20220626

Jeremy Beard

MSDS610 – Final Project

In this document, I am providing the Stack Overflow query information gained from data.stackexchange.com. I decided to keep the data as simple as possible and only get the information which was needed. This information was the post ID and the tags of each post. I also decided to only get data from the year 2020 using wildcards and the ‘where’ command. Using all this, I queried information which I then saved in csv format.

Graphical user interface, text, application

Description automatically generated

Figure 1: Query information from data.stackexchange.com

Graphical user interface, text, application, table

Description automatically generated

Figure 2: Example of the output information gained from the Stack Overflow query

For all other information pertaining to this final project of the MSDS 610 course, please refer to the Jupyter Notebook file (and corresponding pdf file) which goes over the code and the methodology behind the code. Please let me know if you have any questions! Thank you!

All the best,

Jeremy

References:  
1) How to create a pyspark dataframe from multiple lists. (2018, October 12). Stack Overflow. Retrieved June 26, 2022, from <https://stackoverflow.com/questions/52784192/how-to-create-a-pyspark-dataframe-from-multiple-lists>

2) How can I get inverted index? (2018, March 1). Stack Overflow. Retrieved June 26, 2022, from <https://stackoverflow.com/questions/49059096/how-can-i-get-inverted-index>

3) How to export a table dataframe in PySpark to csv? (2015, July 13). Stack Overflow. Retrieved June 26, 2022, from <https://stackoverflow.com/questions/31385363/how-to-export-a-table-dataframe-in-pyspark-to-csv>