ANALY482 MEETING MINUTES WITH SUPERVISOR(S)

Date:	18 January 2017
Time:	14:30 – 15:00
Venue:	Meeting Rm 4-1 (School of Information Systems)
Attendees:	Lu Ning, Song Rui, Dina, Prof Kam
Absentees:	
Agenda:	Finding out if our approach meets Prof Kam's requirements for the project

S/N	Things Discussed/Done	Remark

1.	Updates about our domain summary and methodology	1. Limitation is not referring to data we have
		  1. show the example why we have this kind of limitation
		
		2. Raw data are not made into data file
		  1. Lu Ning: organize data into separate txt files base on the domain name
		  I. Prof: need to translate raw data file into a data table/ extract them into different column that capture certain characteristics, URL, time spend, etc. know how to dealt with missing value. Then will be able to do exploratory (summary frequency count, distribution…)
		  II. based on queries, find a way to put them into column
		
		3. Prof: Capture both behavior (pdf is being viewed online), some people only read pdf online. But some will download directly. Later then see how we can decide on the analysis. (don't eliminate possible behavior first, find first then drop later)
		
		4. prof: Review the URL if we can differentiate view online or actual download. (okay)
		
		5. we can see the link but cannot re-query. this analysis is very unstable. Point to individual publisher via proxy.
		  1. Prof: The current system may create grey areas for the analysis. (proxy on a proxy)

		6. Prof: bench marking is not really a feasible solution. (e.g. same bank on different location cannot reapply the model been built) 7. Proxy Approach - Downloads 1. Prof: URL will lead you back to the publisher. some publisher is expired. 2. Prof: two sets of data (pdf vs non-pdf) cannot draw the conclusion on which set to choose from. PDF has 3 sub-groups there. We should consider each group on their detailed behaviors (e.g. The third bar, what are the characteristics? To determine if it is a noise, maybe is a scan image?) 3. Prof: Take all collection as the base, then look at the file size and draw a fairer conclusion. 4. Prof: There is way to match the URL that (replace the front part of the URL) (via manual sampling to find out the pattern of the) 8. Prof: Do not support (pattern approach), because before we explore our situation how our lib been used, we cannot assume we can use other people approach. We should do a proper understanding our situation first, if ours are identical to theirs, then we can follow their approach. But if our approach is really contradicting, then we must change.

Item Due (Team)/Action(s)
Deadline: 8 February 2017 1. Inform Aaron on our revised methodology