ANLY482
Analytics Practicum

Project Proposal

Prepared for Prof. Kam Tin Seong

Group 20 – MaxiMuM

Members:

Mahima Damani

Manisha Goyal

Mohammed Lukman Bin Mohamed Syed

Overview

Situated at the heart of SMU's city campus, SMU Libraries offers its services through two physical libraries, Li Ka Shing Library and Kwa Geok Choo Law Library as well as virtually through various web based services. SMU Libraries is the centre for academic and professional knowledge resources and services that support the research and learning needs of the SMU community. The library facilitates knowledge creation, discovery through its electronic search platform and access to a wide array of research resources on and off campus. [1]

Motivation

Our project sponsor is the SMU Libraries Analytics and Research Department. Their aim is to constantly derive insights from library usage patterns to drive the decision-making process of higher management.

The library currently aims to optimise its resource availability and distributions channels to maximise the learning effectiveness of its students. This could be in terms of increasing resources available for certain highly searched topics, altering current trainings and workshops to focus on any common mistakes committed by students while using the assets or finding any unexpected trends in user journey through digital and physical touchpoints. They further want to know if usage patterns vary between students based on certain attributes like Programme, Year of Graduation and Education Level. For this purpose, they have conducted an initial survey for the freshman batch of 2017 to evaluate the difference in their confidence level in various research skills before and after joining SMU, factoring in several considerations like modules taken, library workshops attended and so on and so forth. They wish for us to understand if this survey contains any actionable insights.

Objectives

We had an initial discussion with our project sponsor and they would like us to create a visual dashboard to ascertain cause and effect relationship between the initiatives and resources of the library, and student performance (in terms of confidence and optimal usage of resources).

The objectives of the project would be of the following:

1. Analysis of the Library's proxy server logs (usage of online resources) and turnstile logs (usage of physical resources)
    a. Initial exploratory analysis to identify clusters in usage patterns
    b. To understand if these patterns affect confidence level of students (limited to freshmen)
    c. Alternatively, determine if confidence level spurs certain search behaviour.
2. Recommendations based on findings
    a. To help stakeholders understand the analysis of the findings
    b. To validate or suggest changes in current workshops and trainings
    c. To validate or suggest changes in the current availability of resources

The current objectives may be subjected to further changes after we have obtained and look at the actual data.

Data (The data is withheld until we signed the NDA agreement)

The sponsor has provided us with five datasets - student data, request log data, turnstile data, and pre and post survey data.

The student dataset contains information about the current students of SMU across all batches. The record attributes are the following:

- email (hashed to a 64-digit-long hexadecimal number for non-disclosure reasons)
- education level
- faculty
- admission year
- graduation year
- degree program

The request log dataset contains records captured by the library's URL rewriting proxy server throughout the year of 2017. This dataset captures all user requests to external databases. The record attributes are the following:

- user ID
- session ID
- search database
- timestamp
- search query

The turnstile dataset contains records captured by the library's gantries throughout the year of 2017. This dataset captures physical taps on the gantries of the library. The record attributes are the following:

- date
- time
- device name
- email (hashed to a 64-digit-long hexadecimal number for non-disclosure reasons)

The pre and post survey dataset contains responses of students before and after the first semester of freshman year on their confidence level in various research skills. Some of the record attributes are as follows:

- email (hashed to a 64-digit-long hexadecimal number for non-disclosure reasons)
- school
- modules taken
- library workshops attended

Methodology

As we have not obtained the data until the NDA is signed, we will only share our initial thought process of how we will tackle the project. We shall adopt closely to the Data Analytics Lifecycle approach.

Our plan of action is to discern the effectiveness of the library eBook databases in meeting the research needs of students. By analysing the proxy entries, we can define the usage pattern of its users and divide them into distinct clusters based on demographic and behavioural traits. Furthermore, we intend to track student user journey once they start

interacting with the several physical and digital touchpoints sequentially. As such, we have also conducted a secondary research from various university published articles to gain a broad understanding of now turnstile and proxy data could be used to draw insights (Referenced below). [2] [3] [4]

At this phase of the project, we will focus on understanding the given dataset and clean the data. Concurrently, we will decide on the analytical model and prepare the data accordingly.

Project Scope

While the project will be primarily focussed on answering the questions mentioned above, our client has been supportive enough to let us experiment with different analytical tools and present any other significant insights we derive.

- We will be unable to conduct a yearly or seasonal analysis as the dataset is limited to records from 2017 only.
- The dataset pertains to all students of SMU who used the library resources in the said time-period. However, the survey was only conducted for the freshman batch.

Work Plan

The below gantt chart shows our draft timeline for our project. It may be subjected to changes, especially after week 4 when we reviewed our project progress with the project sponsor.

	Tasks	Task Allocation	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15
	Discovery																	
1	Project research	Mahima, Manisha, Lukman	✓	✓														
	Proposal Submission																	
2	Upload Proposal document	Manisha	✓	✓														
3	Create a Wiki page	Mahima, Manisha, Lukman	✓	✓														
Milestone 1: Project Proposal Submission																		
	Administrative Matters																	
4	Sign NDA	Mahima																
5	Obtain data from Sponsor	Lukman																
	Data Preparation and Exploration																	
6	Data consolidation	Mahima, Manisha, Lukman																
7	Data cleaning	Mahima, Manisha, Lukman																
	Model Planning																	
8	Data exploration	Mahima, Manisha, Lukman																
9	Model selection	Mahima, Manisha, Lukman																
	Sponsor Review																	
10	Review findings with sponsor	Mahima, Manisha, Lukman																
11	Finalise project objectives and scope	Mahima, Manisha, Lukman																
	Mid Term																	
12	Finalise interim report in Wiki page	Lukman																
13	Prepare interim presentation	Manisha, Mahima																
14	Finalise analytical sandbox	Lukman																
Milestone 2: Mid-Term Deliverables Submission																		
	Model Building																	
15	Model calibration	Mahima, Manisha, Lukman																
16	Model assessment	Mahima, Manisha																
17	Interpration of analysis results	Mahima, Manisha, Lukman																
	Final Preparation																	
18	Prepare conference paper	Mahima, Manisha, Lukman																
19	Prepare final presentation	Lukman																
20	Update Wiki page	Mahima																
21	Prepare poster	Manisha																
	Sponsor Review																	
22	Update project progress with sponsor	Mahima, Manisha, Lukman																
23	Finalise deliverables with sponsor	Mahima, Manisha, Lukman																
Milestone 3: Final Deliverables Submission																		

References

1. SMU. (2017, April 11). About Us - Overview. Retrieved January 13, 2018, from https://library.smu.edu.sg/about-us-overview
2. LU, Ning; SONG, Rui; HENG, Dina; GOTTIPATI, Swapna; TAY, Chee Hsien Aaron (ZHENG Zhixian); and TAY, Aaron. Using data analytics for discovering library resource insights – Case from Singapore Management University. (2017). 1-8. Research Collection School Of Information Systems.
3. J. (2015, June 26). Cleaning NYC Turnstile Usage Data. Retrieved January 14, 2018, from http://www.jbencina.com/blog/2015/06/25/cleaning-nyc-turnstile-usage-data/
4. Adey, H., & Eastman-Mullins, A. (2017). User engagement analytics case study: how customer behaviour can drive intelligent library decision making. *Insights the UKSG journal,30*(3), 138-147. doi:10.1629/uksg.387