Hotel TULIP Web Server Data Analysis

Assignment 2 - SIT742 Modern Data Science

Alex Cummaudo `<ca@deakin.edu.au>`

Jake Renzella `<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 **??**, 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:

- A majority of requests are from Hong Kong, specifically the Central District (Section **??**).

- The number of requests made to the web server are increasing over time (Section **??**).

- Response times to the server are getting slower (Section **??**).

- There are a number of increasing server errors (Section **??**).

- The web server is most active at 9am UTC, with most requests made in March 2015 (Section **??**).

- iOS is being used more than Android (Section **??**).

- The server is generally secure with relatively few genuine repetitive unauthorised attempts (Section **??**).

- iPhone is being used more than iPad to request the server worldwide, but internally, guests use iPad more than iPhone (Section **??**)

- Windows was mostly used for requests made on PC devices (Section **??**).

- Apple devices are primarily used around the world for smartphone and tablet requests (Section **??**).

- Google is crawling at 16%, over Bing and Facebook at 10% and 4% (Section **??**).

2 Demographics

A Additional Figures