CAPSTONE 1 PROJECT – PULITZER

Milestone Report
Dec 28th 2017

Objective and Clients

The goal is to identify different types of Newspapers Segments based on Pulitzer prize and then identify ways to increase daily circulations and boost readers confidence in print media.

- ❖ Newspaper Segmentation
- ❖ Newspaper with the maximum number of Pulitzer prices
- ❖ What are the top 5 states?
- ❖ Correlation between Crime, GDP, and Population on Pulitzer? For example - higher GDP means more prices (confounding parameter could be more journalists) or crime-prone cities incubate investigative journalism resulting in more Pulitzer

Clients: New York Times (www.nytimes.com) and Machine Learning Community.

Data

The data for this project is collected from different internet resources as listed below. The base also called raw data ("Pulitzer Dataset") is made available by FiveThirtyEight. It is a simple table mapping newspaper with daily circulation and Pulitzer. Pulitzer Dataset is combined with many other dataset in order to find socioeconomic correlations. Additional data source and their formats are described below.

Additional Data Sources used:

1) Raw Data is made available by FiveThirtyEight.
   https://github.com/fivethirtyeight/data/blob/master/pulitzer/pulitzer-circulation-data.csv

2) Crime Data by State and US: The idea is to check if there is any correlation between crime rate and Pulitzer. The dataset is made available by US Departments.
   a. http://www.usa.com/rank/us--crime-index--state-rank.htm: The data is a simple excel sheet mapping crime index to US state by population.
   b. https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-1 : This is also an excel sheet mapping year, US population to crime index and types of crime

Data Contd…

Additional Data Sources used:

3) GDP for US and States: https://www.usgovernmentrevenue.com/download_multi_year_2000_2014USb_19c1li101mcn_F1cF0t: This is a CSV file mapping GDP by state by year dataset. We have 50+1+1 such files.

4) Population by State and US: https://www.census.gov/data/datasets/2016/demo/popest/state-total.htm : This is also an excel based data file mapping US stats to population to year. We have two files to cover 22 decades i.e. 1 file per census.

Data Wrangling

The final Pulitzer dataset went through industry standard data validation and verification(V&V) along with transformation before it was used for exploration and pre-processing. The final dataset was created after merging and transforming 58 datasets. For Audit, timestamp and unique identification(UUID) was added to all the 58 pandas dataframes. And to make it persistent and reproducible, domain level information was stored in Cassandra cluster running in Google Cloud. Please see the below list for different type of operations as performed.

- ❖ Pulitzer Dataset (1 data table) : Datatype Conversion, Removing special char {, % / \} , pandas.join|merge|concat|pivot
- ❖ Crime Data (2 Data tables): Missing Values, Datatype Conversion, Column Hacks
- ❖ GDP for US and States (53 Data tables) : CSV Bad Lines, Missing Data, Column Hacks, pandas.join|merge
- ❖ Population by State and US (2 Data tables) : Selective columns reading, Datatype Conversion

Highlights:

1) Google Cloud Platform and Multi Threading to handle massive volume of data
2) Cassandra Storage for Persistency and Reproducibility
3) Audit details for Traceability

Data Exploration

Based on initial data review, it appears that we have outliers and Pulitzer data tends to follow a 80-20 rule i.e. 80% of the newspaper population is under 10 Pulitzer Prizes and 20% of the population is above 10 Pulitzer Prizes.