SINGAPORE MANAGEMENT UNIVERSITY

Recommendations to Improve Content Viewership Yield for Skyscanner

Team SkyTrek:

Aseem PRABHAT
Jedaiah TAN Jia Le
NGUYEN Viet Huy

ANLY482

Sponsor and Background Information

Skyscanner is leading global travel search site offering an unbiased, comprehensive and free flight search service as well as online comparisons for hotels and car hire.

Skyscanner's flexible search options allows users to browse prices across a whole month, or even a year, allowing users to get the best deals. When you find the perfect deal through Skyscanner, you are redirected to book direct with the airline or travel agent, so you get the lowest price, with no extra fees added.

Skyscanner has been in the travel business for over 10 years, and employs more than 50 different nationalities in its global offices in the Edinburgh, Singapore, Beijing, Shenzhen, Miami, Barcelona, Glasgow, London, Sofia and Budapest. It has over 50 million unique visitors per month who use it to find flights, car hire and hotels in more than 30 different languages.

On its website, Skyscanner has a travel features and news section[1]. This helps attract users to Skyscanner through its content marketing activities. The company constantly publishes news articles relating to travel trends, travel tips, top destinations, best deals and new product features in order to constantly engage its users drive more traffic to the site. The project sponsor is the content manager for APAC at the Skyscanner Singapore office.

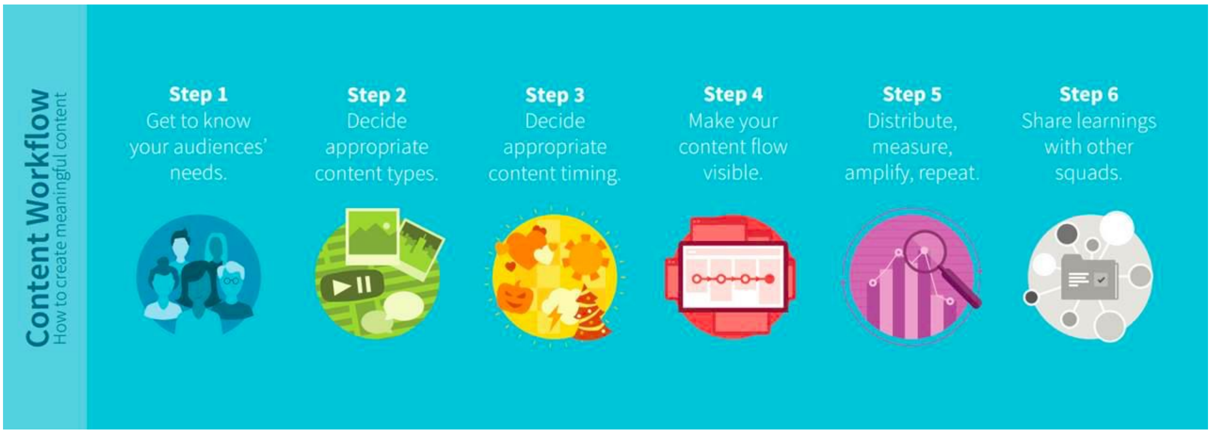
Motivation

One on Skyscanner's goals is to acquire new users and engage its current users through content marketing on its travel features and news site. The goal of the company is to drive more users to the website in order to increase its metric of unique monthly users. This metric has a large impact on revenues as well as the valuation of Skyscanner and similar internet companies. It has been growing at a high quarter on quarter growth rate over the last 2 years and Skyscanner wishes to maintain this high growth rate.

As a lean organization, Skyscanner has limited resources for content marketing and hence must use resources in a way to maximize impact. This impact is measured through page views and engagement metrics. Skyscanner believes in the idea of "Build. Measure. Learn" and hence is constantly conducting experiments such as A/B tests in order to reevaluate and improve its processes.

[1] http://www.skyscanner.com.sg/news/

The Content team has similarly been moving towards a data driven approach over the last year but there is still a lot of room for improvement. Below is the new process flow for content creation. This process is constantly improved through experiments, feedback and learnings.


Content Workflow
How to create meaningful content
Step 1
Get to know your audiences' needs.
Step 2
Decide appropriate content types.
Step 3
Decide appropriate content timing.
Step 4
Make your content flow visible.
Step 5
Distribute, measure, amplify, repeat.
Step 6
Share learnings with other squads.

Objectives

The aim of the practicum is to provide deeper insight into the performance of different content pieces on the Skyscanner travel and news features site.

The client is a content manager who intends to use the results from our analysis in the content planning process in order to make the most optimal use of resources. This will help decide what kind of content is to be created at different times of the year in order to maximize the number of visitors to the Skyscanner news site.

The final deliverables will aim to:

1. Identify the different web content factors that affect content performance in order to differentiate between high and low performing content
2. Facilitate the content planning process by way of an interactive dashboard

In order to achieve these high level objectives, the project will aim to explore some specific questions in order to find new insights within the content database as well as validate previously held intuitions about the performance of different types of content.

The approach to this would be to formulate a list of hypotheses that would be tested through the different stages of our analysis. A high level overview of each hypothesis is provided in the form of unanswered questions relating to content performance. Some of these questions that will be explored over the course of the project are:

1. What are the most popular content themes that resonate with users in each of the given markets namely Singapore, Malaysia and Thailand?
2. What are the common attributes that lead to some content pieces drawing in most of the traffic?
3. Should the strategic focus be on generating more articles to drive traffic or focus on fewer quality articles?
4. What is the role of seasonality and annual trends in the online readership pattern of Skyscanner users?
5. How can the content planning process be streamlined in order to create maximum impact with minimal resources?
6. Is there an ideal standard format for news articles that best caters to the needs of the users?

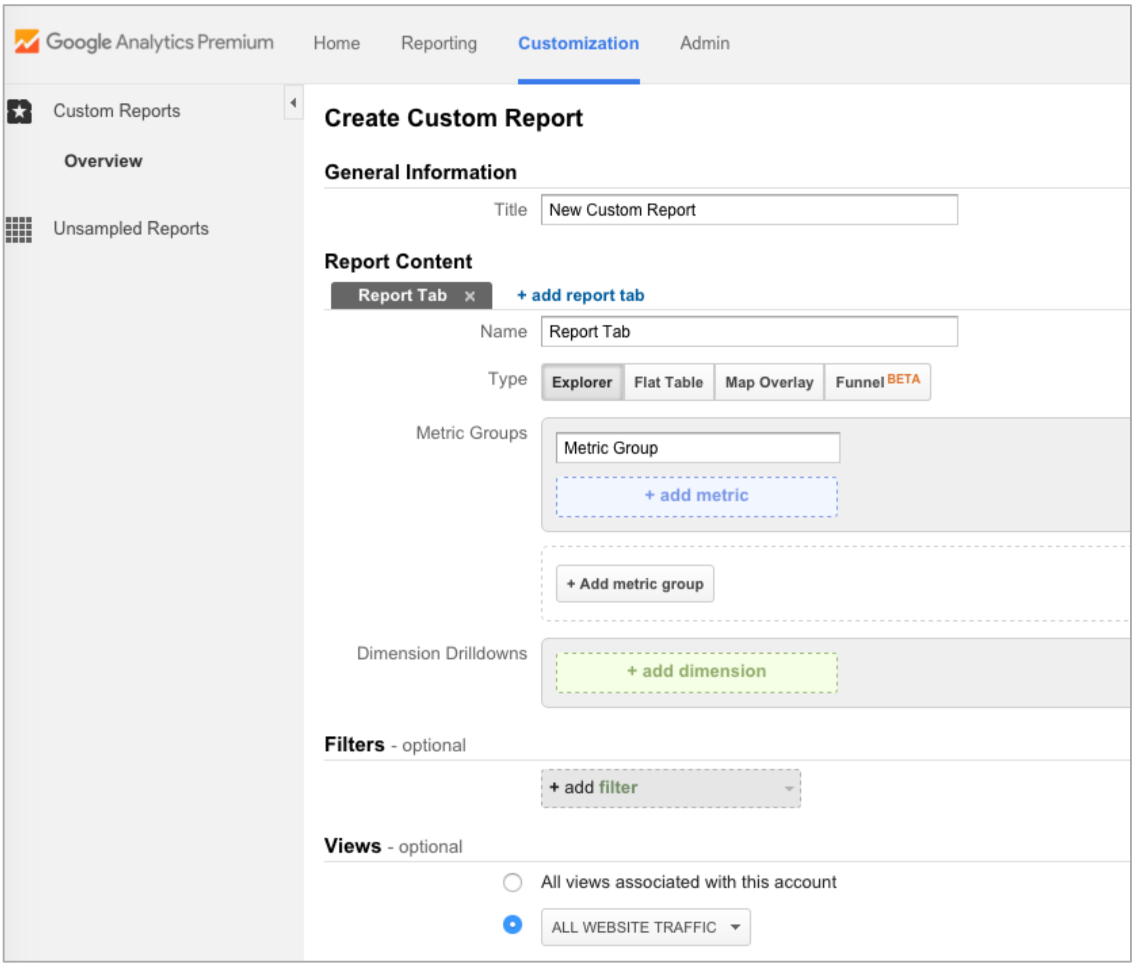
Data (Provide metadata and sample dataset(s))

1. Dataset provided by Skyscanner

Skyscanner provided our team access to their Google Analytics platform, allowing us to pull data about the performance of articles posted on their website. Tracking parameters include:

- Unique page views
- Page views
- Average time spent on page
- Exit rate
- Bounce rate, which measures the occurrences of people coming to Skyscanner website and exit after just reading one page.

Skyscanner uses a Google Analytics premium account in order to track user behavior across all its webpages globally. The Singapore team has created an access account for this project, allowing us to pull all possible combinations of data from Google Analytics relating to the Skyscanner news site. This is also summarized in the form of different dashboard views available on Google Analytics premium. The method for querying any data is through creation of custom reports.



Create Custom Report

General Information

Title: New Custom Report

Report Content

Name: Report Tab

Type: Explorer

Metric Groups: Metric Group

Dimension Drilldowns

Filters - optional

Views - optional

○ All views associated with this account

● ALL WEBSITE TRAFFIC

Custom Report creation process on Google Analytics

2. Crawling

The data provided by Skyscanner is what they currently rely on to determine their content planning. In addition to the parameters mentioned above, our group also want to track various other characteristics of the articles, namely:

- Number of words (stop words removed)
- Number of outbound links references
- Number of images
- Number of videos
- Number of article shares

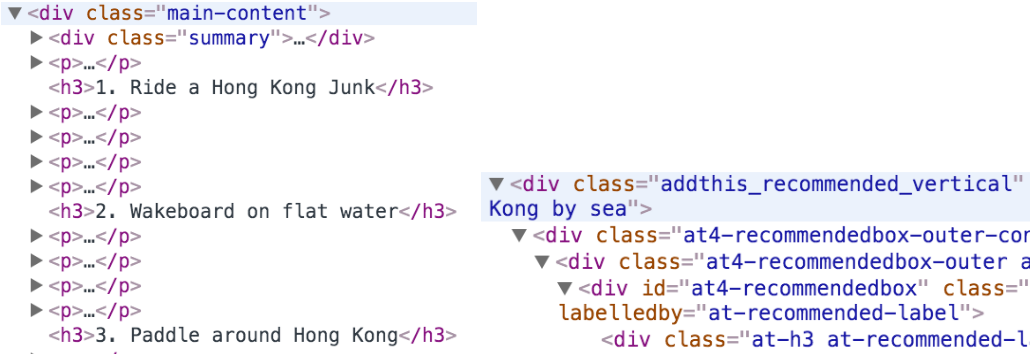
The rationale for choosing these attributes will be mentioned in the methodology section of our proposal.

These attributes are not available for us. Thus we need to manually crawl the data from the website and merge it with the provided data set.

We noticed that Skyscanner website use Javascript to add content to the DOM when the HTML finishes loading. Because of that, normal crawlers without ability to execute Javascript code will not be able to crawl for data within the page after the DOM is modified.

After some research, our group decided to employ the use of a headless browser, namely PhantomJS. It allows for Javascript code execution, DOM access, and programmatically interaction with websites without opening any real web browser. Another option is Selenium, but it needs to proxy through a browser installed on our machine.

Looking through the DOM structure of Skyscanner article, we found that the information we need is easily accessible. For example, the main article content is nested in a div block with CSS class “main-content”, and links to recommend other articles are provided in another div block with CSS class “addthis_recommended_vertical”. The repeating structure makes it easy for us to write code and scrap data from the website without too much trouble.



```
▼<div class="main-content">
  ▶<div class="summary">…</div>
  ▶<p>…</p>
   <h3>1. Ride a Hong Kong Junk</h3>
  ▶<p>…</p>
  ▶<p>…</p>
  ▶<p>…</p>
  ▶<p>…</p>
   <h3>2. Wakeboard on flat water</h3>
  ▶<p>…</p>
  ▶<p>…</p>
  ▶<p>…</p>
  ▶<p>…</p>
   <h3>3. Paddle around Hong Kong</h3>
```

```
▼<div class="addthis_recommended_vertical"
Kong by sea">
  ▼<div class="at4-recommendedbox-outer-cor
    ▼<div class="at4-recommendedbox-outer a
      ▼<div id="at4-recommendedbox" class="
      labelledby="at-recommended-label">
          <div class="at-h3 at-recommended-l
```

Figure 1.1 DOM structure of main content and outbound links

After successfully scraping the DOM data, we can clear out HTML tags easily using a regular expression /<(\/|).+?>/g, then proceed to compute the necessary attributes that we want to collect.

3. Merging of crawled data with Skyscanner provided data

Since the data is provided for each URL, we can easily match the URL between the data given by Skyscanner and the characteristics crawled by us. Thus, we will have a list of attributes mapped to URLs of each specific article.

4. Storing data

Our data needs to be saved in a convenient format so that we can use it as input for other analytic programs.

An option for fast querying is storing the data in a database. This approach provides easy export to other formats that can work with analytic software, and access from both a GUI and code.

Another option is to store data in flat files for easy transport between systems. However, it will reduce accessibility since our code and program need to parse the information again.
With pros and cons in mind, we will proceed with the database approach initially, and make changes as the the project continues.

Methodology

The following table demonstrates the analytical methods proposed for use, in order to achieve our objectives for this practicum.

Objective	Analytical Method(s)
Identify the different web content factors that affect content performance in order to differentiate between high and low performing content	● Multiple Linear Regression on Article Characteristics
Facilitate the content planning process by way of an interactive dashboard	● Data Visualization ● Google Trends Analysis ● Content Themes Analysis

Multiple Linear Regression on Article Characteristics

Based on the merged dataset comprising of attributes from Google Analytics and article attributes scraped directly from the new articles, we will be performing multiple linear regression (MLR) to determine key attributes affecting the number of unique page views.
We will be exploring the following dependent variables in predicting the number of unique page views:

Independent Variable	Intuition for Selection
No. of words (stop words removed)	This measure serves as an indicator of the length

	of the article. Recognizing that readers have a limited attention span, it would be interesting to explore the effect of a lengthy article on its popularity.
No. of outbound links references	Outbound links typically direct readers to more in-depth content. An article with more links might be indicative of more meaningful content, which might translate to greater popularity and better reception amongst its readers.
No. of images	Images and videos make for a more interactive experience with the reader. It might be an important determinant in an article's receptivity.
No. of videos	
No. of article shares	The intuition is that readers share articles that are useful and impactful. Number of article shares is expected to have a positive correlation with the number of unique page views. It would be of interest to assess its importance, hence making an assessment of the importance of social media as a platform of publicity in comparison to other platforms.
Bounce rate *(Percentage of sessions that starting with the page (out of all the other tracked Skyscanner pages) where the reader leaves after visiting the page (i.e. one page views))*	Readers arriving at Skyscanner's news pages are expected to be browsing for information related to a particular destination or related travel content. Since Skyscanner articles are light (bit-sized) reads, we would expect readers to continue browsing other relevant articles via the recommendation engine or the outbound links within the articles themselves. Nevertheless, there will bound to be a point where readers finally exit the site. Hence, we are expecting to see an average bounce rate and exit% rating across the articles. Articles with particularly high ratings would serve as good negative-subjects of study for future
Exit % *(Percentage of sessions involving the page where the reader leaves after reading the page)*	

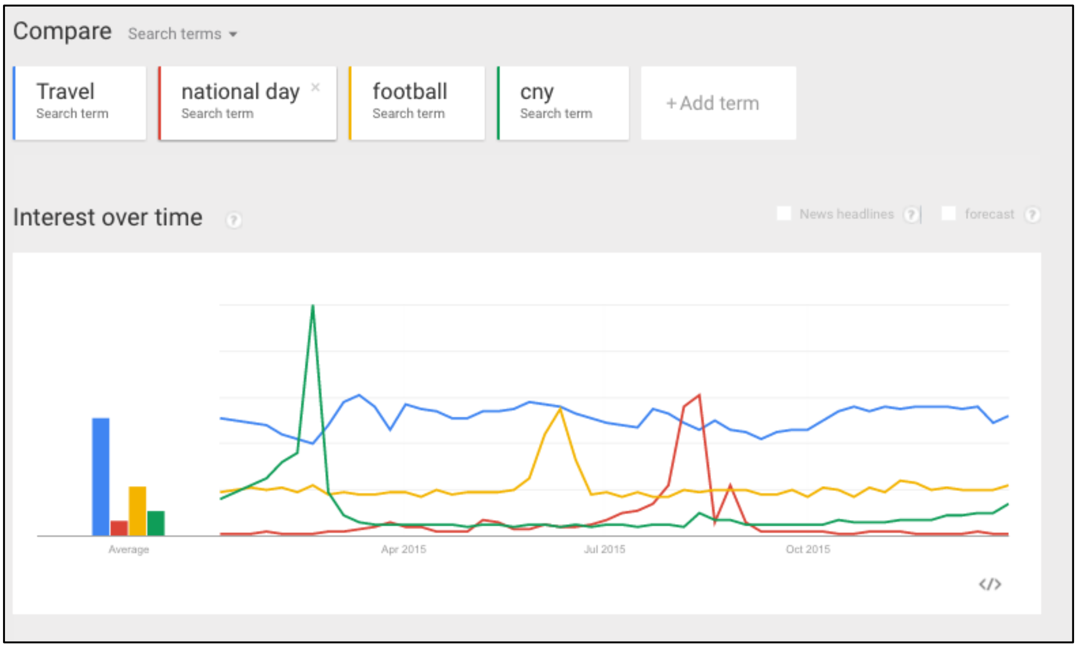
	reference.
Average time on page	Time spent on a page is expected to be indicative of interest levels in an article and possibly the number of unique page views. It would be interesting to validate if time spent is a predictor of unique page views. If so, we could pursue 2 routes: 1. Study articles with long average times to identify the characteristics of a good article 2. Look at techniques to increase average time spent on article pages

Understanding key dependent variables which influence the value of the unique page views will help in the creation of content which have greater tendency of receiving higher page views.

Google Trends Analysis

In planning the content for the upcoming quarter, the content management team typically uses Google Trends to understand consumer trends in both past similar quarters as well as the present. They would also consider the present context of festivities and events. A word cloud of Google trends relevant to each quarter will help incorporate these trends into the content planning process.

This will tie up with our exploration of seasonality and the effect of external events on the content readership. While Google Trends does not have an API, the data can be scraped through manipulation of the URL. This trend data will be aggregated and put into word cloud and put side by side with the quarterly patterns of the different Google Analytics metrics in order to gain a better understanding of seasonality.



Sample of Google Trends data

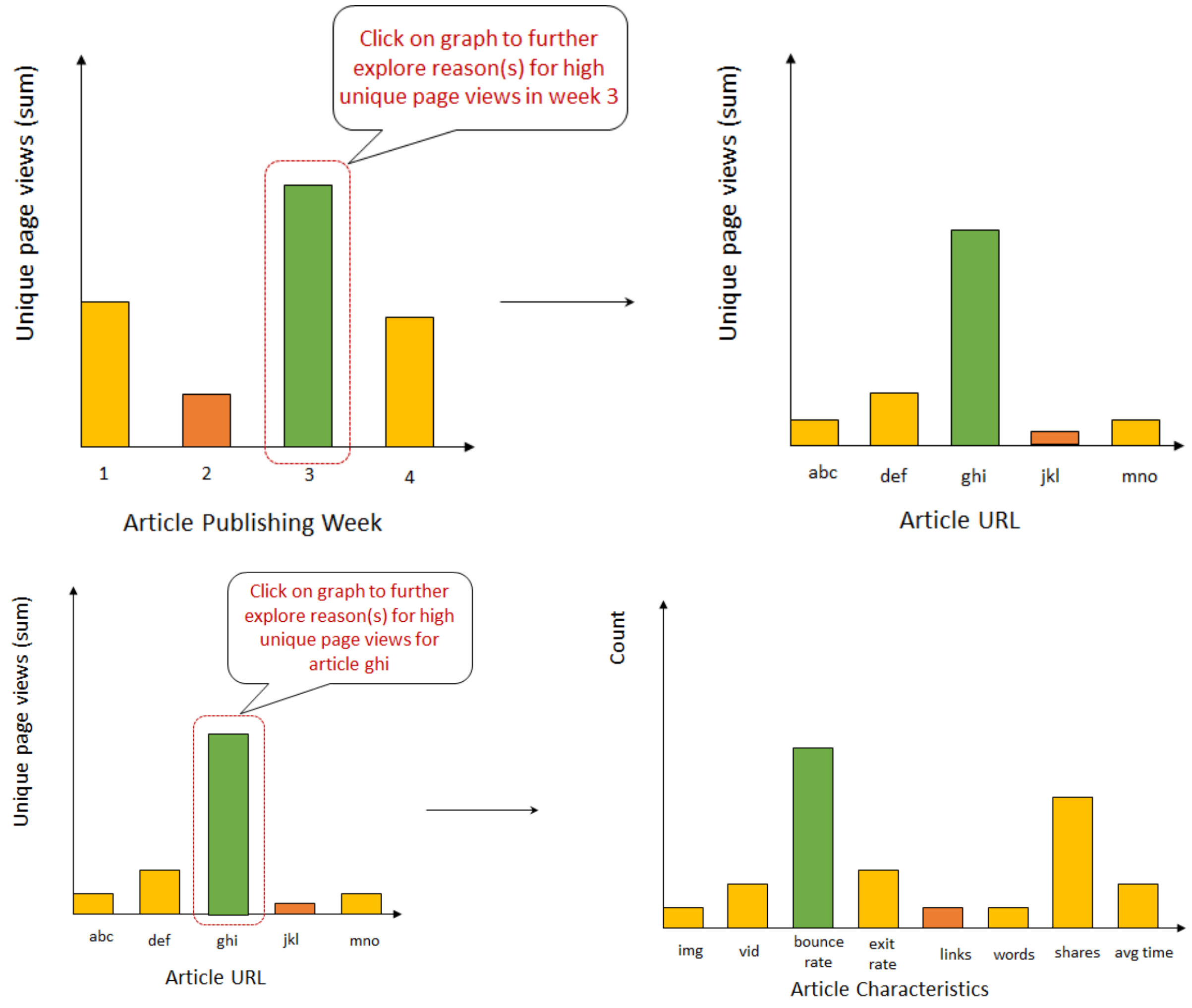
Content Themes Analysis

Skyscanner has identified 7 content themes articles typically belong to. Operating on a lean workforce, it would be helpful to be able to identify which of the 7 content themes reaps the greatest yield. Here, we define yield by the metrics Google analytics tracks. They are the number of unique page views, bounce rate and exit %, as well as the average time spent on page. This will be done via Text Miner by SAS.

Text Miner can generate a number of topics. Each topic will be associated with a set of representative keywords derived from the corpus of articles input to the algorithm. Each article would have a probability rating of belonging to a particular topic. We would tag the topic with the highest probability rating to the article. We would then manually examine the keywords representative of the topic, then classify the topics according to the 7 content themes. Having classified the articles into the 7 content themes, we can then analyze them with the google analytics metrics, thereby identifying popular content themes as an area of focus for the content delivery team.

Data Visualization

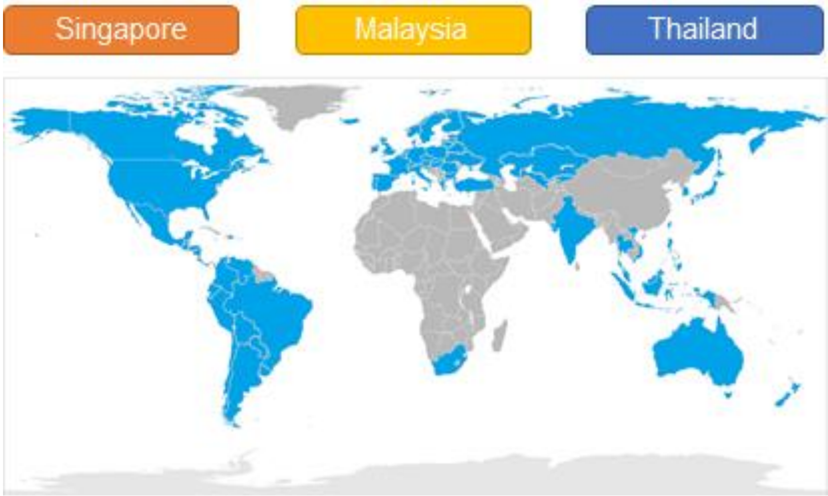
Unique Page Views Exploration



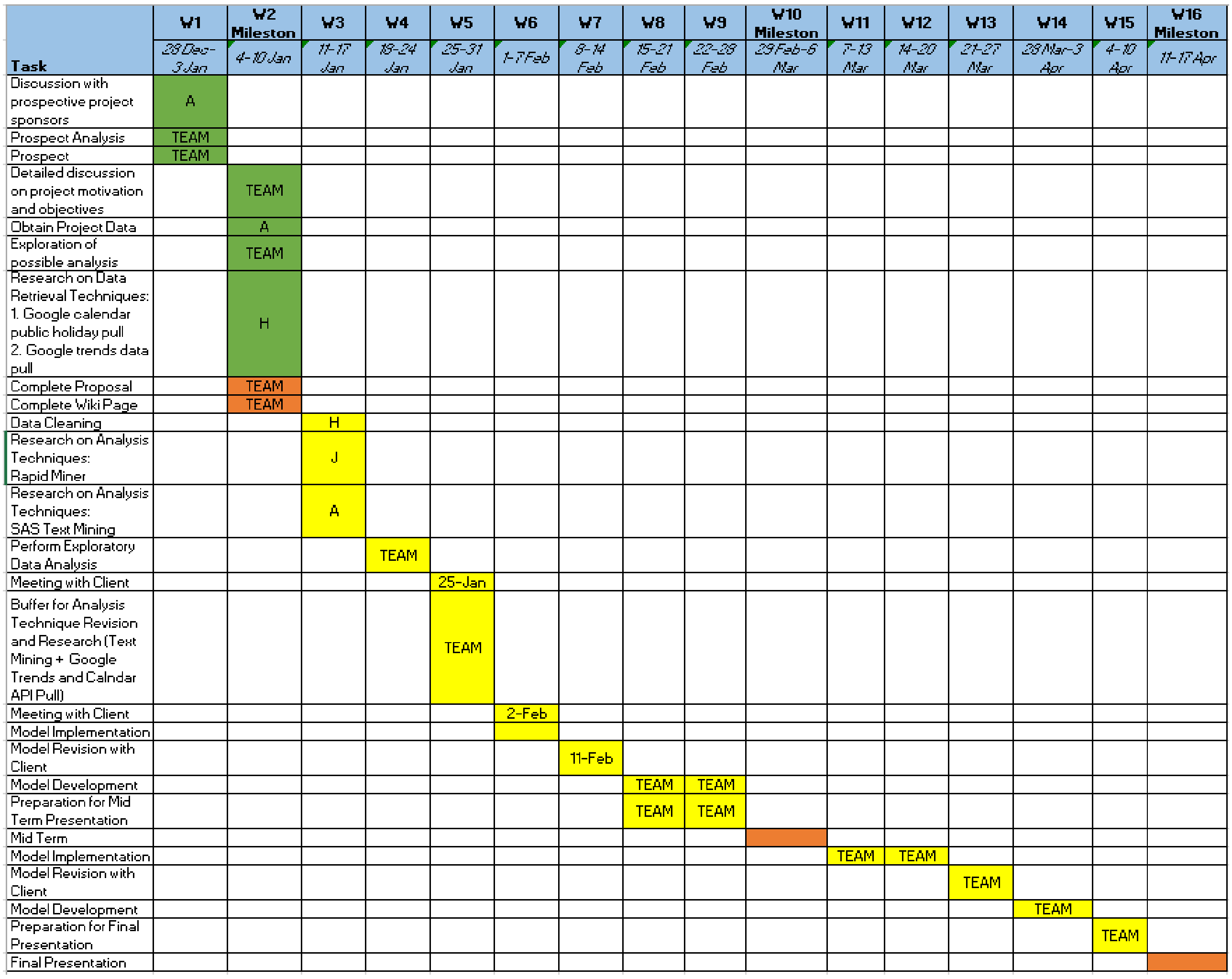
Click on graph to further explore reason(s) for high unique page views in week 3

Click on graph to further explore reason(s) for high unique page views for article ghi

Heat Map of Traffic Source (Country Specific New Page)



Scope of Work



Task	W1 28 Dec-3 Jan	W2 Mileston 4-10 Jan	W3 11-17 Jan	W4 18-24 Jan	W5 25-31 Jan	W6 1-7 Feb	W7 8-14 Feb	W8 15-21 Feb	W9 22-28 Feb	W10 Mileston 29 Feb-6 Mar	W11 7-13 Mar	W12 14-20 Mar	W13 21-27 Mar	W14 28 Mar-3 Apr	W15 4-10 Apr	W16 Mileston 11-17 Apr
Discussion with prospective project sponsors	A															
Prospect Analysis	TEAM															
Prospect	TEAM															
Detailed discussion on project motivation and objectives		TEAM														
Obtain Project Data		A														
Exploration of possible analysis		TEAM														
Research on Data Retrieval Techniques: 1. Google calendar public holiday pull 2. Google trends data pull		H														
Complete Proposal		TEAM														
Complete Wiki Page		TEAM														
Data Cleaning			H													
Research on Analysis Techniques: Rapid Miner			J													
Research on Analysis Techniques: SAS Text Mining			A													
Perform Exploratory Data Analysis				TEAM												
Meeting with Client					25-Jan											
Buffer for Analysis Technique Revision and Research (Text Mining + Google Trends and Calndar API Pull)					TEAM											
Meeting with Client						2-Feb										
Model Implementation						🟨										
Model Revision with Client							11-Feb									
Model Development								TEAM	TEAM							
Preparation for Mid Term Presentation								TEAM	TEAM							
Mid Term										🟧						
Model Implementation											TEAM	TEAM				
Model Revision with Client													TEAM			
Model Development														TEAM		
Preparation for Final Presentation															TEAM	
Final Presentation																🟧

References

Hacking the Google Trends API. (2014, September 25). Retrieved January 9, 2016, from http://techslides.com/hacking-the-google-trends-api

Phantom JS (http://phantomjs.org/)

RapidMiner (http://docs.rapidminer.com/)

Selenium HQ (http://www.seleniumhq.org/)