Meeting minutes

- Team SkyTrek

Date (dd.mm.yyyy)	03.02.2016
Attendee(s)	Jedaiah, Aseem, Huy
Absentee(s)	

Contents

Item	Discussion point	Remark
1	- Dataset to pull from Google Analytics	- URL level with sources - URL level without sources - URL level by organic - URL level by paid - Source level paid - Source level organic - Source level aggregated
2	- Organic Growth Analysis	Regression Input variables from 3 classes: 1. Titles (mapped to content types, then converted to dummy variables, e.g. isCT1_, isCT2, etc) 2. imgs/videos/shares/length of article 3. phrases (from editors)
3	Paid / Advertised Growth Analysis - Is Skyscanner paying for the right source/medium? - Even if the analysis shows the profitable source/medium to invest in, how do we know the result is not influenced by a viral article?	Look at source dimension aggregated data and ask: • Which is the most valuable (UVP and ATOP) source? • Sort and rank by UVP and ATOP • Apply pareto principle (80/20) and identify the most effective and least effective sources Look at source/URL dimension aggregated data (drill down) and plot the distribution graph to verify distribution is not overly skewed. If skewed, 1. look into the bias data points for more exploratory analysis 2. Possibly remove bias data points and re-identify most important source 3. then proceed to 'if not skewed' case If not skewed 1. map articles to content types 2. for each source (within pool of shortlisted questions from analysis1) , rank by content types to identify most valuable content type to pay for

		3. for each content type, rank by source to identify which source is more effective for a content type 4. Hence ask question: what is the cost of switching?
4	Methodology - Odds ratio - K Mean Clustering	- Problem: unable to quantify impact (on Unique Page View & Average Time On Page) of switching between options (FB vs Twitter vs etc.) or within a category - Solution: • Classification of two target metrics- UPVs and TOP into binary form • Comparison of effect of categorical variables, i.e. Source, Content theme on the binary variables • Odds ratio will help compare the advantages of switching between options within a category Objective: - Identify Content Theme (CT) types ○ Since clusters are generated based on the words within the articles themselves - Validate client's set of CT types ○ new CT type will only be proposed if there are sufficient articles within the identified cluster(s) to be used to generate the CT type - Map into respective CT types ○ Compare metrics on CT type level against population mean ○ Perform regression analysis to identify important independent variables within the CT level

Actions

Task	Assigned to	Status
Pull Google Analytics Data	Aseem	Created
Manual classification of content types based on heavy weightage terms generated from k means model	Jedaiah	Created
Explore details of incorporating categorical and numerical data into regression model	Huy	Created
De-aggregate shares count for each URL	Jedaiah & Huy	Done
Build regression model based on current dataset	Aseem	Done
Come up with list of possible title categories to clear with sponsor	All	Done
Clarify with Prof Kam: Is there a way to dynamically assign new articles to the relevant clusters based on the trained model?	All	Done
Characteristics to scrape from article: 1. Number of words (remove stopwords) 2. Number of links 3. Images 4. Videos 5. Number of shares of article (by Jed)	Huy	Done
Retrieve URL share count	Jedaiah	Done
Put topic modelling and google trends research and implementation into Gantt chart timeline	Jedaiah	Done
Input meeting with client in gantt chart for: 1. 25th Jan 2. 2nd Feb 3. 11th Feb	Jedaiah	Done
Create flow chart documenting process how skyscanner team plans for the next quarter's content postings. To send it to client for verification	Aseem	Done
Test Google API public holiday pulling	Huy	Done
Create a mockup for the following visualization: o To plot Unique page views views (count) vs Week of article publishing o Clicking this particular bar of interest (week has high count) would drill into the next bar chart (Number of unique page views, vs identifying URL) o Further clicking on the URL would reveal characteristics of the article (e.g. number of images, videos, etc, bounce/exit rate, etc)	Jedaiah	Done
Create a mockup for the following visualization: o Identify key features of popular posts (MLR against unique page views)	Huy	Done
Create a mockup for the following visualization: o Where do they come from (On the whole country news page level, where do the traffic come from) o How long do they spend at the page (average time on page) o Do they visit other pages? (bounce rate vs exit rate)	Jedaiah	Done

○ How effective are the recommended pages? (If GA does not track the unique page views, then this analysis is not going to be implemented) ○ [Just to prove a logical point] Correlation test between number of shares vs unique page views		
Deal with proposal segments: 1. Sponsor and Background Information 2. Motivation 3. Objectives 4. Data 5. Scope of Work	Aseem	Done
Questions for Prof Kam: 1. Ask Kam about requirement to post sample date. Concern: data confidentiality 2. What is required for scope of work? (different from methodology)	All	Done

---

The minutes were prepared by: Jedaiah Tan and Nguyen Viet Huy