,1,1,1,1,1,1,1”, que representa a contaminção nas seqências de chamadas. A partir destes dois arquivos o score da semelhança é claculado. Foram usados 6 entradas contamindas para testes pois o arquivo de chamadas de sistemas tem 67 syscalls e portanto 58 ngrams de tamanho 10 e com 6 entradas contamindas poderíamos verificar um score de aproximadamente 0,1 ou 10%. O score final foi 0.10344827586206896.

Todos os arquivos gerados e o script feito em python, comentado encontram-se na pasta.