Hotel TULIP Web Server Data Analysis

Assignment 2 - SIT742 Modern Data Science

Alex Cummaudo <ca@deakin.edu.au>

Jake Renzella <jake.renzella@deakin.edu.au>

Deakin Software and Technology Innovation Laboratory

School of Information Technology

Deakin University, Australia

May 3, 2017

Executive Summary

This report summarises findings from a data exploration on the Hotel TULIP web server logs, recorded between the periods of August 2014 and August 2015. Each log contains one *request*, or *hit*, that lists fourteen attributes as described in the attached Data Dictionary spreadsheet. Publicly known client IP addresses were extracted from the MaxMind GeoIP2[1] dataset to analyse the location of requests (narrowed down to city). Additionally, user agent strings were parsed to analyse device and browser statistics using the Python `user-agents` library[2], thereby extrapolating demographics, usage trends, platform information, server performance, and security statistics from the raw logs provided in the dataset. Further details on the extraction of the data is provided in the source code attached in Appendix B, and an interactive version of this file is published on Databricks.

[1] See `http://dev.maxmind.com/geoip/geoip2/`.

[2] See `https://pypi.python.org/pypi/user-agents`.

Contents

List of Figures

1 Key Findings

A list of key findings in the analysis are as thus:

- Foo

2 Introduction

3 Dataset

4 Method

4.1 Assumptions Made

4.2 Extraction Process

4.2.1 Internal/External IP Address Regular Expression

To differentiate between private site visitors and external visitors, a regular expression was used to filter private/public IP address ranges. The regular expression is shown below:

```
c_ip regexp
        '(^127\.)|(^10\.) |
        (^172\.1[6-9]\.) |
        (^172\.2[0-9]\.) |
        (^172\.3[0-1]\.) |
        (^192\.168\.)'
```

Using the not of the regular expression will select only public IP addresses.

4.3 Data Mining

5 Results

5.1 Patterns Identified

5.1.1 Internal Site Visitors

5.1.2 External Site Visitors

5.1.3 Hong Kong Visitors

5.1.4 USA Visitors

5.1.5 Australia Visitors

5.1.6 PC Visitors

5.1.7 Smartphone Visitors

5.1.8 Bots Visitors

5.1.9 Tablet Visitors

A Additional Figures

B Extrapolation Results

Attached on the following pages are the results from Databricks. You may also interact with this online on Databricks.