211Toronto.ca Web Session Analysis and Visualization

by

Yuhong Liu

Supervised by: Prof. Mariano P. Consens

A M.Eng Project Report submitted in conformity with the requirements
for the degree of Master of Engineering

Department of Mechanical and Industrial Engineering

University of Toronto

211Toronto.ca Web Session Analysis and Visualization

Yuhong Liu

Master of Engineering

Department of Mechanical and Industrial Engineering
University of Toronto

2016

Abstract

Findhelp Information Services (Findhelp) is a lead provider of information and referral services in Ontario and across Canada. Through websites like 211Toronto.ca or 2-1-1 helpline, they connect Ontarian to a complete range of government and community-based health and social services close to their communities. The paper discusses the process of analyzing, visualizing and extracting insights from web logs of 211Toronto.ca for a 24 month. The data preparation stage includes preliminary cleaning to remove irrelevant information, sessionize events, identify robot web crawler, match IP addresses for getting geo-location as well as parsing user agents. With cleaned session data set, the paper extracted useful insights and patterns about the website's session visits. In addition, the paper conducted statistical analysis on analyzing the behavioral difference between bot and non-bot, weekend and weekday, as well as mobile and non-mobile sessions. The results showed significant statistical evidences to support the difference in user behavior for these session categories. Moreover, the paper categorized 18 topics listed on the website to understand their popularity and conducted sequential pair association analysis through calculation of support, confidence and interests to investigate the relationship between each topic pairs.

Table of Contents

1. Introduction

Findhelp Information Services (Findhelp) is a lead provider of information and referral services in Ontario and across Canada. [1] They operates the community information 211Toronto.ca and referral helplines 2-1-1. It connects Ontarian to complete range of government and community-based health and social services close to their communities. 211Toronto.ca is one of the services offered by Findhelp. [2] It is a website contains information with over 20,000 community, health, social and related government services information. With more than 4000 visits on average per day, the website serves as a reliable source and alternate channel of community information for residents in Toronto.

The objective of the projects is to utilize the log data received from 211Toronto.ca, extract insights through data cleaning, ip address matching, user agent parsing, and statistical analysis in order to learn different session behaviors and patterns at different time of the day, day of the month, or month of the year. In addition, we would like to know the popularity of 18 topics listed on the website and their association with one another.

Throughout the project, we have first discussed with our client to understand the background information and purpose of the website before coming up with the set of project objectives. Next, after a period of literature reviews, we have identified Google's research paper, Session Viewer: Visual Exploratory Analysis of Web Session Logs [3] and Resul Das' paper on creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method [4] contains very useful and similar website log analysis methods for our project. Therefore, we have decided to replicate part of the visualizations and methods introduced in their paper. Then we have selected a set of data processing and statistical analysis software to cleaning, parse and analyze the given website log data. Using external libraries, we have obtained geo-location of IP address and parsed user agents to understand the platform, operating systems as well as browsers used. Last but not least, through rigorous statistical analysis and data visualizations using tools such as Tableau, Python and Minitab on cleaned session data set, the paper extracted useful insights and patterns about the website's session visits. Moreover, with the categorized 18 topics listed on the website, the paper investigated the popularity of these topics and conducted sequential pair association analysis through calculation of support, confidence and interests to investigate the relationship between each topic pairs.

2. Definitions

Event – one activity performed on the website by a user agent through a unique ip address. [4]

IP address - Internet Protocol Address (or IP Address) is a unique address that computing devices such as personal computers, tablets, and smartphones use to identify itself and communicate with other devices in the IP network. [7] Any device connected to the IP network must have a unique IP address within the network. An IP address is analogous to a street address or telephone number in that it is used to uniquely identify an entity.

User agent - When visiting a webpage, browser sends the user-agent string to the server hosting the site that you are visiting. [6] This string indicates which browser is used, its version number, and details about visitor's system, such as operating system and version. The web server can use this information to provide content that is tailored for visitor's specific browser, such as a mobile version of the website

Session – a series of events performed on the website by a single user agent through a unique ip address at specific period of time. [5] The number of user sessions on a site is used in measuring the amount of traffic a Web site gets. The site administrator determines what the time frame of a user session will be. In general, people use 30 minutes as a benchmark for one session. For our analysis, we have also used 30 minutes to group events into sessions. If the visitor comes back to the site within that time period, it is still considered one user session. However, If the visitor returns to the site after the allotted time period has expired or with inaction period for more than the benchmark (e.g. 30 minutes), then it is counted as a separate user session.

3. Project Objectives

Managers and developer of 211Toronto.ca are interested to understand the following questions based on the web log data provided:

- Identify user behaviors based on count of session visits monthly, daily and hourly
- Differentiate pattern for bot and non-bot, weekend and weekday sessions
- Distribution of website session visits and duration throughout a 24-month period
- Top cities and countries visited the website
- Identify top browser, devices and operating systems used to visit the website
- Visualize sessions visited the 18 topics categorized on the website
- Investigate paired relationship between listed topics on the website

4. Data Source and Schema

211Toronto.ca is currently hosted on a Windows web server, therefore its web logs are automatically generated by the Microsoft Internet Information Service (IIS), the version of IIS used for the website is 7.5 which was included in windows 7 and windows server 2008 R2. [8] IIS is a set of programs for building and administering web sites, a search engine, and support for

writing web-based applications that access databases. [9] It tightly integrated with the Windows servers resulting in faster web page serving.

Here is the data schema from the data received from 211Toronto.ca [10] :

Field Name	Field Code	Field Description
Date	date	Date of the web log
Time	time	Time of the webpage was accessed
Client IP Address	c-ip	The IP address of the client that made the request.
User Name	cs-username	The name of the authenticated user who accessed your server. Anonymous users are indicated by a hyphen.
Server IP Address	s-ip	The IP address of the server on which the log file entry was generated.
Server Port	s-port	The server port number that is configured for the service.
Method	cs-method	The requested action, for example, a GET method.
URI Stem	cs-uri-stem	The target of the action, for example, Default.htm.
URI Query	cs-uri-query	The query, if any, that the client was trying to perform. A Universal Resource Identifier (URI) query is necessary only for dynamic pages.
HTTP Status	sc-status	The HTTP status code.
Win62 Status	sc-win62-status	The Windows status code.
Time Taken	time-taken	The length of time that the action took, in milliseconds.
User Agent	cs(User-Agent)	The browser type that the client used.
Protocol Substatus	sc-substatus	The substatus error code.

5. Methodology

In this section, we will discuss the methodology we have used in the project. The project is divided into 6 main stages; they are data retrieval, data cleaning and data analysis. Most of the emphases are on the data cleaning and data analysis stage. Within data cleaning, we have done session grouping, bot identification, IP address matching, user agent parsing as well as removing any potential outliers. In the data analysis stage, we have used statistical analysis packages in python to obtain insights about those log session data and then the results are shared in the results section.

5.1 Tools and Architecture

This section describes the list if technologies we have used in this project in order to process over 13 GB of data from web logs. Below is the overall architecture of the project:

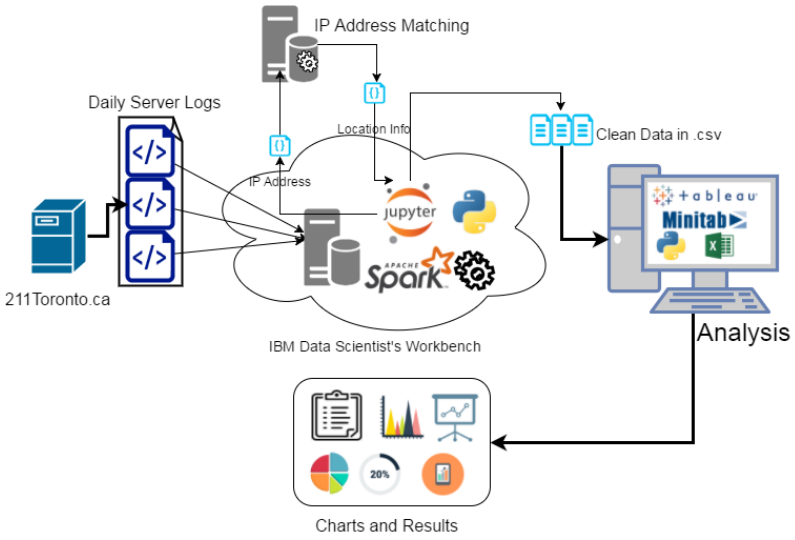


Figure 1 Project Architecture for Data Processing and Visualization

Most data analytics related projects use either Python or R as their main language for data cleaning and processing. Both language are very user friendly and have very good integration with Apache Spark which is our main data processing engine. However, we have picked Python as the main scripting language used in cleaning, processing and creating data outputs because it is one of the three languages used in Apache Spark's official documentation. We can easily utilize sample codes in the documentation without worry about syntax and indentation. In addition, python works better with other steps of the data processing such as IP address matching and user agent parsing.

We have chosen IBM's data scientist's workbench (DSWB) as our main platform for the project because this cloud-based work environment comes with locally installed Apache Spark, a fast and general engine for processing large scale data. DSWB also supports Jupyter notebook in python 2 for iterative workflow and immediate feedback. In addition, IBM has been very generous by assigning large amount of storage space for each user. We were able to upload large amount of log data to the platform, create python notebooks to process data, and save cleaned data to local folders on the server.

Apache Spark is used as our main engine for processing data. Initially, we tested Hortonwork's single node Hadoop VM. However, none of our computer is able to quickly process the large number of log data we received. Apache Spark enables us to conduct data processing of web logs a lot faster than Hadoop's MapReduce. Spark also has very good documentation in python and high level operators available to be called through its python API. The ability to use SQL is also a major deciding factor, because it is the main language for the community to conduct ETL of the data. Lastly, Spark is able to pass dataframe over to pandas. It allows us to conduct iterative querying for specific data in the dataframe.

Although Python and Apache Spark are great at processing, transforming and cleaning the data, it is very difficult to conduct data visualization or statistical analysis using them. The analysis of

data requires us to produce many charts and graphs; therefore, Tableau, Excel and Minitab are used to satisfy those requirements.

5.2 Data cleaning

In this section, we discuss the detail steps taken in order to clean and process the data. In order to obtain clean and aggregate data from weblogs, we went through steps in preliminary cleaning, session grouping, IP address matching and user agent parsing. The details of these steps are described below.

5.2.1 Preliminary Cleaning

Once we have retrieved the raw web logs from the server of 211Toronto.ca, they are unzipped and uploaded individually to the same folder of DSWB. A script (processdata.py) is run to conduct the preliminary cleaning of weblogs. Here are steps done by the script:

1. Remove irrelevant IIS header line in the weblog files
2. Retrieve column headers
3. Remove picture files, graphics, JavaScript, stylesheet files, etc. [13]
4. Save the cleaned data in .csv format
5. Append all cleaned log data into one master file

A typical raw weblog would look like as following:

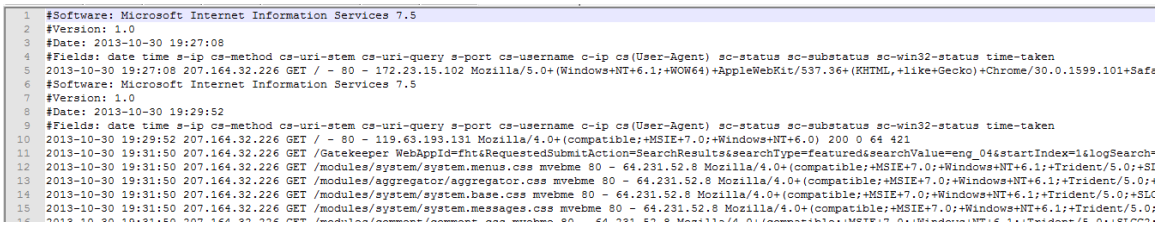


Figure 2 Sample Web Log Data

The top 8 lines contain Microsoft IIS' information which is irrelevant in our analysis. Therefore, we have removed them in our step 1. In step2, we have retrieved the column header in line 9 and store them as a array in python. Step 3 parses all the rest of data and removes any non-text files because they are not part of our project objectives and targets for analysis. [4] Here are the exact codes used and list of file extensions to be removed:

```
for line in log:
    if (not line.startswith('#')) \
    and (".css" not in line) \
    and (".js" not in line) \
    and (".png" not in line) \
    and (".jsp" not in line) \
    and ("/Gatekeeper" not in line) \
    and (".php" not in line) \
    and (".aspx" not in line) \
    and (".gif" not in line) \
    and (".jpg" not in line) \
    and (".ico" not in line) \
    and ("/system/ajax" not in line): \
    #and ("/robots.txt" not in line)
    #and (" / " not in line):
        output.writelines(line)
```

Figure 3 List of non-text file removed

5.2.2 Session Grouping

Session grouping is an exercise to group events based on a set of logics to differentiate visitors of the website throughout the day. [8] Most web log data contains only transactional data. It could tell us at a particular time, what page has been access by which user agent at certain ip address. However, by just looking at the data, we won't be able to identify what are pages view by a particular user and his/her user behavior. Therefore, session grouping needs to be conducted based on the useragent, ip address as well as the time gap between user's activities.

We have described session briefly in the definition section, it is a collection of events, from the same user, grouped together based on certain criteria. A session can often be viewed as the ordered list of a user's actions in completing a task. [8] A lot of analytics tools such as Adobe Webtrends or Google Analytics groups hits together based on activity. These analytics tool detects the activities of a user and when the user becomes inactive or closes his/her browser window, the session has considered to be ended. In general, many analytics tools consider 30 minutes of inaction as a benchmark for timeout and restarting a new session.

We are not given any 'user id' or 'machine id' that we can easily use as a identifier for a user session. Therefore, we have considered the combination of user agent and ip address will be the unique identifier for a user. For a sorted daily web log text file, we consider a new session has started when either ip address or user agent has changed when next event happened, or there is more than 30 minutes of inaction between the two events.

We have used HiveSQL's window function lag() in order to obtain the time difference between the previous event and the event after in a new column called 'previous_timestamp'. [14] Any time difference greater than 1800s (30 minutes) are marked in another new column called 'is_new_session'. Then the sum of 'is_new_session' column as 'id' will give us the number of sessions happened within each user agent and ip address combination. Lastly, concatenate the date, ip address and id column will give us the session id in column 'session_id'. Here are actual codes and data used to sessionize events:

Create previous_timestamp column:

```
In [9]: lag_timestamp = sqlContext.sql("SELECT *, \
              LAG(time) OVER (PARTITION BY date, cip, useragent ORDER BY date, cip, useragent, time) AS previous_timestamp\
            FROM weblogSessions")
        lag_timestamp.registerTempTable("weblogSessions2")

In [10]: lag_timestamp.show()
```

date	cip	csuristem	useragent	time	previous_timestamp
2015-06-01	173.33.177.15	/detail/en/203265	Mozilla/5.0+(Wind...	23:51:26	null
2015-06-01	180.76.15.5	/fr/detail/en/101283	Mozilla/5.0+(comp...	05:30:41	null
2015-06-01	180.76.15.5	/fr/detail/en/144944	Mozilla/5.0+(comp...	06:54:54	05:30:41
2015-06-01	180.76.15.5	/fr	Mozilla/5.0+(comp...	08:41:25	06:54:54
2015-06-01	180.76.15.5	/detail/en/168152	Mozilla/5.0+(comp...	11:04:33	08:41:25
2015-06-01	180.76.15.5	/fr/detail/en/144487	Mozilla/5.0+(comp...	16:24:32	11:04:33
2015-06-01	180.76.15.5	/detail/en/176514	Mozilla/5.0+(comp...	16:45:00	16:24:32
2015-06-01	180.76.15.5	/fr/detail/fr/144506	Mozilla/5.0+(comp...	18:56:06	16:45:00

Figure 4 Sample output for creating previous_timestamp column

Identify events with more than 30 minutes of inaction:

```
In [10]: get_ind = sqlContext.sql("SELECT *,\
                CASE WHEN unix_timestamp(time) - unix_timestamp(previous_timestamp) >= (60 * 30)\
                     OR previous_timestamp IS NULL\
                THEN 1 ELSE 0 END AS is_new_session\
            FROM weblogSessions2")
         get_ind.registerTempTable("weblogSessions3")
         get_ind.show()
```

```
+----------+-------------+--------------------+--------------------+--------+------------------+--------------+
|      date|          cip|           csuristem|           useragent|    time|previous_timestamp|is_new_session|
+----------+-------------+--------------------+--------------------+--------+------------------+--------------+
|2015-06-01|173.33.177.15|   /detail/en/203265|Mozilla/5.0+(Wind...|23:51:26|              null|             1|
|2015-06-01| 180.76.15.5|/fr/detail/en/101283|Mozilla/5.0+(comp...|05:30:41|              null|             1|
|2015-06-01| 180.76.15.5|/fr/detail/en/144944|Mozilla/5.0+(comp...|06:54:54|          05:30:41|             0|
|2015-06-01| 180.76.15.5|                 /fr|Mozilla/5.0+(comp...|08:41:25|          06:54:54|             0|
|2015-06-01| 180.76.15.5|   /detail/en/168152|Mozilla/5.0+(comp...|11:04:33|          08:41:25|             0|
|2015-06-01| 180.76.15.5|/fr/detail/en/144487|Mozilla/5.0+(comp...|16:24:32|          11:04:33|             0|
|2015-06-01| 180.76.15.5|   /detail/en/176514|Mozilla/5.0+(comp...|16:45:00|          16:24:32|             0|
|2015-06-01| 180.76.15.5|/fr/detail/fr/144506|Mozilla/5.0+(comp...|18:56:06|          16:45:00|             0|
```

Figure 5 Sample output for identifying two events with more than 30 minutes of inaction

Summarize is_new_session for each user agent and ip address:

```
In [11]: session1 = sqlContext.sql("SELECT *, \
                sum(is_new_session) OVER (PARTITION BY date, cip, useragent \
                ORDER BY date, cip, useragent, time) AS id\
             FROM weblogSessions3")
         session1.registerTempTable("weblogSessions4")
         session1.show()
```

```
+----------+-------------+--------------------+--------------------+--------+------------------+--------------+---+
|      date|          cip|           csuristem|           useragent|    time|previous_timestamp|is_new_session| id|
+----------+-------------+--------------------+--------------------+--------+------------------+--------------+---+
|2015-06-01|173.33.177.15|   /detail/en/203265|Mozilla/5.0+(Wind...|23:51:26|              null|             1|  1|
|2015-06-01| 180.76.15.5|/fr/detail/en/101283|Mozilla/5.0+(comp...|05:30:41|              null|             1|  1|
|2015-06-01| 180.76.15.5|/fr/detail/en/144944|Mozilla/5.0+(comp...|06:54:54|          05:30:41|             0|  1|
|2015-06-01| 180.76.15.5|                 /fr|Mozilla/5.0+(comp...|08:41:25|          06:54:54|             0|  1|
|2015-06-01| 180.76.15.5|   /detail/en/168152|Mozilla/5.0+(comp...|11:04:33|          08:41:25|             0|  1|
|2015-06-01| 180.76.15.5|/fr/detail/en/144487|Mozilla/5.0+(comp...|16:24:32|          11:04:33|             0|  1|
```

Figure 6 Aggregation of ids to create is_new_session column

Concatenate date, ip address and id as session id:

```
In [12]: session = sqlContext.sql("SELECT *, concat_ws('_',date,cip,id) as session_id \
               FROM weblogSessions4")
         session.registerTempTable("sessionized_log")
         session.show()
```

```
+----------+-------------+--------------------+--------------------+--------+------------------+--------------+---+--------------------+
|      date|          cip|           csuristem|           useragent|    time|previous_timestamp|is_new_session| id|          session_id|
+----------+-------------+--------------------+--------------------+--------+------------------+--------------+---+--------------------+
|2015-06-01|173.33.177.15|   /detail/en/203265|Mozilla/5.0+(Wind...|23:51:26|              null|             1|  1|2015-06-01_173.33...|
|2015-06-01| 180.76.15.5|/fr/detail/en/101283|Mozilla/5.0+(comp...|05:30:41|              null|             1|  1|2015-06-01_180.76...|
|2015-06-01| 180.76.15.5|/fr/detail/en/144944|Mozilla/5.0+(comp...|06:54:54|          05:30:41|             0|  1|2015-06-01_180.76...|
|2015-06-01| 180.76.15.5|                 /fr|Mozilla/5.0+(comp...|08:41:25|          06:54:54|             0|  1|2015-06-01_180.76...|
|2015-06-01| 180.76.15.5|   /detail/en/168152|Mozilla/5.0+(comp...|11:04:33|          08:41:25|             0|  1|2015-06-01_180.76...|
|2015-06-01| 180.76.15.5|/fr/detail/en/144487|Mozilla/5.0+(comp...|16:24:32|          11:04:33|             0|  1|2015-06-01_180.76...|
|2015-06-01| 180.76.15.5|   /detail/en/176514|Mozilla/5.0+(comp...|16:45:00|          16:24:32|             0|  1|2015-06-01_180.76...|
|2015-06-01| 180.76.15.5|/fr/detail/fr/144506|Mozilla/5.0+(comp...|18:56:06|          16:45:00|             0|  1|2015-06-01_1        |
```

Figure 7 Sessionized events

5.2.3 Bot identification

A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. [13] This process is called Web crawling or spidering. Many legitimate sites, in particular search engines, use spidering as a means of obtaining up-to-date data for indexing purpose.

We would expect sites like Google, Yahoo or Bing would use their bots to crawl the website's information for their search engine users. [14] We are not surprised to see a large volume of traffic for the website are referral from major search engines because majority of us might not heard of Findhelp or 211 services in Toronto.

Therefore, it makes sense for us to identify whether an event is conducted by a bot or an actual human being during the session grouping. There are many databases or libraries online that store any possible bot's IP address or user agents. In our case, the most convenient way of identifying bot is to use event's user agent and search in a library to identify whether it is a bot activity. [15] When a software agent operates in a network protocol, it often identifies itself, its application type, operating system, software vendor, or software revision, by submitting a characteristic identification string to its operating peer.

We have used the python library robot_detection [16], its method *is_robot* function's return value is True, then we assign 1 to the bot column for the event and identifies it has a bot activity. In addition, 211Toronto.ca web site owners use the /robots.txt file to give instructions about their site to web robots or spiders. Whenever a robot wants to visits a Web site URL, say /welcome.html. Before it does so, it firsts checks for /robots.txt. The file will tell the robot whether it is approved to visit certain pages of the website. This method is called The Robots Exclusion Protocol. Usually, human users will not be visiting the robots.txt file, only bots will be programmed to visit this file before any action on the website. Therefore, we could easily conclude that any user agent who visited the robots.txt file could be classified as a robot.

5.2.4 IP Geo Location

One of the objectives of this project is to understand whether the website is fulfilling its purpose of serving local resident of Toronto on providing local community and social information. Is there a lot of people from Toronto really using the website? Are people outside of Toronto also interested in website's information? These are important questions for the website's maintenance, improvement as well as contents available. All in all, we will need to find ways to identify the geo location of sessions with a granularity to city or community level.

The only item gives us possible geo location is the ip address associated with each event. Internet Protocol Address (or IP Address) is a unique address that computing devices such as personal computers, tablets, and smartphones use to identify itself and communicate with other devices in the IP network. Any device connected to the IP network must have an unique IP address within the network. An IP address is analogous to a street address or telephone number in that it is used to uniquely identify an entity. Because of these important property of IP address, we were able to use event's ip address to identify the geo location of website visitors.

In another script, we have extract all the unique ip addresses from raw data, and using an external HTTP request service, http://ip-api.com/batch [17], through a post request method in python, we were able to batch query 100 ip addresses at a time. However, the free service is limited to 150 query per minute. In total, there are approximately 110900 unique ip addresses.

Therefore, we have grouped 100 ip addresses together as a JSON query to be sent over to this web service and have the result stored in another JSON object. Three new fields are created in summary table, city, state, and country.

5.2.5 User Agent Parsing

User agent serves as a software agent and it acts on behalf of a user in computing. a user agent acts as a client in a network protocol used in communications within a client–server distributed computing system. In particular, the Hypertext Transfer Protocol (HTTP) identifies the client software originating the request, using a "User-Agent" header, even when the client is not operated by a user. The SIP protocol (based on HTTP) followed this usage. [18]

User agents in the raw data give us important information on the operating system, used browser, type of platform, as well as identifying whether the visitor is a human being or bot. We have already utilized user agent to identify bot activities in the session grouping stage. Now, we would like to understand more about how the visitor is accessing the website. Therefore, we have used an external python library, httpagentparser [19], to parse user agent for getting important visitor information.

A typical human user agent would look like:

Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_5) AppleWebKit/567.66 (KHTML, like Gecko) Chrome/68.0.2125.111 Safari/567.66

Through parsing this string, we were able to obtain the following information [20]:

Chrome 68.0.2125.111	
Mozilla	MozillaProductSlice. Claims to be a Mozilla based user agent, which is only true for Gecko browsers like Firefox and Netscape. For all other user agents it means 'Mozilla-compatible'. In modern browsers, this is only used for historical reasons. It has no real meaning anymore
5	Mozilla version
Macintosh	Platform
Intel Mac OS X 10_9_5	Operating System: OS X Version 10_9_5 : running on a Intel CPU
AppleWebKit	The Web Kit provides a set of core classes to display web content in windows
567.66	Web Kit build
KHTML	Open Source HTML layout engine developed by the KDE project
Description:	Free open urce web browser developed by Google. Chromium is the name of the open source project behind Google Chrome, released under the BSD license.

Figure 8 Sample decomposition of a user agent

These are valuable insights when the developers of the website are working on improve the website layout. They might want to understand the most popular browsers or operating system used to access the website, therefore the website could be optimized for that browser or operating system. Especially when there are a lot of users are accessing the website using their smart phones, it makes sense for the website to show a mobile friendly version instead of the full desktop version. We have also included browser, operating system (os) and device as additional fields to store parsed information from user agents.

5.3 Dimension and Fact

In this session, we discuss the method and tools we have used to visualize and analyze the data we have cleaned in the previous stages. Through session grouping, bot identification, outlier removal, ip address matching, and user agent parsing, we were confident enough to believe the data contains only human sessions and the ip address and parse user agents are fairly accurate and complete. Dimension is defined as a collection of reference information about a measurable event. In this context, events are known as "facts." Dimensions categorize and describe data warehouse facts and measures in ways that support meaningful answers to business questions. They form the very core of dimensional modeling. [21]

Based on the above definition, we have defined the following dimension for our data analysis:

- city - city name of the ip address matched
- state - province or state name of the ip address matched
- country - country name of the ip address matched
- browser - browser used to access the website based on parsed user agent
- os - operating system used to access the website based on parsed user agent
- device - device type used to access the website based on parsed user agent
- mobile - whether a session visit is from a mobile device
- date - date of the session visit
- hour - Hour of the session visit
- weekend - whether a session visit happened on weekend
- bot - whether a session visit is identified as a bot

A fact, sometimes also called KPI (key performance indicator) is a value or measurement, which represents a fact about the managed entity or system. Facts at raw level are further aggregated to higher levels in various dimensions to extract more service or business-relevant information out of it. These are called aggregates or summaries or aggregated facts. [22] In our project, we have defined the following facts for the performance of the website:

- session count - count of grouped session visits
- duration - length of session visit in seconds

5.4 Statistical Testing

The cleaned data we obtained show distinct characteristics for different categories such as bot vs. nonbot sessions, weekend vs. weekend traffic and mobile vs. non-mobile session durations. We have employed a couple statistical analysis methods in our analysis in order to identify key differences in two categories.

The two-sample t-test is one of the most widely used hypothesis tests. It is applied to compare whether the average difference between two groups is really significant or if it is due instead to random chance. It helps to answer questions like whether the average success rate is higher after implementing a new sales tool than before or whether the test results of patients who

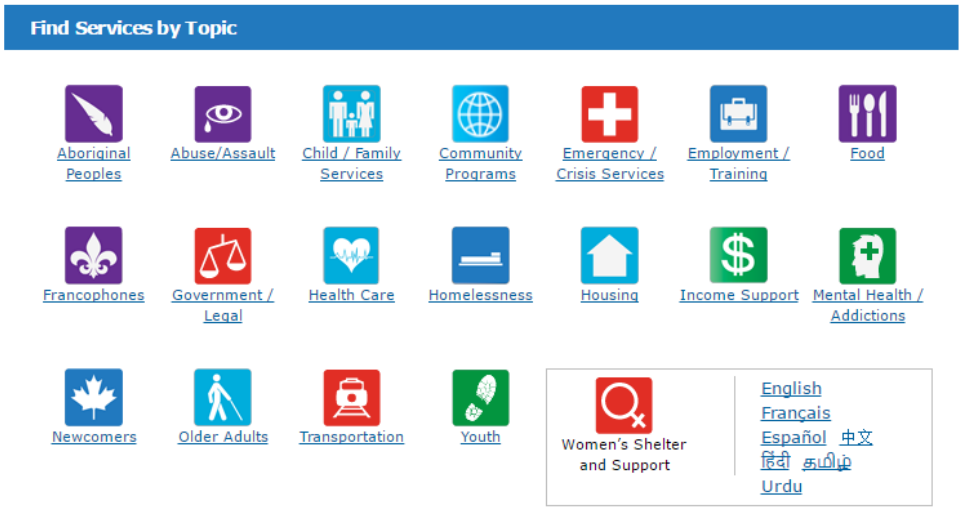
received a drug are better than test results of those who received a placebo in pharmaceutical industry. In our case, we have used to this test to identify the difference between number of daily session counts for bot and non-bot sessions, the difference in daily session counts for weekday vs. weekend. It is also used to measure whether a user would spend more time on the website if he uses a non-mobile device. In addition, we have also used analysis of variance (ANOVA) to prove whether the test results we received from the 2 sample t test are actually correct. ANOVA is used to test general rather than specific differences among means.

5.5 Seasonality Analysis

After plotting weblog session data in tableau, we could see clear seasonality pattern for daily and hourly session count by month for both bot and non-bot sessions. Therefore, we have conducted seasonality analysis using Minitab. The software allows us to decompose the data and extract seasonality factors for defined seasonal periods. The software went through steps in de-trending, smoothing, and uses the median of raw seasonal values to generated estimated seasonal indices. Then if the user is interested in predicting future values, Minitab uses the least squares regression trend line as well as seasonally adjusted data to compute predictions.

5.6 Sequential Pair Association Analysis

There are total of 18 topics listed on the front page of the website:



The topic women's shelter and support divides into different languages and needs to go through additional encryption for privacy, there it is not included in our list of topics for analysis. We are interested to understand the behavior of users visiting those topics. Some of questions that we are very interested to know:

- What is the most popular topic? What is its probability to be visited in a session?
- Does the probability of visiting a topic increase when user has already visit another topic?
- Does user conduct text search and visiting topics at the same time?

Therefore, we have decided to employ the method of sequential pair association analysis described in the paper of Wagner A. Kamakura. [21] In the paper, the author described the method of sequential market basket analysis in order to characterize the relationship between products bought by consumers as substitute, compliment or independent. Similarly, in this project, we would like to understand the same kind of characteristics for our website topics. In order to get the relationship between topics, we will have to first identify association rules. In general association analysis or market basket analysis, we will have to identify the following for each pair of products or topics A and B:

Support - the joint probability of finding the pair AB across all sessions. A low support means that the pair is not relevant because it is not visited together frequently enough. We are generally interested in pairs with high $P(A \cap B)$.

Confidence - the conditional probability $P(B|A)$, when a topic A is visited in a session, what is the condition probability the user will visit session B. It is often interpreted as the probability that visiting of topic A will lead to visiting of topic B.

Interest - the ratio between the joint probability and the probability of joint occurrence under independence $\frac{P(A \cap B)}{P(A)P(B)}$ It is an important measure for the popularity of two topics in all session visit pairs.

We also believed that the sequence of session visits will have profound effect. If a pair of topics is substitute of each other, then a user would already receive information from the topic he visited first and would have lower probability of visiting other topics. On the other hand, if a topic pair is compliment of each other, then visiting one of the topics in the pair will actually increase the probability of visiting the other. The paper used the measure of 'gain' to incorporate the sequential effect in the market basket analysis.

$\boldsymbol{P(A \rightarrow B)}$ - the probability that within a session, the user first visited topic A and then topic B

$\boldsymbol{Gain[A \rightarrow B]}$ – the % increase in probability for visiting topic B when topic A is already visited. In mathematics: $Gain[A \rightarrow B] = \frac{P(A \rightarrow B)}{P(A)}\%$

The measure of $Gain[A \rightarrow B]$ allows us to identify topics pairs with the highest probability increase when topic A is visited. Vise versa, $Gain[B \rightarrow A]$ is also calculated throughout the analysis. With calculated $Gain[A \rightarrow B]$ and $Gain[B \rightarrow A]$, we can make the following conclusion based on our calculation:

Strong $\boldsymbol{A \rightarrow B}$ relationship – when $Gain[A \rightarrow B]$ number is large positive number and $Gain[B \rightarrow A]$ is large negative number

Weak $\boldsymbol{A \rightarrow B}$ relationship – when $Gain[A \rightarrow B]$ number is small positive and $Gain[B \rightarrow A]$ is small negative number

Compliment relationship – when both $Gain[A \rightarrow B]$ and $Gain[B \rightarrow A]$ are both positive and more than 1

Substitute relationship – when both $Gain[A \rightarrow B]$ and $Gain[B \rightarrow A]$ are both negative and less than -1

Independence or no clearly defined relationship - when both $Gain[A \rightarrow B]$ and $Gain[B \rightarrow A]$ are both between (-1 , 1)

The project picked the last three month's data in order to characterize all 18 topics' relationship with each other. In order to avoid inflation of results, we have used the benchmark of 0.05% for the joint probability of topics pairs. In other words, topic pairs needs to be visited in a session 5 times in 10,000 sessions.

6. Results

In this section of the report, we will discuss results from the analysis of cleaned find help log data. We first started by investigate whether there is a difference between identified bot and non-bot sessions in session count and duration. Then detail session analysis has been done for both bot and non-bot sessions. Lastly, we have used the method of sequential pair association analysis to characterize the relationship between topics.

6.1 Facts about 211Toronto.ca

Total Number of sessions between December 2013 and November 2015: **818444**

Month with the highest number of session counts: **May, 2015**

Weekday with the highest number of session counts: **Tuesday**

Busiest hour of the day: **2:00 pm**

Date with highest non-bot activity: **12/12/2015**

Average sessions per month: **34102**

Average session per day: **1138**

Table 1 Web session count and average session duration for the 24 month period between 2013 and 2015

Year	Month	Sessions	Avg. Session Duration (s)
2013	December	21839	4800
2014	January	27111	4548
	February	24563	5227
	March	29911	4809
	April	27070	4854
	May	26259	4649

	June	25994	4721
	July	30234	4483
	August	29671	4613
	September	35715	4758
	October	36318	4644
	November	35501	4632
	December	33260	4561
2015	January	36141	4612
	February	36070	4457
	March	40674	4330
	April	39168	5251
	May	46803	6720
	June	45520	6508
	July	40486	5964
	August	40045	6288
	September	40934	6195
	October	41739	6474
	November	27418	6087
Grand Total		**818444**	**5263**

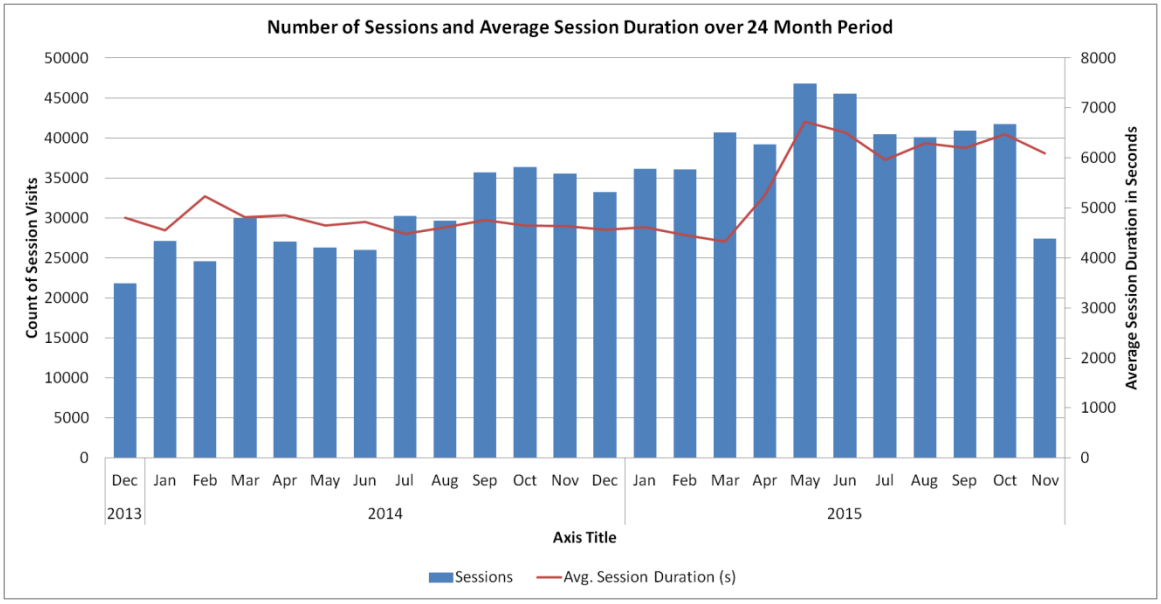


Figure 9 Number of Sessions and Average Session Duration over 24 Month Period

Table 1 and Figure 9 shows the growing trend in number of sessions visited the website everyday and average length of the session in second for the period between December 2013 and November 2015. We could clearly identify May 2015 is the month with the highest number of session visits as well as longest session durations.

Table 2 % Bot vs. % non-bot sessions over weekday and weekend

Session	Weekday	Weekend	Total
Non-bot	544431	112056	656487
Bot	118566	43391	161957
Total	662997	155447	818444

%	Weekday	Weekend	Total
Non-bot	67%	14%	81%
Bot	14%	5%	19%
Total	81%	19%	100%

Table 3 also shows interesting results. It is a surprise to us that the % of bot vs non-bot session follows closely with the 20/80 rule. Each month, almost 20% of the traffic is generated by bot activities and the rest of 80% are coming from non-bot users. If we split the result by weekday and weekend as well as bot and non-bot, then we can find the following probabilities listed in the tables above. The overall weekday and weekend split also very closely follows the 80/20 rule similar to non-bot vs. bot.

6.2 Top URLs

The following are top URLs visited by both non-bot and bot user agent. They are sorted based on total number sessions per day over 24-month period. The meaning of these URL is manually checked on the website to find out the page these URLs actually represents.

Table 4 Top Non-bot URL

Top Non-bot URL	Page	Session	% Total
/	Home	230179	34.72%
/findhelp-soap-search/community/autocomplete/M	Page contains no content	20302	3.06%
/topic/community-programs	Community Program	19665	2.97%
/topic/child-family-services	Child Family Services	18469	2.79%
/topic/employment-training	Employment Training	16456	2.48%
/topic/Central+Region/ORGANIZATION/fht142/Toronto+(City+of)	Food Bank	13277	2.00%
/topic/abuseassault	Abuse Assault	12233	1.85%
/topic/housing	Housing	12068	1.82%
/soap-query/food-banks-0	Food banks located in Toronto (City)	12018	1.81%
/detail/en/81049	Ontario Council of Agencies Serving Immigrants	12005	1.81%
Total		**662997**	**100.00%**

Table 3 shows the top URLs visited by non-bot sessions. Besides the home pages with more than 34% of sessions visited it, the 2nd highest page session visited contains no content. Community program is the most popular topic by session count.

Table 5 Top Bot URL

Top Bot URL	Page	Session	% Total
/	Home	39159	25.19%
/robots.txt	Robots.txt	27741	17.85%
/fr	French	22052	14.19%
/ont	Ontario	9526	6.13%
/topic/Central+Region/ORGANIZATION/fht142/Toronto+(City+of)	Food Bank	5477	3.52%
/node/23	Help Page	4936	3.18%
/print/node	Print Website Home Page	4831	3.11%
/topic/community-programs	Community Program	4554	2.93%
/soap-query/food-banks-0	Food Bank	4436	2.85%
/basic-page/site-map-0	Site Map	4052	2.61%
Total		**155447**	**100.00%**

Table 4 shows the top URLs visited by bot sessions. Different from non-bot sessions, a lot of not sessions visited the robots.txt page. As we explained previously, robots.txt file gives the permission and rules for robots to crawl the website, therefore we are expected to see a high % of sessions visited the file. Similar to its popularity in non-bot sessions, community program is also listed in the top 10 URLs visited by bot sessions.

6.3 Bot vs. Non-bot in session count and duration

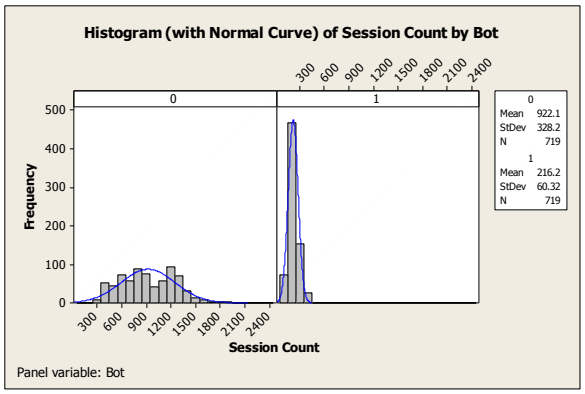
Results below are calculated using Minitab to evaluate the descriptive statistics of daily session count and average session duration in second for bot and non-bot sessions.

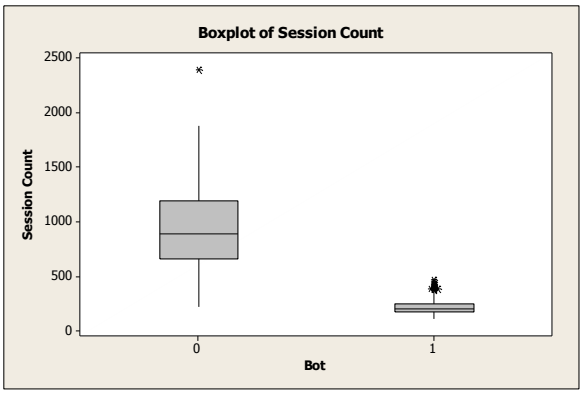
Descriptive Statistics: Session Count, Avg duration (s)

```
Variable          Bot    N     Mean  SE Mean  StDev  Minimum      Q1  Median
Session Count     0    719   922.1     12.2  328.2    219.0   657.0   888.0
                  1    719  216.20     2.25  60.32   112.00  171.00  203.00

Avg duration (s)  0    719  2755.8     23.5  629.6   1229.4  2328.7  2674.6
                  1    719   15750      198   5314     5174   11869   14895

Variable          Bot      Q3  Maximum   Range    IQR
Session Count     0    1192.0   2390.0  2171.0  535.0
                  1    250.00   467.00  355.00  79.00

Avg duration (s)  0    3111.1   6859.5  5630.1  782.4
                  1     18852    35816   30642   6982
```

6.3.1 Analysis of Session Count

It is statistically significant to show there is a difference between number of session by bot and by nonbot. First of all, as shown on the ANVOA table, the p-value, which is practically zero, is less than 5%, indicating rejection of null hypothesis of equal mean session count. Lastly, from both the histogram and the boxplot, it's evident that the dispersion of data is wildly different between the two groups, with the bot group having a relatively low dispersion.





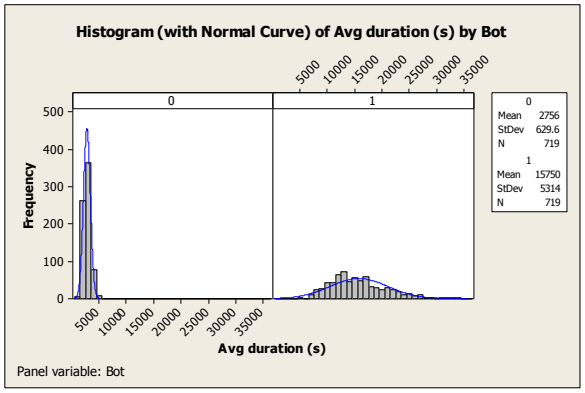
One-way ANOVA: Session Count versus Bot

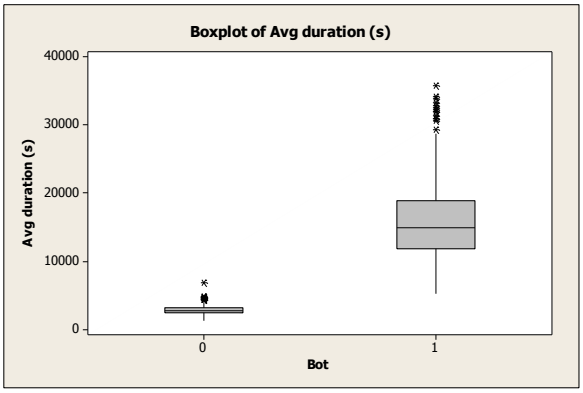
```
Source     DF          SS          MS        F      P
Bot         1   179142561   179142561  3217.37  0.000

Error    1436    79956169       55680
Total    1437   259098730

S = 236.0    R-Sq = 69.14%    R-Sq(adj) = 69.12%

                              Individual 95% CIs For Mean Based on Pooled StDev
Level     N    Mean  StDev    +---------+---------+---------+---------
0       719   922.1  328.2                                        (*)
1       719   216.2   60.3    (*)
                              +---------+---------+---------+---------
                             200       400       600       800
Pooled StDev = 236.0
```

The large F value and nearly zero p value indicates significant statistical evidence that the two category bot and non-bot have different number of sessions to the website. The confidence internal diagram also shows the evidence clearly with no overlapping of confidence interval.

6.3.2 Analysis of Session Duration

The two sample t-test also rejects the null hypothesis of equal mean session duration. In addition, the variation of session duration for the non-bot group is significantly lower than that of the bot group, as indicated by both the boxplot and the histogram.





One-way ANOVA: Avg duration (s) versus Bot

```
Source    DF           SS           MS        F      P
Bot        1  60704207365  60704207365  4240.11  0.000
Error   1436  20558717977     14316656
Total   1437  81262925343

S = 3784   R-Sq = 74.70%   R-Sq(adj) = 74.68%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level    N   Mean  StDev  ---+---------+---------+---------+------
0      719   2756    630  (*)
1      719  15750   5314                                      (*)
                          ---+---------+---------+---------+------
                          3500      7000     10500     14000

Pooled StDev = 3784
```

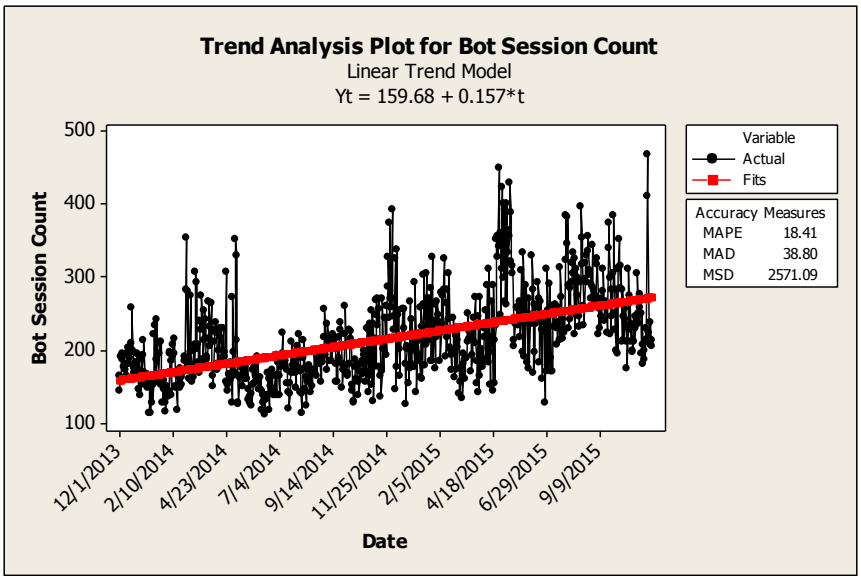
Again, the large F value and nearly zero p value indicates significant statistical evidence that the two category bot and non-bot have different session duration in seconds when visiting the website. Bot sessions has significantly longer sessions compare to non-bot sessions. In addition, the distribution of session durations for bot sessions are more concentrated around the mean, non-bot sessions' duration are more scattered.

6.4 Bot sessions

Based on our previous analysis, we discovered distinctive characteristics between bot and non-bot sessions for the amount of daily traffic as well as the duration spent on the website. In this section of the report, we want to investigate further into bot behaviours and understanding if there is any seasonality daily or hourly. In addition, using matched ip addresses, we were able to identify most frequently used browsers for crawling activities as well as top cities generated highest amount of traffic because of web crawling.

6.4.1 Daily Bot session count trend analysis

There is no clear indication of seasonality throughout the 24-month period as the variation around the trend line appears to be somewhat random. But the time series data has an increasing trend over time as indicated by the red linear trend line shown above, which may indicate either a growing number of bots crawling the web in general or a growing interest for the bots to crawl this particular site.



6.4.2 Bot session activities throughout the day

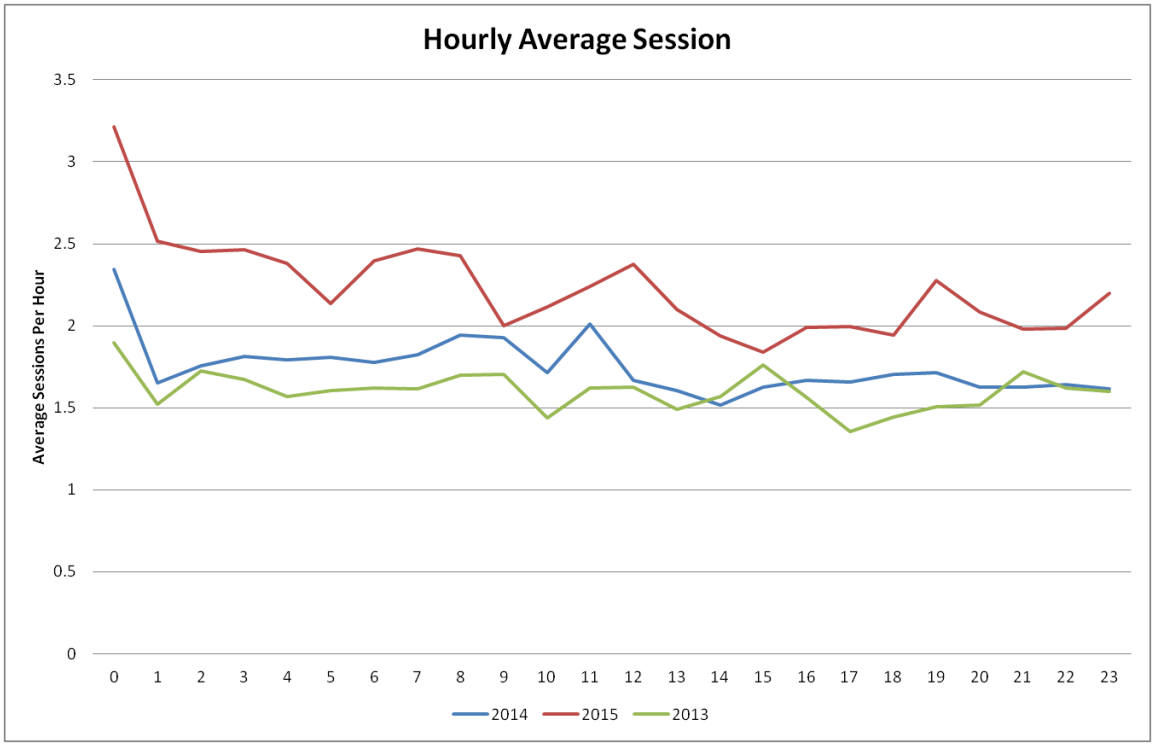


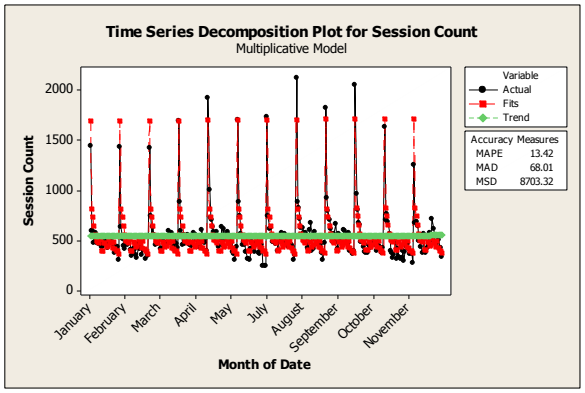
Figure 10 Hourly average bot sessions

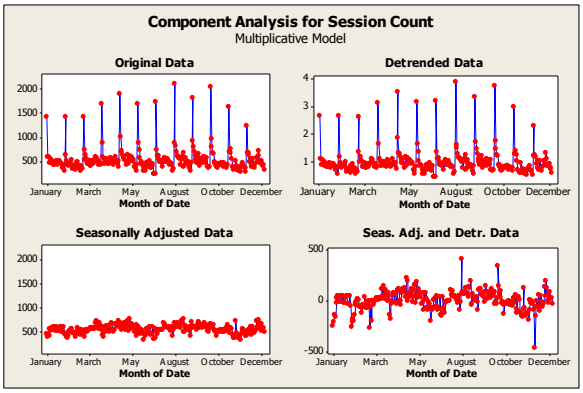
The graphs above and the time series decomposition below provides supporting evidence that strong seasonal pattern exists for bot session count data, as the daily times series graphs are similar in both the peak and the trough for bot sessions. Bot sessions are most active in early morning (12 - 1 am) and drops to a steady level throughout the day.

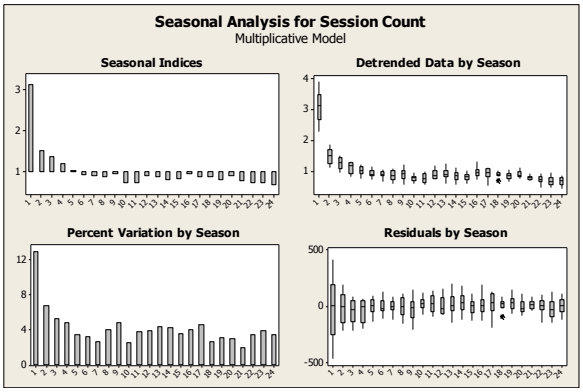
Time Series Decomposition for Session Count

Seasonal Indices

Period	Index				
1	3.13975	9	0.95209	18	0.86804
2	1.50071	10	0.72939	19	0.78982
3	1.35378	11	0.73226	20	0.90395
4	1.18807	12	0.88644	21	0.77203
5	1.02687	13	0.87945	22	0.72173
6	0.91984	14	0.79926	23	0.72501
7	0.90269	15	0.82571	24	0.67483
8	0.87498	16	0.95342		
		17	0.87991		







From the component analysis above, it can be shown that no clear trend exists since the trend line is flat as well as the de-trended data appears not significantly different from the original data. From the Seasonal Indices plot, the number of sessions is most active around midnight from 0 to 1 AM and quickly decays afterwards, and the de-trended shares the same story. From the residuals plot, only minor decay in variances, and the mean is around zero.

6.4.3 Top Location and Browsers used by Bot sessions

Table 6 Top 10 cities by bot activity

city	country	Sessions	% Bot Sessions	% Total Sessions
Beijing	China	49700	31.97%	6.07%
Mountain View	United States	21902	14.09%	2.68%
Redmond	United States	14531	9.35%	1.78%
Zhengzhou	China	10908	7.02%	1.33%
Shanghai	China	9215	5.93%	1.13%
Cologne (Innenstadt, Cologne)	Germany	3506	2.26%	0.43%
Ashburn	United States	3023	1.94%	0.37%
Seoul	Republic of Korea	2879	1.85%	0.35%
Issy-les-Moulineaux	France	2803	1.80%	0.34%

Table 7 Top 10 browsers by bot activity

browser	Session	% Bot Sessions	% Total Sessions
Baiduspider	49658	31.95%	6.07%
360Spider	20129	12.95%	2.46%
Googlebot	16910	10.88%	2.07%
bingbot	15500	9.97%	1.89%
Googlebot-Mobile	5767	3.71%	0.70%
XoviBot	5252	3.38%	0.64%
Daumoa	4990	3.21%	0.61%
OrangeBot	4714	3.03%	0.58%
Java	3075	1.98%	0.38%
favicon	2190	1.41%	0.27%

6.5 Non-bot Session Analysis

In this section, we deep dive into the user behaviours of non-bot sessions. Using statistical analysis such as 2 sample t test and seasonality decomposition, we were able to identify the difference in session count between weekday and weekend as well as seasonality indices from daily and hourly trend.

6.5.1 Weekend vs. Weekday Session Count and duration

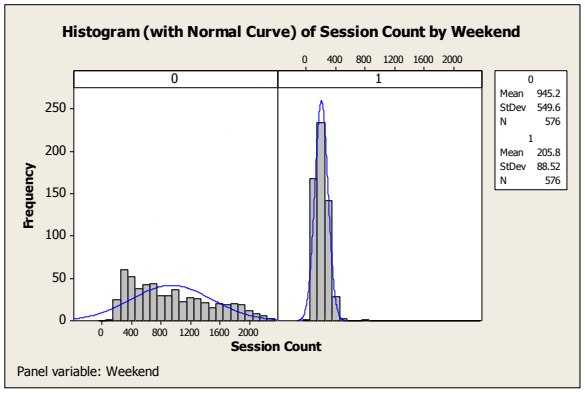
Significant statistical evidence to prove there is a difference in session count between weekend and weekday groups for non-bot agents. The p-value of zero (less than 5%) leads to the rejection of the null hypothesis of equal session counts between weekdays and weekends. Both the mean and the variance are significantly higher for the weekday group than the weekend group.

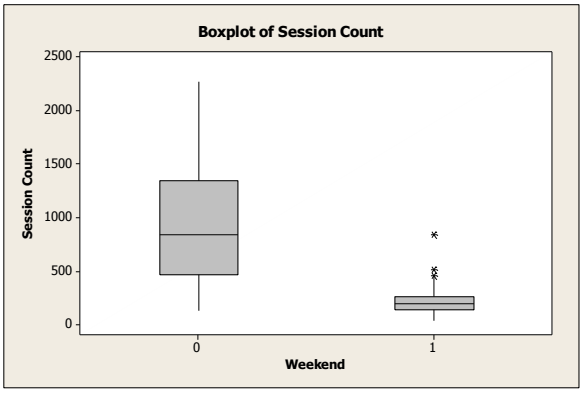
Descriptive Statistics: Session Count, Avg Duration

```
Variable       Weekend    N     Mean  SE Mean  StDev  Minimum      Q1  Median
Session Count  0        576   945.2     22.9  549.6    132.0   463.5   837.5
               1        576  205.84     3.69  88.52    40.00  138.25  195.00
```

Avg Duration	0	576	3183	109	2609	64	933	2967
	1	576	2673	113	2719	52	752	1750

Variable	Weekend	Q3	Maximum	Range	IQR
Session Count	0	1344.3	2276.0	2144.0	880.8
	1	263.00	843.00	803.00	124.75
Avg Duration	0	4488	13451	13386	3554
	1	3656	14075	14023	2904





One-way ANOVA: Session Count versus Weekend

Source	DF	SS	MS	F	P
Weekend	1	157431422	157431422	1015.95	0.000
Error	1150	178204298	154960		
Total	1151	335635720			

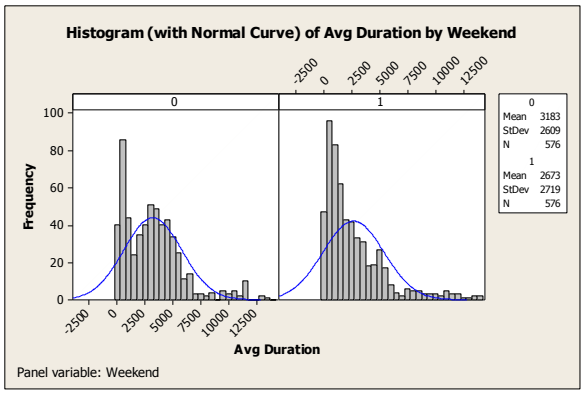
S = 393.6 R-Sq = 46.91% R-Sq(adj) = 46.86%

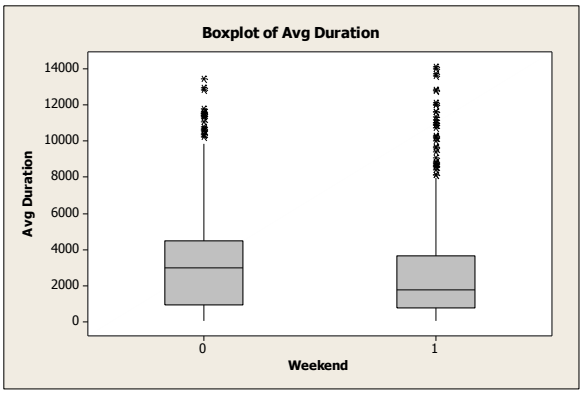
Level	N	Mean	StDev	Individual 95% CIs For Mean Based on Pooled StDev
0	576	945.2	549.6	(*)
1	576	205.8	88.5	(*-)

250 500 750 1000

Pooled StDev = 393.6

Significant statistical evidence to prove there is a difference in session duration between weekend and weekday. Even though the variance of session duration between the two group is relatively close, and both group exhibits right skewness as shown on the boxplot, the p-value of 0.1% (less than 5%) leads to the rejection of the null hypothesis of equal mean, indicating higher average session duration for the weekday group.





One-way ANOVA: Avg Duration versus Weekend

```
Source      DF          SS         MS      F      P
Weekend      1    74878401   74878401  10.55  0.001
Error     1150  8162154965    7097526
Total     1151  8237033366

S = 2664    R-Sq = 0.91%    R-Sq(adj) = 0.82%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level    N  Mean   StDev   --+---------+---------+---------+-------
0      576  3183    2609                        (-------*--------)
1      576  2673    2719   (--------*--------)
                           --+---------+---------+---------+-------
                          2500      2750      3000      3250

Pooled StDev = 2664
```

6.5.2 Analysis of Mobile Sessions

Year of Date	Month of Date	Not Mobile	Mobile	Grand Total	% Mobile
2013	December	14745	1427	16172	8.82%
2014	January	19831	2003	21834	9.17%
2014	February	17881	2063	19944	10.34%
2014	March	20723	2366	23089	10.25%
2014	April	19026	1970	20996	9.38%
2014	May	18452	2396	20848	11.49%
2014	June	18745	2706	21451	12.61%
2014	July	21383	3467	24850	13.95%
2014	August	20564	3805	24369	15.61%
2014	September	25196	4514	29710	15.19%
2014	October	26288	4453	30741	14.49%
2014	November	24319	4502	28821	15.62%
2014	December	22160	4036	26196	15.41%
2015	January	24128	4997	29125	17.16%
2015	February	24717	4834	29551	16.36%

2015	March	28215	6261	34476	18.16%
2015	April	26121	5536	31657	17.49%
2015	May	31830	5917	37747	15.68%
2015	June	31993	6461	38454	16.80%
2015	July	26409	6298	32707	19.26%
2015	August	24878	5888	30766	19.14%
2015	September	26170	6502	32672	19.90%
2015	October	27557	6487	34044	19.05%
2015	November	18640	4137	22777	18.16%
Grand Total		**559971**	**103026**	**662997**	15.54%

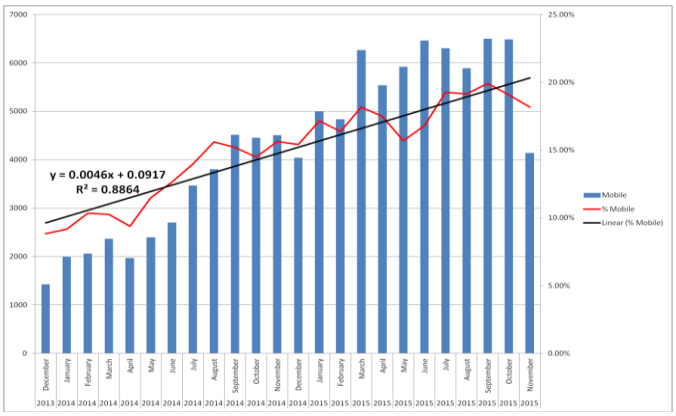


Figure 11 Mobile session count and % mobile over 24 month period

We have noticed a clear Increasing trend for % mobile device/browser used to visit the website, as shown with the clearly upward trend line above. The R squared of 0.886 also shows a very good fit.

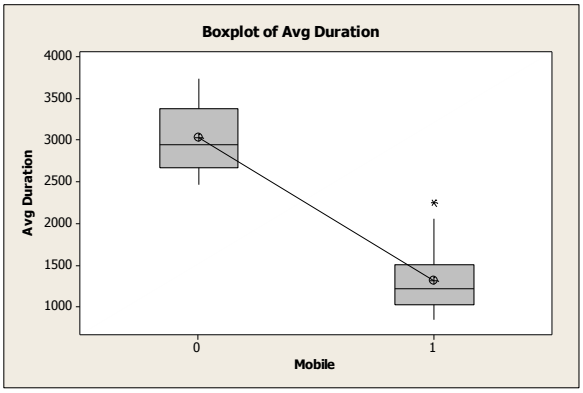
Two-Sample T-Test and CI: Avg Duration, Mobile

```
Two-sample T for Avg Duration

                          SE
Mobile   N  Mean  StDev  Mean
0       24  3028    383    78
1       24  1307    392    80


Difference = mu (0) - mu (1)
Estimate for difference:  1721
95% lower bound for difference:  1533
T-Test of difference = 0 (vs >): T-Value = 15.37  P-Value = 0.000  DF = 45
```

One-way ANOVA: Avg Duration versus Mobile

```
Source  DF        SS        MS       F      P
Mobile   1  35541434  35541434  236.39  0.000
Error   46   6916198    150352
Total   47  42457631

S = 387.8   R-Sq = 83.71%   R-Sq(adj) = 83.36%
```

```
                                Individual 95% CIs For Mean Based on Pooled StDev
Level    N     Mean   StDev    -+---------+---------+---------+--------
0       24   3028.3   383.3                                    (-*--)
1       24   1307.3   392.2    (--*-)
                                -+---------+---------+---------+--------
                               1200      1800      2400      3000

Pooled StDev = 387.8
```



Significant statistical evidence to prove that non-mobile users spend more time on the website compared to mobile users, as the null hypothesis of equal session duration is rejected, with a p-value of zero.

6.5.3 Trend Analysis of non-bot sessions

We can clearly identify a monthly seasonality for decreased number session counts for Saturday and Sunday. The green line represents sessions without using a mobile device and the red line represents sessions from user who used a smartphone. We can also see a general increasing trend of session count generated by mobile users is shown by the chart below, and it is applicable to both mobile and non-mobile sessions.

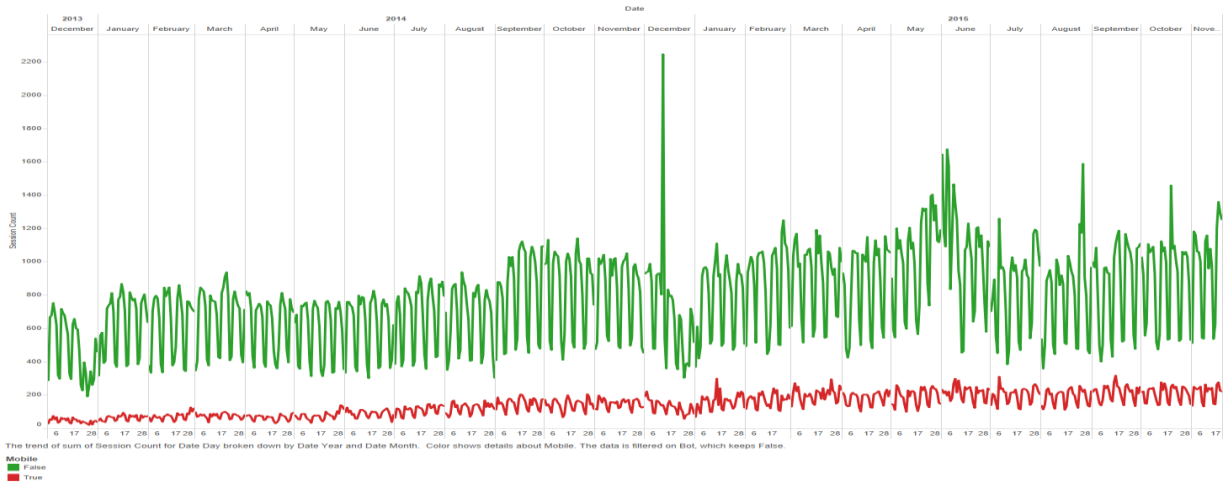


Figure 12 Daily session count over 24 month period

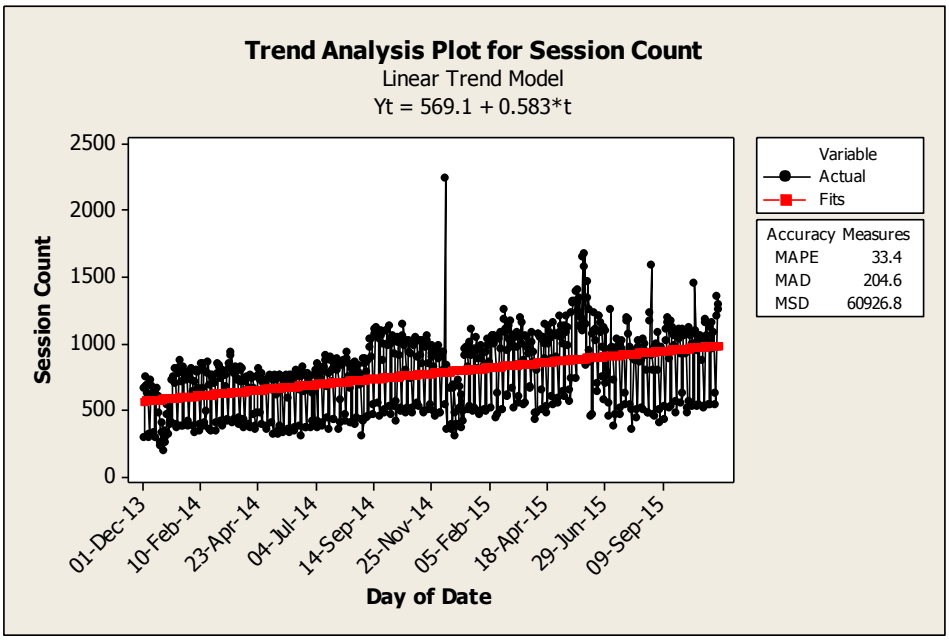
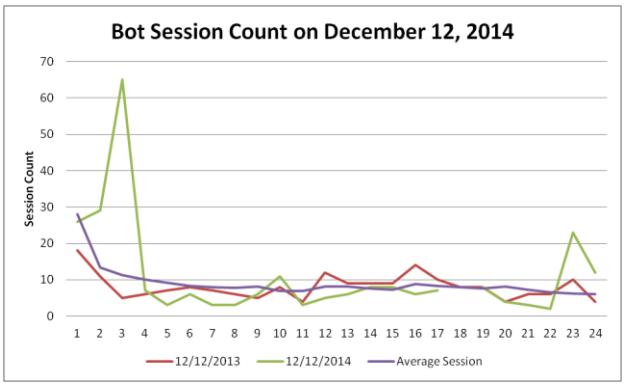
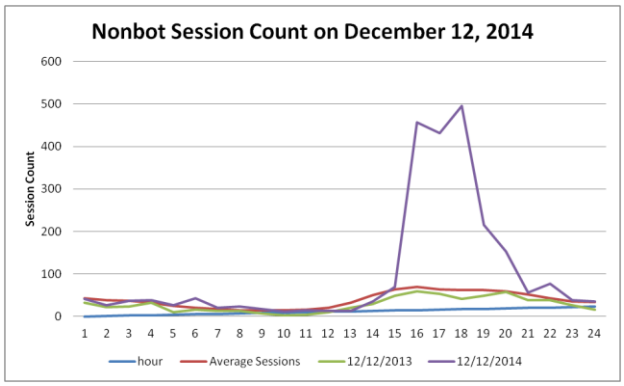


Figure 13 Trend Analysis Plot for Daily Session Count

Figure 12 and Figure 13 not only showed a clear upward trend of increasing session visits to the website. Additionally, we were able to spot some abnormal traffic during mid December 2014 and early June 2015. With some investigations, we identified December 12, 2014 and first week of June 2015 are time periods with abnormally high session visits. Therefore, we have further analyzed their sessions.



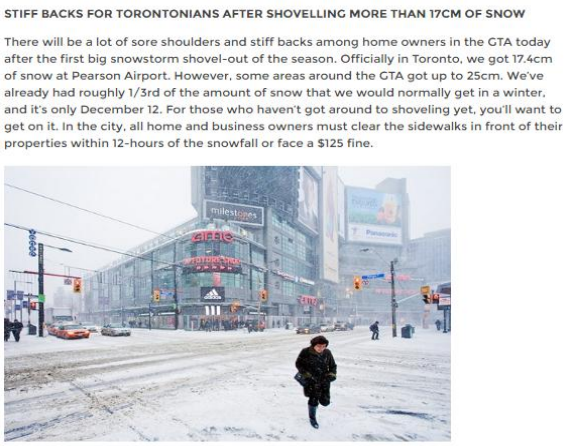


The two figure at the top showed there are abnormally high number of sessions visited the website from both bot and non-bot sessions. The traffic of bot sessions spikes between 2 am and 3 am and levels off to the normal level. On the other hand, the traffic of non-bot sessions starts off with the same level and the average, yet it spikes sharply around 3 pm in the afternoon to almost 5 times to the average.

Topic	12/12/2013	12/12/2014	Average
communityprograms	29	26	27
childfamilyservices	31	20	26
employmenttraining	20	19	23
abuseassault	22	14	17
healthcare	19	14	16
mentalhealth	0	0	0
housing	0	21	17
youth	18	8	14
food	0	19	13
olderadults	17	11	13
governmentlegal	19	11	11
aboriginalpeoples	19	9	9
newcomers	0	20	12
incomesupport	11	13	12
homelessness	0	13	11
emergencycrisisservices	0	11	8
francophones	7	3	6
transportation	5	7	5
quicksearch	182	159	159
Other	577	2286	823

The distribution of topic count didn't show much abnormal activity except for the "other" category which it is almost 3 times the normal average.

A quick search on the internet for December 12, 2014 showed there was a big snow fall with more than 17 cm of snow. [25] The first big snowstorm of the year is measured officially in Toronto with 17.4cm of snow. However, some areas around the GTA got up to 25cm. We believe the snowfall resulted many people needed the city service information from the website in order to deal with the harsh weather and condition. As a consequence, we have had this abnormal one day spike in session count.



STIFF BACKS FOR TORONTONIANS AFTER SHOVELLING MORE THAN 17CM OF SNOW

There will be a lot of sore shoulders and stiff backs among home owners in the GTA today after the first big snowstorm shovel-out of the season. Officially in Toronto, we got 17.4cm of snow at Pearson Airport. However, some areas around the GTA got up to 25cm. We've already had roughly 1/3rd of the amount of snow that we would normally get in a winter, and it's only December 12. For those who haven't got around to shoveling yet, you'll want to get on it. In the city, all home and business owners must clear the sidewalks in front of their properties within 12-hours of the snowfall or face a $125 fine.

Figure 14 News about Toronto's snowfall on December 12, 2014

In addition, we also noticed the high website traffic during the first week of June 2015. During that week, the number of sessions on the website almost doubles the 2014's number. Search on the internet showed there as a potential possibility to have a strike for the public elementary school teachers. [26] Parents might be worried about their kids have no school to go to, therefore searching on the website to find community programs that will be able to assist them taking care of their kids.

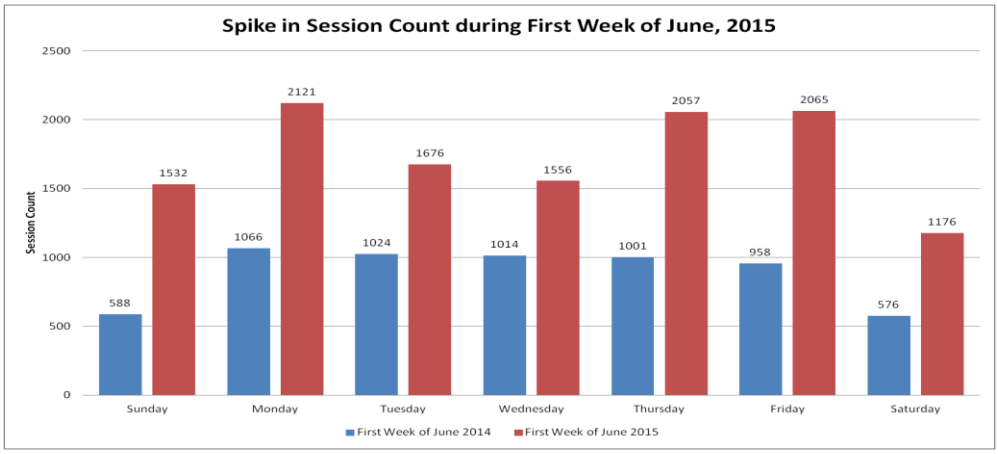
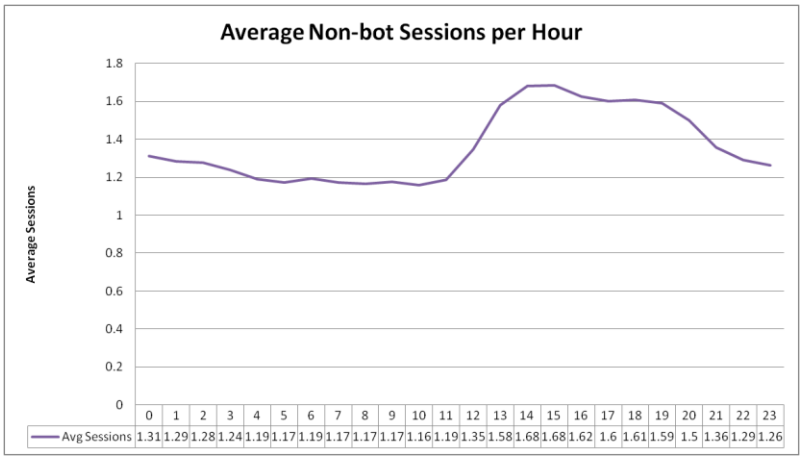


Figure 15 Spike in Website Traffic during first week of June 2015

6.5.4 Hourly Trend and Seasonal Analysis by Month

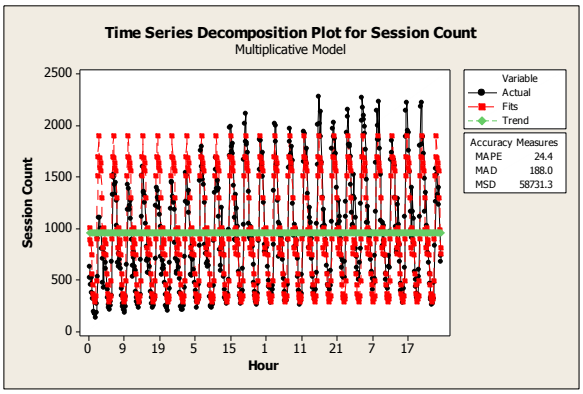


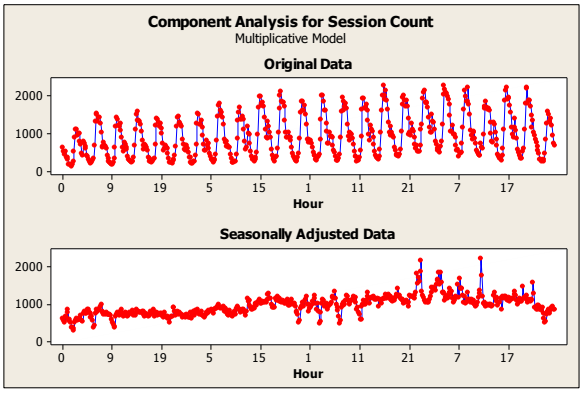
The analysis shows a clear seasonal pattern for the weekday session counts during the day, with the peak around the afternoon to midnight and trough around the late night to before noon. The flat trend line indicates no clear trend during the day. The variation of session counts also increases during the peak hours and decreases during the trough. The residuals have the mean around zero and high variance around 2-3 PM in the afternoon.

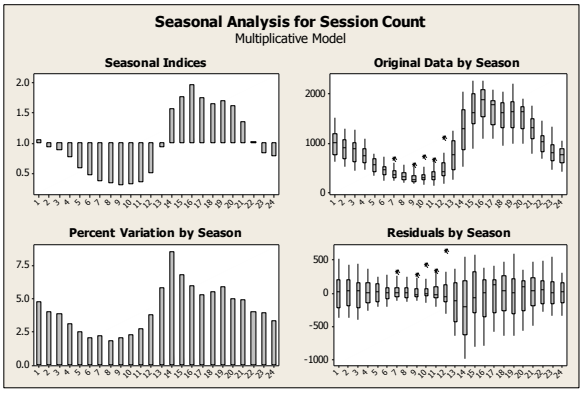
Time Series Decomposition for Weekday Session Count

```
Seasonal Indices

Period     Index                          3   0.89387
     1   1.05101                          4   0.77554
     2   0.93254                          5   0.58154
```

6	0.47729
7	0.37778
8	0.33989
9	0.30318
10	0.32751
11	0.35929
12	0.51233
13	0.94147
14	1.57870
15	1.77145
16	1.98104
17	1.76009
18	1.66368
19	1.71415
20	1.63362
21	1.35449
22	1.03020
23	0.84822
24	0.79111







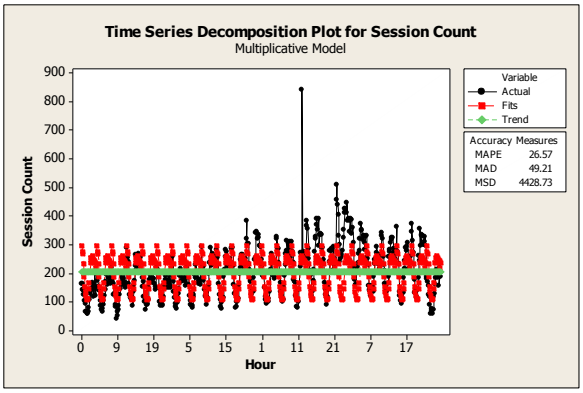
The weekend data also shows seasonal pattern, though the seasonality is much weaker compared to the weekday data. The positions of the peak and trough are similar to that of the weekday data. The trend line is flat and the percent variation by season on average is smaller too as compared to the weekday data. The residuals are centered around zero and have lower variations between them than the weekday residuals. This may indicate that non-bot users are less sensitive to the time of day when browsing the website during weekends than during weekdays.

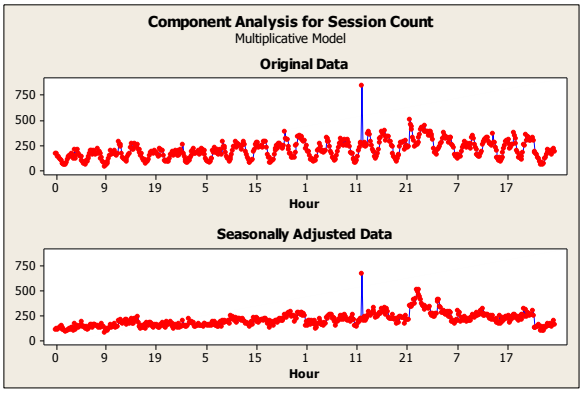
Time Series Decomposition for Weekend Session Count

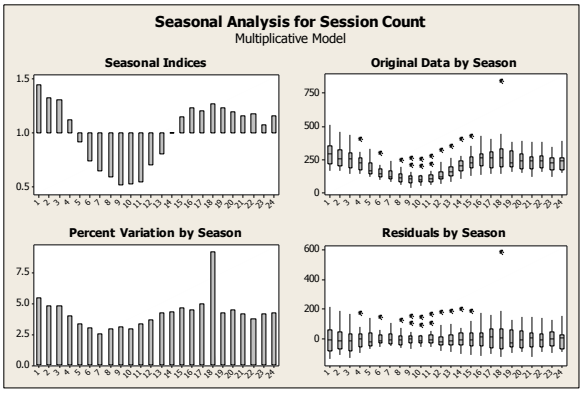
Seasonal Indices

Period	Index
1	1.44246
2	1.32664
3	1.30753
4	1.12288
5	0.91396

6	0.73831	16	1.22943
7	0.64618	17	1.19841
8	0.59102	18	1.26245
9	0.51847	19	1.22560
10	0.52738	20	1.19137
11	0.54230	21	1.15168
12	0.70488	22	1.17759
13	0.80755	23	1.06861
14	1.00171	24	1.15481
15	1.14879		







6.5.5 Top location, OS, device and browser used for non-bot sessions

Table 8 Top 10 cities with highest non-bot activity

City	Sessions	% Total
Toronto	253930	38.30%
Scarborough	34562	5.21%
Mississauga	19362	2.92%
Gunzenhausen (Frickenfelden)	14839	2.24%
North York	13560	2.05%
Beijing	13339	2.01%
Etobicoke	12269	1.85%
Ashburn	10767	1.62%
Brampton	10328	1.56%
Montreal	8396	1.27%

Table 9 Top 10 OS with highest non-bot activity

os	Sessions	% Total
Windows	432074	65.17%
Other	160645	24.23%
Android	53832	8.12%
Linux	6868	1.04%
Ubuntu	4514	0.68%
BlackBerry OS	4154	0.63%
BlackBerry Tablet OS	345	0.05%
Windows 95	78	0.01%
Nokia Series 40	65	0.01%
Symbian OS	52	0.01%

Table 10 Top 10 browsers with highest non-bot activity

browser	Sessions	% Total
Chrome	196089	29.58%
Other	152140	22.95%
Firefox	96793	14.60%
Mobile Safari UI/WKWebView	73603	11.10%
IE	62192	9.38%
Safari	54661	8.24%
Opera	5395	0.81%
BlackBerry WebKit	4030	0.61%
Chrome Mobile iOS	3181	0.48%
Samsung Internet	2671	0.40%

6.6 Sequential Pair Association Analysis

As we have discussed previously, there are total of 18 topics listed on the website for visitors to obtain useful information from each one of them. We would like to understand the popularity of all 18 topics as well as their relationship with each other.

Figure 14 shows the top 5 topics' session count over a 24 month period. We could clearly see a pattern throughout the time period that we have data. The number of sessions visited these topics raises in January of each year then gradually lowered to the minimum level in August. Then the number of sessions spikes sharply in September and maintains its level throughout September and October. The number of sessions declines in November and December, and the cycle repeats in next year. We can also summarize that all top 5 topics follows the same seasonality and moves in the same direction throughout the 24 month time period.

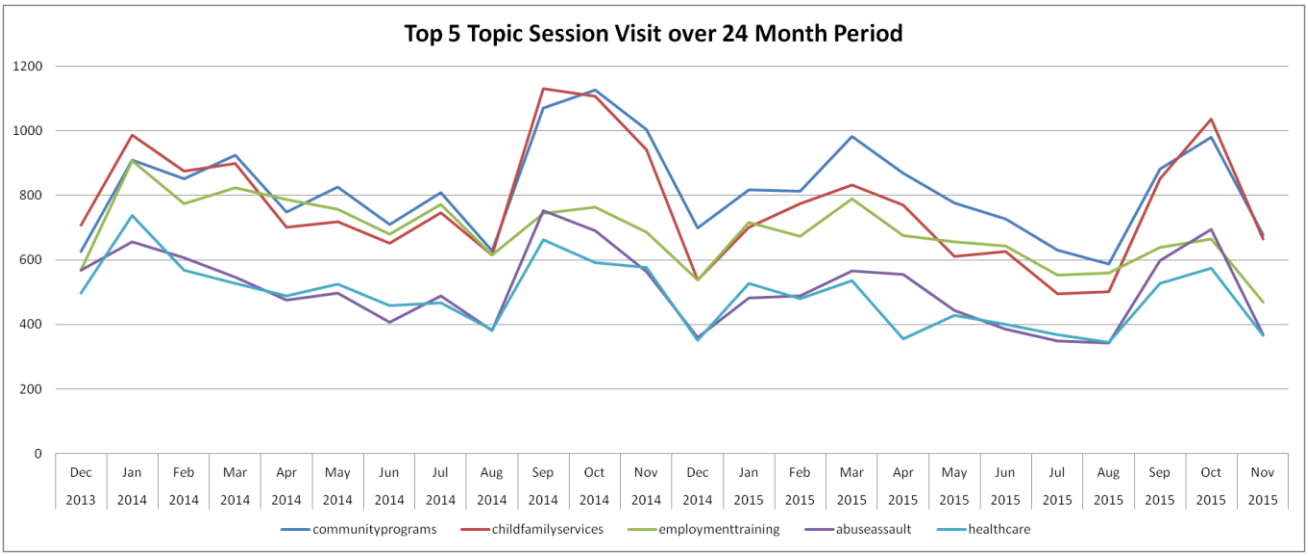


Figure 16 Top 5 topics' session count over 24 month period

We have chosen the last three month's data for our analysis in order to capture the most recent interests for non-bot users. For each combination of topic pairs, their support, confidence, interest as well as % gain will be calculated based on the method describe in methodology section. A threshold of 0.5% for the joint probability is implemented to filter out low count topic pairs.

Measure	2015-09		2015-10		2015-11		Total
Topics	Session	% Total	Session	% Total	Session	% Total	Session
quicksearch	5416	16.38%	6227	18.07%	4078	17.58%	15721
childfamilyservices	898	2.72%	1105	3.21%	705	3.04%	2708
communityprograms	922	2.79%	1036	3.01%	712	3.07%	2670
employmenttraining	665	2.01%	675	1.96%	494	2.13%	1834
housing	681	2.06%	651	1.89%	493	1.69%	1825
mentalhealth	651	1.97%	702	2.04%	442	1.91%	1795
abuseassault	626	1.89%	739	2.14%	393	2.13%	1758
healthcare	554	1.68%	609	1.77%	385	1.91%	1548
youth	529	1.60%	639	1.85%	341	1.66%	1509
food	458	1.39%	582	1.69%	380	1.64%	1420
homelessness	464	1.40%	509	1.48%	334	1.10%	1307
newcomers	488	1.48%	491	1.42%	323	1.26%	1302
incomesupport	591	1.79%	376	1.09%	262	1.13%	1229
aboriginalpeoples	648	1.96%	272	0.79%	219	1.39%	1139
olderadults	388	1.17%	434	1.26%	292	1.47%	1114
governmentlegal	371	1.12%	430	1.25%	255	1.44%	1056
emergencycrisisservices	363	1.10%	398	1.15%	221	0.95%	982
francophones	418	1.26%	173	0.50%	110	0.94%	701
transportation	155	0.47%	192	0.56%	146	0.63%	493

Table 11 Count of sessions visited the 18 topics in September, October, and November 2015

6.6.1 September 2015

Table 12 shows topic pairs with joint probability more than 0.5% and it is sorted by interest from high to low. We noticed that topic pairs with 'francophones' are ranked high on the list in September, but in any other month. Topic 'francophones' itself provides no particular useful contents, therefore we suspect that many users will exist the topic and start browsing some other topics instead. Consequently, we observed high association for pairs with 'francophones'. Other than the top 2 topic pairs, the rest of the list is intuitively accurate for describing people's interests in Toronto.

A	B	interest	GainA->B	GainB->A	Relationship
incomesupport	francophones	45.77388	5.558246	37.27874	COMP
aboriginalpeoples	francophones	43.09023	-0.02345	39.89299	B -> A
aboriginalpeoples	incomesupport	31.77172	1.331077	6.94293	COMP
abuseassault	emergencycrisisservices	24.59015	1.910077	17.9155	COMP
childfamilyservices	emergencycrisisservices	18.25765	2.347236	11.98322	COMP
housing	homelessness	18.10241	9.882374	5.278293	COMP
communityprograms	emergencycrisisservices	16.89328	2.852853	9.175484	COMP
abuseassault	aboriginalpeoples	16.38333	10.49278	2.260364	COMP
childfamilyservices	abuseassault	15.17485	7.999039	3.646563	COMP
abuseassault	mentalhealth	14.92856	2.73214	10.11529	COMP
childfamilyservices	youth	12.45881	2.967329	7.421874	COMP
communityprograms	abuseassault	11.57182	6.676353	1.807025	COMP
communityprograms	childfamilyservices	10.4229	5.269715	2.314563	COMP
communityprograms	employmenttraining	10.29998	1.157063	6.387941	COMP
childfamilyservices	aboriginalpeoples	10.28449	6.159366	1.272815	COMP
childfamilyservices	mentalhealth	10.06742	1.771368	6.239491	COMP
communityprograms	mentalhealth	10.0257	2.911126	5.059491	COMP
childfamilyservices	employmenttraining	9.246426	0.937874	5.478035	~COMP
youth	quicksearch	3.196692	0.904167	0.292525	IND
communityprograms	quicksearch	3.039195	0.158734	0.88046	IND
mentalhealth	quicksearch	2.963349	0.875537	0.087812	IND
newcomers	quicksearch	2.902317	0.701358	0.200959	IND
homelessness	quicksearch	2.81561	0.605161	0.197292	IND
childfamilyservices	quicksearch	2.685329	0.264484	0.420845	IND
housing	quicksearch	2.680407	0.389509	0.290899	IND
abuseassault	quicksearch	2.301518	0.267785	0.033733	IND
healthcare	quicksearch	2.203926	0.410513	-0.20659	IND
employmenttraining	quicksearch	1.982937	-0.21968	0.193434	IND
aboriginalpeoples	quicksearch	1.667535	0.036321	-0.36879	IND

Table 12 Topic pairs with joint probability more than 0.5% and sorted by interest in September 2015

6.6.2 October 2015

In October, we noticed that the ranking of topic pairs by interest shuffles. Popular topics such as ("homelessness", "housing") or ("aboriginalpeoples", "incomesupport") disappeared from the list.

Instead, we have a lot of topic pairs contain “communityprograms”. It is a good indication that people’s interest shifts from period to period.

A	B	interest	GainA->B	GainB->A	Relationship
abuseassault	emergencycrisisservices	22.26406	3.335633	14.11613	COMP
childfamilyservices	emergencycrisisservices	16.14359	2.291412	9.893006	COMP
abuseassault	youth	15.76474	4.181928	9.43684	COMP
communityprograms	incomesupport	15.57196	4.131669	8.555522	COMP
communityprograms	emergencycrisisservices	15.54706	2.259868	7.776569	COMP
childfamilyservices	incomesupport	14.93141	3.81123	8.456556	COMP
mentalhealth	youth	14.0602	3.763564	7.605148	COMP
childfamilyservices	abuseassault	13.59025	5.795125	3.895866	COMP
communityprograms	governmentlegal	13.53905	1.630444	9.289676	COMP
childfamilyservices	youth	13.42295	3.246532	8.078793	COMP
abuseassault	mentalhealth	13.41986	3.318273	7.902286	COMP
childfamilyservices	governmentlegal	13.41897	1.32112	9.299968	COMP
communityprograms	youth	11.76593	3.060806	6.653058	COMP
communityprograms	food	10.9748	2.029503	5.973573	COMP
communityprograms	childfamilyservices	10.89845	4.900819	2.733171	COMP
communityprograms	abuseassault	10.66897	5.527428	1.926088	COMP
childfamilyservices	food	10.45027	1.733148	5.752484	COMP
communityprograms	employmenttraining	10.15271	1.16854	5.99847	COMP
communityprograms	healthcare	9.55959	0.966544	6.319915	~COMP
childfamilyservices	mentalhealth	9.019344	1.621386	5.264667	COMP
communityprograms	mentalhealth	8.861822	1.46425	5.065846	COMP
childfamilyservices	employmenttraining	8.086308	0.709677	4.544897	~COMP
communityprograms	quicksearch	2.804781	0.0952	0.704238	IND
youth	quicksearch	2.572498	0.498458	0.074039	IND
childfamilyservices	quicksearch	2.484385	0.147024	0.327343	IND
abuseassault	quicksearch	2.209413	0.160878	0.048535	IND
mentalhealth	quicksearch	2.160294	0.198411	-0.03812	IND
newcomers	quicksearch	2.130491	0.127244	0.003247	IND
housing	quicksearch	2.057471	0.020234	0.037238	IND
homelessness	quicksearch	2.044276	0.120002	-0.07573	IND
food	quicksearch	1.978061	0.01756	-0.05852	IND
healthcare	quicksearch	1.944894	0.072418	-0.12752	IND
employmenttraining	quicksearch	1.869522	-0.19643	0.065955	IND

Table 13 Topic pairs with joint probability more than 0.5% and sorted by interest in October 2015

6.6.3 November 2015

A	B	interest	GainA->B	GainB->A	Relationship
childfamilyservices	youth	11.8669	2.280281	7.200703	COMP

childfamilyservices	abuseassault	10.88272	5.194777	2.013675	COMP
communityprograms	food	10.20138	1.571777	5.429443	COMP
communityprograms	healthcare	10.0689	2.553728	4.58443	COMP
communityprograms	childfamilyservices	9.842063	4.637238	2.003446	COMP
communityprograms	employmenttraining	9.825507	1.901492	5.132699	COMP
communityprograms	abuseassault	9.698152	5.465434	1.238035	COMP
communityprograms	housing	8.457825	1.973454	3.823603	COMP
communityprograms	quicksearch	2.69202	0.086394	0.58965	IND
housing	quicksearch	2.249656	0.049839	0.199817	IND
mentalhealth	quicksearch	2.187535	-0.02204	0.209578	IND
youth	quicksearch	2.16829	0.217578	-0.04929	IND
childfamilyservices	quicksearch	2.145957	-0.10451	0.250463	IND
healthcare	quicksearch	2.023896	-0.01021	0.034108	IND
employmenttraining	quicksearch	1.957268	-0.21709	0.174361	IND
food	quicksearch	1.855951	0.092616	-0.23667	IND
abuseassault	quicksearch	1.736669	-0.07378	-0.18955	IND

Table 14 Topic pairs with joint probability more than 0.5% and sorted by interest in November 2015

7. Discussion

The statistical analysis and data visualization in the result section showed some very interesting findings for both bot and no-bot sessions. We can slice and dice the data to see different trends and seasonality. In this section, we would like to discuss some of expected reasons and explanations for those results and trend we analyzed.

7.1 Bot vs. Non-bot Sessions

The first statistical analysis conducted is to analyze the difference between the behavior of robots and human sessions. It is easy for us to make the assumption that robots will have a different behavior. The two sample t tests on session count and duration both proved it with high statistical significance. We noticed that approximately 19 to 20 percent of sessions are generated by robots and the number of bot sessions per day varies very little from each other. The standard deviation of daily bot sessions is only 60.32 sessions. On the other hand, the spread of non-bot distribution for number of sessions per day is wider and has more variability. The standard deviation of non-bot daily session counts is 328.2 sessions which is more than 5 times of robots' standard deviation. It is expected that robots will have consistent visit patterns without much variation throughout a long period of time. However, the traffic generated from human or non-bot users will vary based on many factors such as day of the week, month of the year or hour of the day, etc. The 2 sample t test for number of sessions per day for bot and non-bot showed significant result. It indicates there is strong statistical evidence for us to conclude the number of sessions per day from non-bot users is larger than bot sessions.

However, it is a different case for the average session duration in seconds. Although there are less bot sessions every day, these bot sessions spend longer period of time on the website. On average, bot sessions spend 15,750s or more than 4 hours per session crawling the website. Non-bot users spend

significantly less time on the website. On average, they spent 2755s or approximately 45 minutes browsing the website. This is a very reasonable measurement and matches our assumption for non-bot users. Another interesting find is that bot session's duration varies greatly from each other. The time difference between shortest bot session and longest bot session could be more than 8 hours. It presents us evidence that companies or search engines design their web crawl and rules for spidering very differently. However, human users only have a standard deviation of average session duration for approximately only 10 minutes (629s).

In summary, here are our learnings from our analysis of bot vs. non-bot sessions:

- Approximately 19 – 20% of total sessions are generated by bot activity
- Bot sessions usually spend 4 hours browsing the website and non-bot sessions spend only 45 minutes on the website
- Bot spends less sessions but more time on the website, vise versa for non-bot sessions
- Bot sessions have low variation in number of sessions per day but high variation on time spent on the website, vise versa for non-bot sessions

7.2 Bot Sessions Analysis

Once we have identified distinctive characteristics between bot and non-bot sessions, we would want to investigate them separately. We are interested to know if the website is generating more attention from search engines or companies over time. Will there be a growing traffic over time? We first plotted the number of sessions per day for bot sessions on a chart to see if it has any clear indication of trend or seasonality. Then we tried to fit a trend line using software in order to find the formula of the trend line. Based on our analysis, we didn't find any clear seasonality patters, data is randomly scattered. However, fitting the trend line gives us a clear vision for growing number of bot sessions over the 24-month period.

The daily pattern of bot session activity is also interesting topic for us to investigate. Therefore, we plotted the count of bot sessions vs. hours of the day for all month. The plot shows a clear hourly pattern with the numbers of session peaks up at midnight and steeply dropped to a lower and steady level throughout the day. The seasonality decomposition analysis shows at 12 AM, the number of bot sessions is more than 3 times the hourly session count average. We could easily understand this pattern because developer of robots would like to avoid heavy traffic time for any crawling activity. If the crawling activity is affecting the performance of the website, robots will face the risk of being banned by website administrator.

Operations such as ip address matching and user agent parsing allow us to generate additional dimensions other than date and time. We were quite surprise to see many of bot activities are coming from outside of Canada, mostly from China and USA. Chinese website Baidu.com and 360.com as well as Google and Bing are regular visitors of the website. Moreover, approximately 10% of bot sessions are coming from a mobile user agent. Overall, bot session analysis tell us that:

- Increasing trend of bot visiting the website over time
- Bot is most active between 12 AM and 1AM to avoid human traffic

- Most of bot sessions are generated by user agents outside of Canada

7.3 Non-bot/Human Session Analysis

Plotting of non-bot session data shows a clear cyclic pattern for each week within a month. The website would start the week with a lower number of sessions on Sunday and then the session counts per day gradually increases and peaks on Thursday, after that it goes back down to restart the cycle. Therefore, we believed it will be a good idea to analyze whether non-bot user's behavior will be different between weekdays and weekends. Using the same method as in the bot vs. non-bot anlaysis, we conducted two sample t test as well as the analysis of variance. Both test results on session count and duration are statistically significant.

The descriptive statistics show that there are steadily less sessions on weekend compares with weekdays. In addition, on average, non-bot users spends less time on the website on weekend than weekdays. The duration for non-bot sessions on weekends is around 44 minutes. During weekdays, non-bot users spends 53 minutes on the website. The two sample t test on average session duration is also significant enough for us to conclude that non-bot users in fact spends less time on the websites on weekend.

Through user agent parsing, we were able to identify sessions from users who visited the website using their mobile devices. We would also be interested to investigate if there is any difference between mobile and non-mobile users. Our analysis showed mobile sessions on average spend 21 minutes on the website. This is less than half of the time spent on the website by a non-mobile user. Statistical test shows strong evidence that the mobile user spends less time on the website compares to non-mobile users. This result is intuitively understandable because most people who use mobile device to browse the website are in a rush and wanted to find a quick answer. They don't have the leisure of time to spend more than 40 minutes on properly researching a topic listed on the website. Another interesting result we received from our analysis is that the % of mobile sessions increase steadily throughout the 24-month period. This percentage has quickly grow from less than 10% in 2013 to almost 20% in later 2015.

Similar to what we see from bot sessions, we spotted an increasing trend for non-bot sessions over time as well. The website is attracting more and more session visits steadily throughout this 24-month period. Moreover, we also spotted there is an hourly seasonality for non-bot sessions. It is distinctively different from the hourly seasonality of bot sessions. For non-bot sessions, there is less sessions on average for early morning hours of the day. 2 pm - 3pm usually has the highly number of non-bot sessions visiting the website.

Most of non-bot sessions has ip addresses within Canada and Toronto is the city with highest number of sessions. This result is total within our expectation. But we are surprised to see foreign cities such as Beijing and Gunzenhausen are in the top 10 cities with highest session count. People from these cities could be very interested in moving to Toronto, therefore many of them spent a lot of time researching about the city and its programs. However, we would also have reason to suspect that sessions from these foreign cities are bot sessions. With the size of web log data and the large number of user agents,

it will be impossible to identify all bot user agents. Especially Beijing is also within the top cities with highest bot session count. Google chrome is the most popular browser used by more than 29% of non-bot users and Windows dominates the operating system used by non-bot sessions with more than 65%.

Therefore, our analysis of non-bot session shows us:

- There are less non-bot sessions and less time spent on the website for sessions on weekend
- % mobile session has increased from less 10% in 2013 to almost 20% in 2015
- The website has the highest number non-bot sessions between 2pm and 3pm

7.4 Sequential Pair Association Analysis

The sequential pair association analysis in the result section shows top topic pairs sorted by interest, and indication of the popularity for two topics in all session visit pairs. We have placed a constraint for the joint probability of two topics happening together to be more than 0.5%. From observation, we can find majority of topics shown an independent relationship with Quick Search. It is not a surprising result, as quick search topic serves as a global feature on the website across all topics and the number of sessions uses quick search significantly more than visiting some other topics. The quick search bar sits in dominate position of the website and gains high popularity among users. Because almost every session will use the quick search feature, it makes the combination of any topic with quick search independent of each other. There is no statistical evidence for us to prove a topic will lead to a high probability of using quick search or vise versa.

After filtering out all quick sort related data, we found most of remaining sequence pairs have both $Gain[A \rightarrow B]$ and $Gain[B \rightarrow A]$ positive and greater than 1, which indicates a compliment relationship. It is also a natural visitor behavior, as visitor tends to visit related topics together in a session. And most visitors do not visit related topics in a particular order. In other situations, some of the pairs shows one of the gains slightly less than 1 whereas another greater than one, which indicates an "likely complement relationship" donated as ~COMP in data sheet above (For example, "childfamilyservices" and "employmenttraining"). We can infer that these topics are loosely or implicitly related in some way for some visitors only. Although we only used 3 month's data to conduct our association analysis, we can easily spot users' behavior and interests change from month to month. Certainly topics are consistently ranked at the top of user's interests. Topic 'communituprogram' seems to have the highest interests for pairing with almost any other topics. Many of top interested topic pairs are intuitively related to each other, such as (aboriginalpeoples, incomesupport), (abuseassault, emergencycrisisservices), (childfamilyservices, emergencycrisisservices), etc. They reflect some of social problems and issues within the region. Here are our learnings from this analysis:

- Quick search is independent of all topics
- User's interests vary from month to month
- Top interested topic pairs are intuitively related to each other

8. Conclusion

In this project of 211Toronto.ca web log analysis and visualization, we have utilized the web log data provided by 211Toronto.ca for getting important descriptive statistics and visualizations about the website. The data provided by 211Toronto.ca is generated by Microsoft's Internet Information Service. We were able to extract human visits among the all activities recorded in logs between December 2013 and November in 2015 after a series of transformation and cleaning. Session grouping is performed to group events into sessions using a set of common rules. While performing the session grouping, we have also used external python library to identify activities from a bot or web crawler using the user agent provided in the raw data. IP address matching enable us to add location dimensions such as city, state and country to our data for understanding the geo location of visitors. User agents are parsed to extract browser, operating system and platform information. Last but not least, outliers are removed to ensure the data used is realistic. A set of dimensions and metrics are defined for the ease of data analysis. Then the clean data has been dice and slice to obtain descriptive statistics for different dimension and metrics. The findings show us that the website is serving considerable amount of Toronto residents well over the 24-month period, on average 1138 sessions per month. In addition, the website also receives many international visits from China, US, Germany etc. The most popular browser used is the Google Chrome and Windows is the top used operating system to access the website. Users are spending an average of approximately 45 minutes on the website. Overall, we believe the website is serving its purpose very well, a lot of residents of Toronto and Ontario are benefiting from the social and community information shared on the website. Many international visits might be useful insight for operator of the website to start thinking of translating the website into different languages. Last but not least, visits from different browsers might also suggest website testers to run usability tests against those browsers when they release a new version of website.

Reference

[1] F. I. Service, "Findhelp Information," 2016. [Online]. Available: http://www.findhelp.ca/what-we-do. [Accessed 12 4 2016].

[2] 2. toronto, "211Toronto - About us," 2016. [Online]. Available: http://www.211toronto.ca/basic-page/about-us. [Accessed 12 4 2016].

[3] D. R. D. T. T. M. Heidi Lam, "Session Viewer: Visual Exploratory Analysis of Web Session Logs," in *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, 2007.

[4] Wikipedia, "Wikipedia - logfile," 18 1 2016. [Online]. Available: https://en.wikipedia.org/wiki/Logfile. [Accessed 12 04 2018].

[5] eloone, "How does a web session work ?," 29 10 2013. [Online]. Available: http://machinesaredigging.com/2013/10/29/how-does-a-web-session-work/. [Accessed 23 4 2016].

[6] U. A. String.Com, "List of user agent strings," User Agent String.Com, 2015. [Online]. Available: http://www.useragentstring.com/pages/useragentstring.php. [Accessed 12 4 2016].

[7] www.iplocation.net, "What is an IP address?," www.iplocation.net, 2006. [Online]. Available: https://www.iplocation.net/ip-address. [Accessed 13 4 2016].

[8] Google, "How a session is defined in Analytics - Analytics help," Google2016, [Online]. Available: https://support.google.com/analytics/answer/2731565?hl=en. [Accessed 15 4 2016].

[9] Microsoft, "Learn: The official Microsoft IIS site," Microsoft, 2016. [Online]. Available: http://www.iis.net/learn. [Accessed 15 4 2016].

[10] M. Rouse, "What is IIS (Internet information server)? - definition from WhatIs.com," 10 2008. [Online]. Available: http://searchwindowsserver.techtarget.com/definition/IIS. [Accessed 14 4 2016].

[11] sateesharveti, "Look at logging in IIS 7/7.5," Microsoft, 13 9 2013. [Online]. Available: http://blogs.technet.com/b/sateesh-arveti/archive/2013/09/13/look-at-logging-in-iis-7-7-5.aspx. [Accessed 15 4 2016].

[12] WinSCP, "Introducing WinSCP: WinSCP," WinSCP - Free SFTP, SCP and FTP client for Windows, 1 10 2014. [Online]. Available: https://winscp.net/eng/docs/introduction. [Accessed 15 4 2016].

[13] T. G. S. &. E. J. HOM, "What is A web Bot? - Tom's guide," Tom’s Guide, 2 11 2015. [Online]. Available: http://www.tomsguide.com/us/web-bot-definition,review-1961.html. [Accessed 16 4

2016].

[14] http://www.robotstxt.org/, "The web robots pages," Robots Database, 2007. [Online]. Available: http://www.robotstxt.org/db.html. [Accessed 16 4 2016].

[15] http://www.useragentstring.com/, "List of Crawler user agent strings," User Agent String.Com, 2005. [Online]. Available: http://www.useragentstring.com/pages/Crawlerlist/. [Accessed 17 4 2016].

[16] R. McCann, "Robot-detection 0.3," Python, 2013. [Online]. Available: https://pypi.python.org/pypi/robot-detection/0.3. [Accessed 17 4 2016].

[17] ip-api.com, "IP Geolocation API - batch JSON," ip-api.com, [Online]. Available: http://ip-api.com/docs/api:batch. [Accessed 17 4 2016].

[18] J. Massey, "HTG explains: What's a Browser user agent?," http://www.howtogeek.com/, 27 4 2015. [Online]. Available: http://www.howtogeek.com/114937/htg-explains-whats-a-browser-user-agent/. [Accessed 17 4 2016].

[19] S. Tiwatne, "Httpagentparser by shon," [Online]. Available: Httpagentparser by shon. [Accessed 17 4 2016].

[20] U. A. String.Com, "User Agent String explained," User Agent String.Com , 2005. [Online]. Available: http://www.useragentstring.com/index.php. [Accessed 17 4 2016].

[21] W. A. Kamakura, "Sequential market basket analysis," *Marketing Letters,* vol. 23, no. 3, p. 505, 2012.

[22] M. A. Poolet, "Data warehousing: Dimension basics," Mount Vernon Data Systems, 13 9 2007. [Online]. Available: http://sqlmag.com/business-intelligence/data-warehousing-dimension-basics. [Accessed 17 4 2016].

[23] M. Rouse, "What is business metric? - definition from WhatIs.com," SearchCRM, 6 2015. [Online]. Available: http://searchcrm.techtarget.com/definition/business-metric. [Accessed 17 4 2016].

[24] E. Franz, "Using SQL to define, measure and analyze user sessions," Segment Blog, 29 1 2015. [Online]. Available: https://segment.com/blog/using-sql-to-define-measure-and-analyze-user-sessions/. [Accessed 16 4 2016].

[25] J. Cutroni, "Hits, Sessions & Users: Understanding Digital Analytics Data," 5 FEBRUARY 2014. [Online]. Available: http://cutroni.com/blog/2014/02/05/understanding-digital-analytics-data/.

[26] Minitab, "How does Minitab do decomposition?," 2016. [Online]. Available:

http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/time-series/basics/how-does-minitab-do-decomposition/.