**DNSC 6290 Project - Working With Large Datasets**

James (jamesworrall@gwu.edu), Neel (nbarge96@gwu.edu), Suyash (suyash65@gwu.edu)

**Executive Summary:**

We created dataframes from the seven tab-delimited text files that were provided for the advertising dataset. We then formulated several investigative and analytical questions and queried the data to draw conclusions from the dataset. We investigated the top 10 search categories and locations based on clickthrough rates for paid contextual ads, categories with the highest average cost of advertisements and which region are they based out of, and the count of devices: UserDeviceID = 2019 used to search Non-Contextual Ads.

**Methods:**

We read the text files into separate dataframes using Spark’s *read.csv* command. We set the delimitation to \t for tab and set *inferSchema* and *header* to true. This resulted in seven dataframes with schemas accurately reflecting the data contained in the text files.

For the analysis of the top ten search categories and location by clickthrough, the ‘search click’ and ‘search info’ data frames were joined on the common SearchID field. Null values were removed from the *IsClick* and *HistCTR* due to only focus on the clickthrough rate of the paid contextual ads. They were then grouped by their *CategoryID* and *LocationID*. The average was found for *IsClick* and *HistCTR* using an aggregation. Then the top ten results were sorted in decreasing order. The results were visualized in a bar chart using the display command.

To know about the top 10 categories with the highest average cost of advertisements and which region they are based out of, we used Spark’s SQL API to create a *price\_analytical\_table* object and used inner join on *location\_df*, *category\_df,* and *ad\_info\_df* to find the average of the *Price* of the advertisements grouped by *CategoryID* and *RegionID* and selected the top 10 results of the *price\_analytical\_table* dataset.Spark’s SQL API on DataFrames makes this process super fast and easy to work with.

To find out the number of Non-Contextual Ad searches by the *UserDeviceID=2019*, data frames search click and combined stream were joined, using the common *AdIDs*. Further, the newly formed dataframe was joined on the UserID and filtered for 2019 user device type. The main idea was to understand if the different device types had different counts for landing on Non-Contextual Ad Pages.

**Challenges:**

In analyzing the clickthrough rates for the contextual searches two of the fields in the entity relation diagram were mislabeled *SearchLocationID* and Search *CategoryID*. This resulted in an error when trying to run a query on the data however this was easily solved by investigating the dataframes and extracting the correct field names of *LocationID* and *CategoryID*. Another challenge was in analyzing the location data for the ads clicked. The time period was short for the collection and therefore the sample size was not large enough for some locations. This resulted in locations having 100% of their ads clicked if for example there was only 1 ad and it was clicked on. The historical clickthrough rates provide a more accurate picture of the locations that had the most ads clicked. Since data columns like *CategoryID, RegionID,* etc don't have descriptive attributes, we won't know the key-value pair of those results.

**Results/Conclusions:**

For the analysis of the top ten search categories, it was found that categories 18, 2, 0, 36, 500001, 46, 25, 51, 13, and 29 had the highest average number of contextual ads clicked. This can be compared with the top categories based on historical click-through rates of 43, 12, 0, 29, 36, 17, 7, 28, 2, and 18. There is a commonality between 50% of the top ten historically high click-through categories and the average number of ads clicked in the 16-day sample categories. One category from the sample set, 46, was also on the list of the top 10 categories by average ad price. For the location analysis, the sample set in the search click sample dataframe was not robust enough to draw conclusions based on the number of ads clicked per location. Using historical clickthrough rates the top ten locations for contextual ads clickthrough were 1090, 2084, 263, 2961, 3462, 766, 4278, 1084, 2348, 2567.

*RegionID*s 83, 14, 69, 3, 72, 41, 41, 27, 67, 77 are the regions which have the highest average costliest advertisements with *CategoryID*s 250005, 31, 60, 19, 46, 250004, 250001, 37, 4, 250006 respectively.

**Future work:**

We would look into the performance of other types of ads such as non-paid and location based to compare with contextual as well as define what types of users are generating the most traffic.

**Takeaways from the course:**

The biggest takeaway from the course has been the effectiveness of distributed computing and how these tools can be employed to analyze large datasets not possible on a personal computer. Utilizing Azure and Spark we are able to tap into computing resources and a distributed architecture to analyze a large amount of data. This makes it possible to load data, perform ad-hoc analysis as needed and in a cost effective manner.