**Name: Snehal Yadav**

**Website:** [**https://github.com/snehal-y**](https://github.com/snehal-y) **Repository : AI2\_insights**

I used Python programming to do the analysis of the given dataset, named as ‘BIE\_homework\_assignemnt.csv’. Before doing analysis, I checked what all types of variables are in the dataset and checked if there are any missing values in the given dataset.

I found that session\_id is the only variable which has integer type of data and remaining all variables are of object type. Generally, columns with ‘object’ types are possible categorical variables.

Also, none of the variable except landing\_page, had a missing value. So, I decided to remove all those rows which had missing values. Landing\_page had 770 missing values. So, after removing 770 rows, I left with 9,99,230 rows of data.

**Top 100 papers in the dataset:**

My first step to find top 100 papers was to segregate different types of landing pages. In the given dataset, landing page had different types because a user looks for topics, alerts and papers through a landing page. I created df1 dataframe which enlist only those rows which contains ‘paper’ keyword in the landing\_page column. After that I used value\_counts() function of Pandas library, which returns objects containing counts of unique papers. The resulting list of counts is in descending order, meaning that the first element is the most frequently occurring element. I saved that in df\_papers dataframe and then in the third step I extracted the list of all papers in ‘Top100Papers.csv’ file.



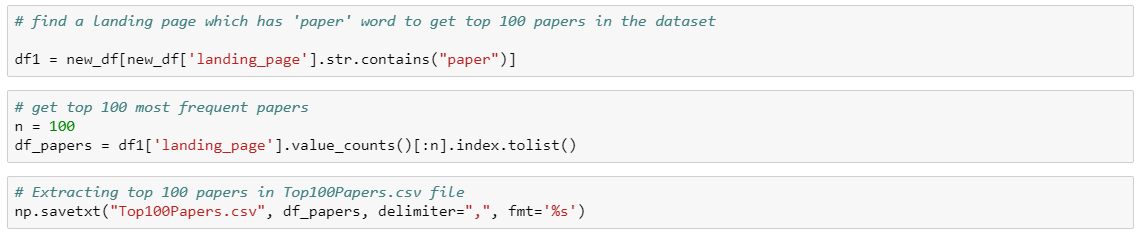
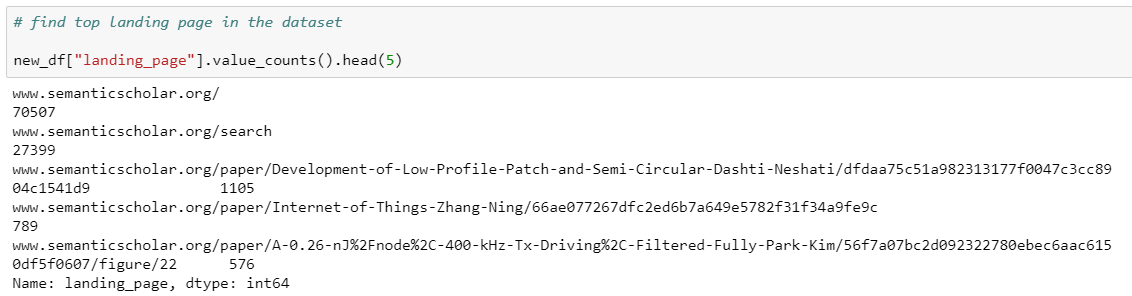


Figure 1: Snippet of code to find Top 100 Papers

**Top landing page type in the dataset:**

Assumption: I assume that the type of landing page is the link on the main webpage. In this case, ‘www.semanticscholar.org/’ is top most landing page visited by the users. I believe that this is the home page and on this hope page there are different links. So, I considered those different links as different types of the landing page. I followed below steps to know top landing page type.

Step 1: Determined top landing page in the dataset



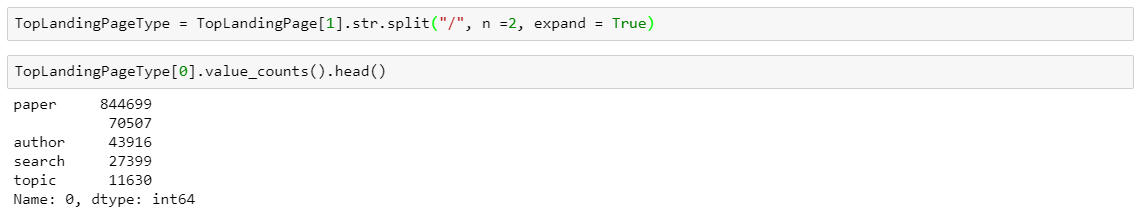
Step 2: Split the URL in different strings and saved the output in ‘TopLandingPage’ data frame



Step 3: Split URL more to get the type of the top most landing page. Stored the output in ‘TopLandingPageType’.



Step 4: As we see the type of landing page is stored in index 0, I took a count of distinct values of type of landing pages and found that ‘paper’ is the most visited landing page type.



**Interesting insights:**

It looks like users vistited the most on June 05, 2018 as compared to all days. The frequency is 214047.

