CS 506 – NAACP Deliverable 0

Project Description:

The main goal for this project is to evaluate the media coverage of Black Americans in Boston over the past five years. This entails focusing on overall coverage, homicide coverage and coverage of predominantly black neighborhoods and sub-neighborhoods, while also expanding on the previous analysis by tackling new topics. The overall analysis will provide media outlets an objective analytical model to self-assess bias in their coverage.

Data sets:

First Dataset: The predominantly black neighborhoods and sub neighborhoods

Second Data set: WGBH and WBUR online articles that we need to scrape

Third Data set: List of homicide victims and race/gender/age

What has been done before:

The Boston globe data set was used in the previous semester. Topic modeling was to focus on:

- How much the black American sub neighborhood were covered
- Education and crime in Black American communities
- Coverage of the death of Black Americans
- When race is mentioned. Why is it mentioned?

What needs to be done:

The two mains things that have to be done this semester:

- Applying the previously developed model after web scraping the WGBH and WBUR online articles. In addition to improving the topic modeling while forming visualizations that better explain the clustering's that are being formed.
- Performing entity recognition on people mentioned in the articles (mostly focusing on the Boston Globe data set) and matching the individuals with their real-life profile using the Wikipedia API and LinkedIn API.

Step by step approach:

1- Review the topics and analysis that was done in Summer 2020

2- Web scraping online articles from WGBH and WBUR: To do this we will use tools like selenium, scrapy and Beautiful soup.

3- Repeat analysis that was done last semester on the new data set, while focusing on performing topic modeling on articles about sub-neighborhoods and large neighborhoods.

4- Apply a new method: The new method we need us to work with ML practicum class to train a BERT model to understand semantic similarity and understand when black people are referred to without a direct Black reference.

5- Use the Boston Globe data set to extract the articles mentioning individuals and create a list of the names and the respective article. Then perform entity recognition by matching the individual using Wikipedia and LinkedIn APIs. Carifi API can be used in this process.

6- Complete visualizations for a more impactful presentation and to better explain the findings to non-computer scientists.