Capstone 1 – Project Proposal – Pulitzer

Newspapers' Segmentation: Clustering

1. **Objective**: 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.

   As part of this project I would like to find:

   a. Newspaper segmentation

   b. Newspapers with the maximum number of Pulitzer prices?

   c. What are the top 5 states?

   d. To find if there is any 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.

   *Project Extension Ideas: This model can further be extended to cluster newspaper article **SOURCE** for pattern recognition followed by classification (chaining model outputs) to boost news quality. (This is not done as part of this project)*

2. The Client

The client for this project is New York Times (www.nytimes.com) and the Machine Learning Group. The purpose is to find an ML model which can be used to correctly cluster newspapers. It will not only increase circulation (e-commence) but also boost readers confidence in news.

3. Data source and Credits

a) Raw Data Source and Credits: Pulitzer Data is available FiveThirtyEight GIT HUB site
https://github.com/fivethirtyeight/data/blob/master/pulitzer/pulitzer-circulation-data.csv

b) Crime Data by State and US:
http://www.usa.com/rank/us--crime-index--state-rank.htm
https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-1

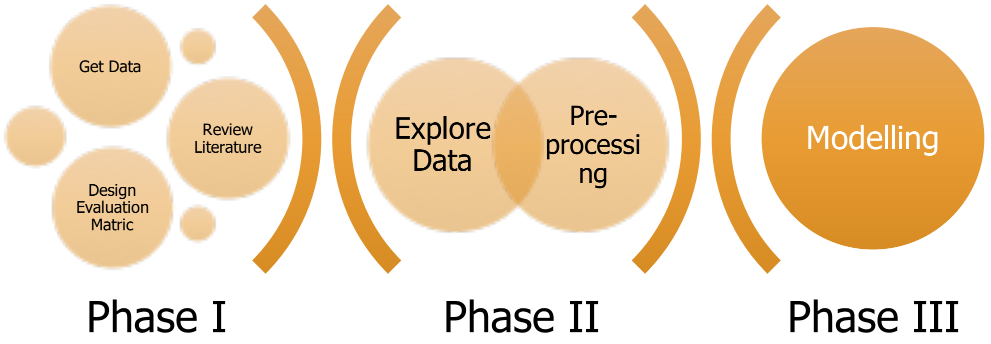
c) GDP for US and States:
https://www.usgovernmentrevenue.com/download_multi_year_2000_2014USb_19c1li101mcn_F1cF0t

d) Population by State and US: https://www.census.gov/data/datasets/2016/demo/popest/state-total.html

4. Solution Approach

I plan to use K-Means and PCA to help with newspapers' segmentation.

The solution approach is specifically designed to achieve a controlled and structured approach to minimize data quality issues that may be present or introduced to protect personally identifiable information(PII). The solution is sub-divided into three phases as listed below.



a) **Data Assembly - Phase I**: This phase of the project is designed to gather and do basic cleanup like join, merge, add or update attributes.

b) **Explore and Preprocessing – Phase II**: This phase of the project is designed to validate and explore the dataset for all the problems listed in the "Problem" section of this proposal.

c) **Modelling and Evaluation Phase III**: In this phase of the project I will be exploring various machine learning algorithms to find the best model to cluster the newspapers.

5. Project Deliverables

a) Project deliverables are listed below.

1. An analysis report (a .pdf document) on:
    a. Newspapers' Segmentation
    b. Identify top 5 categories under which of a fraudulent transaction is happening.
    c. Detailing the most common modus operand of a fraudulent transaction
    d. The min and max time elapsed before a fraudulent transaction is detected.
    e. ML model to detect a credit card fraud
2. All project artifacts like – IPython Notebook with code, description and charts.
3. Project Presentation (.pptx)