HOMEWORK 2

PROCESSING IRC LOGS

OVERVIEW

The purpose of this homework is to analyze textual log data from an online chat forum related to the Anonymous hacktivist group. You will learn how to apply regular expressions, summarize log data, quantify text data, and summarize time trends.

DATA

IRC is an early protocol for instant messaging developed in the early years of the Internet. The openness and ability to remain anonymous has made IRC a popular channel for hacker networks to collaborate and share ideas.

The data comes from https://www.azsecure-data.org/internet-relay-chat.html. It contains two years of chats between hackers associated with the hacktivist group Anonymous. In these logs they share information about malware, setting up servers to deploy attacks, and other information related to hacking systems.

The collection and analysis of these chats is a form of cyber-threat intelligence. The analysis of these chats and other dark web data sources enable proactive defense against attacks.

ANALYSIS

1. Many users log in and view the chat without commenting. Which users spent the most time in the logs? (3pts) Which users logged in the most (2pts)
2. Find the most common words (3 pts)
3. Count the total number of written messages (only those with actual text content) (2 pts). Summarize the users that posted the most messages (2pts)
4. Find and rank (by count) words not in an English dictionary (3 pts). This is a simple method that can identify some names of malware tools
5. Which hours of the day had the most messages (2pts)? Which days had the most traffic (or messages) (2pts)?
6. Find and list the URLs posted in the chat. (2pts)

GRADING

This analysis portion of the assignment is graded out of 10 points. The maximum score for analysis is 15 points.

Your code should also be well-documented with comments, sources, and explanations of what is happening. Fully documented code will receive full credit. Mostly complete documentation will receive a deduction of a point, minimal documentation will result in a deduction of 2 points, and no documentation will result in a deduction of 3 points from your score.

SUBMISSION

Submit your code and accompanying documentation and evidence that your program works in a PDF or Word document. The instructor may wish to see a demonstration of your code.

TIPS

<+evilbot> This user is a bot. If possible, filter this user's posts from the chat

You can identify changes in days with the messages "--- Day changed Mon Sep 26 2016". There are some instances of this measure missing. It is possible to correct this issue by looking at the times of the day (i.e. the hour rolls over to 00).

Users can change their usernames. An alternative to usernames for login-logout behavior is to use their login identifiers (for example: [androirc@AN-l8e.7dp.8hdu3q.IP]).