Project 2 Proposal

Title: Visualization of Status of Scholarship

Author: Mengping Yu

Supervisor: Henna Shangufta

Degree: MSc Big Data Analytics

Problem Description

This project attempts to use the tool of big data analysis named hadoop to analyse the information of scholarship, and this project intends to use HUE to execute the job of visualization. The problem of this project is to how to make the diagrams of visualization in the most concise and intuitive way according to the different needs.

This project consider to use the way of data visualization to help Higher Education Services Corporation HESC institutions and students and parents who are interested in scholarships to get the detail information more intuitive and more specific. The institution(HESC) can make a more reasonable plan for the allocation of scholarships after analysis, and for people who want to obtain scholarships, they can have a clearer direction when applying for colleges.

Description of Data

This Project select the status of scholarship of the HESC institute in New York City from 2009 to 2017 as the dataset to do analysis. The dataset came from the official data of New York, and the link of these data is https://data.ny.gov/Education/Scholarship-Recipients-And-Dollars-By-College-Code/ww7t-w8k9. This project uses five sets of data from this dataset, academic year, TAP sector group, TAP collage name, scholarship headcount and scholarship dollars, to complete the data analysis.

that the distribution of scholarships in different schools has received little attention for a long time. In this unpopular situation, because there are large number of schools located in the New York City and the sector groups of school are also complex, which cause the demand of scholarships required by institutions for different schools is varying each year. Therefore, analysis this dataset is important though there are few people care about it.

Approaches

This project uses hadoop and hue as the analysis tools. The definition of wikipidia of Hadoop is an open source software for big data analysis. This software uses the MapReduce model for efficient data processing and analysis(Judge and Peter,2012). Wikipidia also point out thta Hue is an open source, and this tool can visualize data using SQL language (wikipidia).

The data of maximum, minimum and total amount can be shown using MapReduce and HUE. For the aim of analysis the status of scholarship recipients in 2017, this project selects K-means algorithm to do this work. This is because as recent research has pointed that the K-means algorithm is a way of clustering data sets. This algorithm is very suitable for finding the similarity of data in a dataset, and this algorithm is also applicable to MapReduce used in this project (Sreedhar, et al.,

2017). This project attempts to use this method to get several clusters that includes a bunch of data on each cluster, and one of these cluster that includes the maximum amount of data will point out a range of the amount of scholarship recipients, and this is the answer of this question.

Goals

This project intends to use this dataset to obtain the college name with the maximum and the minimum number of scholarship recipients each year,the total amount of scholarships per year and the status of scholarship recipients in 2017. The result can be displayed using visualization tools.

Conclusion

In a nutshell, this project chooses the status of scholarship in New York from 2009 to 2017 as the dataset to do the processing of big data analysis, and because the data visualization is the most suitable one for displaying the result, so this project chooses HUE as the visualization tool to complete the project.

link of Github

References

[1] Sreedhar, C., Kasiviswanath, N. and Chenna Reddy, P (2017). Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. J Big Data 4, 27 doi:10.1186/s40537-017-0087-2

[2] Judge, Peter (October 2012). "Doug Cutting: Big Data Is No Bubble". silicon.co.uk.

[3] hue release, https://en.wikipedia.org/wiki/Hue$_{(software)}$