









# Meeting Notes - 4 Feb 2022



## Meeting Minutes for 4 Feb 2022

<b>Date:</b>	 <b>04 Feb 2022</b>
<b>Time:</b>	 <b>5:00PM</b>
<b>Location:</b>	Microsoft Teams
<b>Present:</b>	<p> @ Ow Ling Jia</p> <p> @ Sarah Chin</p> <p> @ ONG JHIN YEE _</p> <p> @ Chen Jian Yu</p> <p> @ Joshua Wong</p> <p> @ Tian Le Cheow</p>
<b>Absent:</b>	None

### Goals

Clarify with Prof Jisun the plans for AWS services used moving forward, discuss other cheaper alternatives and feedback on proposed ML methodology.

### Agenda

No.	Agenda Item	Remarks
1.	<b>Plans on using AWS services and other alternatives</b> a. Current AWS services used, and estimated cost b. Alternatives to AWS services used, cheaper options	Joshua, Jian Yu
2.	<b>Updates on Project</b> a. Current progress b. ML methodology <ul style="list-style-type: none"><li>• Keyword analysis</li><li>• Trending topics</li></ul>	Ling Jia, Jian Yu




Agenda	Action By
Plans on using AWS services and other alternatives	
<p><b>Current AWS services used, and estimated cost</b></p> <ul style="list-style-type: none"> <li>Informed Prof Jisun about the situation with AWS <ul style="list-style-type: none"> <li>Accidentally set to constant throughput on the 30th after increasing the capacity for the rows, which was increased initially because some of the rows were too big. Increasing the capacity increased the constant throughput, instead of using the dynamic pricing option <ul style="list-style-type: none"> <li>Constant throughout: charging a fix amount per hour (regardless of traffic)</li> <li>Dynamic pricing: charges vary depending on the traffic</li> </ul> </li> <li>Settings were changed last Sat and we did not realise that changing the constant throughput would have affected the price <ul style="list-style-type: none"> <li>Prof Jisun: Will try to get the server back, at the meantime, collect the data manually ourselves on our own local server</li> </ul> </li> </ul> </li> <li>Prof to vet through our drafted reply to AWS and mentioned some preventive measures we can take in the future <ul style="list-style-type: none"> <li>Regular checks</li> <li>Turn on emails notifications and updates</li> <li>AWS asking for more information as part of a standard protocol <ul style="list-style-type: none"> <li>Seems promising that we can appeal for a waiver of the costs, as such situations have occurred before, especially with student projects. Prof Jisun mentioned that if we have to cover the costs eventually, she would have to use funding from her other projects, so it would still be the most ideal if the costs are waived</li> </ul> </li> </ul> </li> <li>Checking of the prices of AWS services (~\$800USD-\$1000USD/month) <ul style="list-style-type: none"> <li>Estimated costs when we use all the services on AWS (calculated by dividing the total memory by the number of records we will collect)</li> <li>Dynamodb (~\$600USD/month) <ul style="list-style-type: none"> <li>Based on Facebook data scraped from Facepager, each record takes up ~0.5kb of memory</li> <li>For Twitter data, each record takes up ~3kb of memory</li> <li>Upper limit of 10kb per record, collecting around 3-4 years worth of historical data, estimated 20 million read and write requests for all platforms estimation of \$600 USD for the scraping of data and uploading to dynamodb</li> </ul> </li> <li>Lambda function (~\$200USD/month) <ul style="list-style-type: none"> <li>In order to scrape the amount of data we require, the lambda service might have to run for a long time, but it's hard to estimate the amount of time required to run the scraper</li> <li>Estimated 5-6 hours of scraping required per platform per day</li> <li>Might be an underestimation, uncertain as of now</li> </ul> </li> </ul> </li> </ul> <p><b>Other alternatives to using AWS services - to collect 1 year's worth of data first</b></p> <ul style="list-style-type: none"> <li>Prof Jisun: Using EC2 server, S3 storage together with MongoDB/SQL or shifting everything to local machines first, before migrating over to the server when it is ready</li> <li>Team's reasons for choosing current services <ul style="list-style-type: none"> <li>DynamoDb instead of S3 <ul style="list-style-type: none"> <li>DynamoDb is the NoSQL version of database in AWS</li> <li>S3 is not really a database but a file storage, no cloud database which means we cannot connect to any cloud database, unless eg we host MongoDB on firebase</li> </ul> </li> <li>Lambda service instead of EC2 server <ul style="list-style-type: none"> <li>Lambda is a serverless service which will be easier to set up, compared to using EC2 which is more complicated, sort of like a PaaS</li> <li>Once off bill for using Lambda service for scraping the historical data, will not require Lambda for further processes</li> </ul> </li> </ul> </li> <li>Prof Jisun: Take care of the current bill and stick to current AWS plan; keep track of the account and costs that incurred <ul style="list-style-type: none"> <li>Alternative would be to just use AWS Lambda for the scraper because running the scraper on our local machines would require too much computational power, and AWS would be more ideal <ul style="list-style-type: none"> <li>Scraping of real time data can be done on the SMU server after it is back up</li> </ul> </li> <li>Database would be shifted to local machines, using MongoDB instead, but the drawback would then be that we will not be able to sync the database across our local machines since the data is now not on the cloud</li> <li>Another alternative would be to explore EC2 and if that works, we can directly deploy the web application on EC2 since it works like a server and it would be a more stable solution</li> </ul> </li> <li>Web dev - use local server; data collection - EC2</li> </ul>	<p>@ Chen Jian Yu - Research on EC2 services and MongoDB</p>
Updates on Project	


<p><b>Current Progress</b></p> <ul style="list-style-type: none"> <li>Scrapers for historical data for the other platforms have been completed</li> <li>Frontend has start coding</li> <li>Team has discussed on ML methodology and are trying out</li> </ul>	-
<p><b>ML methodology for keyword analysis</b></p> <ul style="list-style-type: none"> <li>Current proposed methodology: <ul style="list-style-type: none"> <li>Use NER to identify the different entities that have been mentioned in online posts and comments</li> </ul> </li> <li>Prof's feedback: Using NER is a great way to start, but since we are only considering the word frequency, we might see the same data everyday. <ul style="list-style-type: none"> <li>Once we get 1 year's worth of historical data, aggregate the data per week to see if there are any changes in the keywords over time. If there is no change, then we might have to find another method to show the newly appearing keywords as compared to the previous week's.</li> <li>Apply NER weekly instead as applying it daily might not be very useful (keywords may be the same)</li> </ul> </li> </ul>	<p>@ Ow Ling Jia</p> <p>@ Chen Jian Yu</p> <p>@ Joshua Wong</p>
<p><b>ML methodology for trending topics</b></p> <ul style="list-style-type: none"> <li>Current proposed methodology <ul style="list-style-type: none"> <li>Will not be using topic modelling (unsupervised), since it's hard to generate applicable topics due to the vast amount of data to process, and it takes time to finetune the data</li> <li>Will manually come up with keywords associated with pre-defined topics eg politics (as a topic), then do a match against all the posts, and assign a topic to the post containing the keyword <ul style="list-style-type: none"> <li>Limitation: cannot deal with newly appearing topics</li> </ul> </li> </ul> </li> <li>Prof Jisun: Trending topics - difference between topics and keywords <ul style="list-style-type: none"> <li>Would politics be trending all of the time? As a user, she would like to see what topics would be trending more often at that point in time.</li> <li>Pre-defined topic - we intend on putting more keywords that are related to that topic <ul style="list-style-type: none"> <li>do some topic modelling instead of having the texts all over the place (generalization)</li> <li>Prof: how do we deal with new topics <ul style="list-style-type: none"> <li>if we were to have new topics, we would need to manually scrape and re-update the topics (limitation)</li> <li>if we were to create an unsupervised version now, it will take up too much time for the group to perfect; we still need someone to manually filter and come up with the topics - not sustainable in the long run</li> <li>currently, we will think of current topics that can be easily accessed by the user</li> </ul> </li> </ul> </li> <li>Will think about other methodology for trending topics, but for now, the best approach would be to test out on some sample data and see if we can get the insights and visualization as expected</li> </ul> </li></ul>	<p>@ Ow Ling Jia</p>


:note: Meeting Notes - Updates

Update	Risk Level	Mitigation Plan
Unexpected bill charged on AWS account	<b>HIGH</b>	<p>Contact AWS Support centre to explain the situation and attempt to get a waiver for the charges</p> <p>Moving on:</p> <ul style="list-style-type: none"> <li>Be more wary of the bills and check regularly on the billing dashboard to ensure the costs are within expectations</li> <li>Rethink about the services we will be using from AWS</li> </ul>

 Decisions

-  Reply AWS on billing and waiving of charges
-  Will not use DynamoDB if possible; Use Lambda service
-  Research on using EC2 and S3

 Administrative Matters

<b>Date of Next Meeting:</b>	<b>18 Feb 2022</b>
<b>Time of Next Meeting:</b>	 5PM