|  |  |
| --- | --- |
| **NAACP Media Research (DS) - V2 - Spring 2020** | |
| **Contact** | Tanisha Sullivan  President  NAACP Boston Chapter  617-433-7409  [NAACPBostonPresident@gmail.com](mailto:NAACPBostonPresident@gmail.com)  Paul Singer  [Paul\_Singer@wgbh.org](mailto:Paul_Singer@wgbh.org)  202-746-6047 |
| **Organization** | NAACP Boston |
| **Organization Description** | The mission of the National Association for the Advancement of Colored People is to ensure the political, educational, social, and economic equality and rights of all persons and to eliminate racial hatred and racial discrimination. |
| **Project Description** | The NAACP seeks to understand the coverage of Boston Media in covering Black people and Black communities in Boston, specifically to assess if and how coverage has changed over time in different news outlets pertaining to specific neighborhoods and sub-neighborhoods. |
| **Other Strategic Questions to be Answered** | 1. How much coverage, how many stories are there about Black people and Black communities? 2. How does coverage of Black people and Black communities vary by news outlet and over time? 3. What are the most common topics of stories about Black people and Black neighborhoods and sub-neighborhoods? 4. Who are the Black individuals most commonly profiled in the media in Boston? 5. Is there a difference in the coverage of Black victims of homicides vs. non-Black victims? 6. What percentage of stories about Black neighborhoods are negative vs. positive compared to articles about non-Black neighborhoods? 7. Is there a difference in the number of obituaries of Black individuals vs. non-White individuals? |
| **Approach** | **Step One: Define Neighborhoods and Sub-Neighborhoods**  Compile a list of the names of predominantly Black neighborhoods and sub-neighborhoods based on a standardized, transparent and defensible method. We suggest the following approach:   * Boston has [23 official neighborhoods](https://www.boston.gov/neighborhoods) but these are large and have diverse sub-neighborhoods * Download the list of sub-neighborhoods from <https://geonames.usgs.gov/apex/f?p=138:1:0::NO::P1_COUNTY,P1_COUNTY_ALONG:n>[,](https://geonames.usgs.gov/apex/f?p=138:1:0::NO::P1_COUNTY,P1_COUNTY_ALONG:n,)   + US Board on Geographic names: Can be used to find sub-neighborhoods instead of Google * Collect the demographics using the [Mass Census data](https://docs.digital.mass.gov/dataset/massgis-data-datalayers-2010-us-census) or Census’ [Download Center](https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml) at the Census block level (if the census tract/zip code is too large) to determine the sub-neighborhoods with the largest concentration of Black people. * Map all neighborhoods and sub-neighborhoods on ArcGIS (you will have a workshop on ArcGIS in March)   **Step Two: Collect Articles from Major Boston News Outlets**   * Review the compilation of articles from last semester’s teams to determine if additional collection is required. If so, you will use the [Internet Archive’s Wayback Machine](http://wayback.archive.org) to view the past five years of coverage on Boston news sites. The current available data included the following time periods:   + - Team one did Jan 1st, 2014 - 2019 (Boston Herald)     - Team two did Jan. 1st ,2014 - Dec. 31st, 2018 (Boston Globe)     - We suggest focusing on Focus on January 1, 2014 - December 31, 2018 * The [scrapy-wayback-machine](https://github.com/sangaline/scrapy-wayback-machine/blob/master/scrapy_wayback_machine/__init__.py) Python library is specifically designed to retrieve timestamped archival copies of webpages. It is easy to set up and use, and can definitely be adapted to work with the Globe, the Herald, and WGBH (and any other site posting stories about Boston). * Next, compile a list of articles that reference Black people\* and Black neighborhoods (the neighborhoods with the highest concentrations of Black people per **Step One**).   News source scrapes should focus on the following, in order of priority:   * **High priority: print media**   + [Boston Globe](https://www.bostonglobe.com)   + [Boston Herald](https://www.bostonherald.com/)   + [Boston.com](https://www.boston.com/rss-feeds) * **Second priority, public radio**   + [WGBH - Boston Coverage](https://www.wgbh.org/news/boston-public-radio)   + [WBUR - Boston Coverage](https://www.wbur.org/topic/boston) * **Lower priority, local TV**   + [ABC - WCVB - Local News](https://www.wcvb.com/local-news)   + [NBC - Channel 10 - Local news](https://www.nbcboston.com/news/local/)   + [FOX - Channel 25 - Local News](https://www.boston25news.com/news/local)   + [CBS - Channel 4 - Local News](https://boston.cbslocal.com/)   + [NECN](https://www.necn.com/results/?keywords=boston)   **Step Three: Topic Analysis**  **Topics Part A: General Content**   * Identify primary topics covered in the content of the articles (using word count frequency: TF:IDF vectorizer and Word2Vec to find word associations combined with clustering for visualization. * Clean data by removing stop words and filler words * Analyze what words show up more frequently than other neighborhoods; try to understand why, i.e. if there is an outlier, look up the relevant articles for context * Pay special attention to people’s names, create a sub-list of most-mentioned names to see if there is a pattern by race. (M= Male; F= Female; B=Black; A=Asian; W=White; N=Non-Hispanic; H=Hispanic)   **Topics Part B: Coverage of homicides**   * Collect articles on all homicides by name/data for 2014-2019. See data [HERE](https://drive.google.com/open?id=1v_SJrHwHzoRQHE34DjyLj_4-DbNAewg4). * Analyze difference in the volume of coverage based on race (Black vs. Non-Black vs. White vs. Asian vs. Hispanic) * Analyze sentiment of the coverage toward the deceased based on race   **Topics Part C: Explicit coverage of Black People**   * Newspapers are not supposed to mention race/ethnicity unless it is relevant to the story. Given this, we want to analyze the content and sentiment of articles where the race of individuals is explicitly mentioned. Identify what are the story topics? * Apply same analysis used in **Topics Part A** above (e.g. word associations, clusters, etc. * Pay special attention to people’s names, create a sub-list of most-mentioned names to see if there is a pattern by race.   **Step Four: Sentiment Analysis**   * We recommend using NLTK’s Brown Corpus that draws from news and press articles gathered in the 1960s. * Last semester we used the VADER Sentiment analysis tool based off a lexicon of Encyclopedias, Social media, and news articles. VADER draws heavily from social media and we want to avoid that data since social media tends to have very inflammatory language. However, we encourage using both VADER an NLTK and comparing the results. * Explore other Sentiment Analysis Models to see if you get different results. Others to consider include: LabMT sentiment analysis (wrapper [Dharm](https://github.com/weirdindiankid/dharmSentiment) Sentiment) * *Note: please explain how algorithm methodologies vary and what bias they may contain* |
| **Analysis Techniques** | * [Correlation network](https://drsimonj.svbtle.com/how-to-create-correlation-network-plots-with-corrr-and-ggraph) of racial/ethnic language with other terms * Modeling: how predictive is a neighborhood’s black population of certain language in coverage? * This time we can use the Word2Vec library to vectorize words and discover with what words they form associations. * Additionally we can also use clustering to visualize which words are most similar extended from the Word2Vec model |
| **Tools and approaches** | Apply simple sentiment analysis problem.   * To get a broad comparison of local coverage of majority-black neighborhoods versus others, and to do entity recognition, this [NLP engine](https://cloud.google.com/natural-language/docs/basics) (Natural Language Processing Engine from google) could be used, although it is a blunt instrument * NLTK Brown Corpus for sentiment analysis, a new source that was recently made [Wiki Neutrality Corpus (WNC](https://github.com/rpryzant/neutralizing-bias)). WNC draws from 180,000 biased and neutralized sentence pairs drawn from Wikipedia edits. Published by the Stanford NLP group, it seems promising as it may provide a way to single out loaded words. * Word2Vec to find word associations combined with clustering for visualization * NLTK Part of Sentence tagging to find combinations including nouns and names (by gender, race) * [IBM NLU API](https://cloud.ibm.com/apidocs/natural-language-understanding/natural-language-understanding) - create an instance - keywords, entity identification, sentiment, emotion on entities and overall articles |
| **Relevant Datasets** | Link to [final reports and posters from Fall 2019 teams](https://drive.google.com/open?id=1qlM4EgZq9uAweckRmPNSAa_0Gz9mBuUk) working on the NAACP Project:   * Team Yufeng Data (Boston Globe): <https://drive.google.com/drive/folders/1B82_YZmT1HhK2BNwLuGcZZ-ZhcN06ESn> * Team Yufeng Project Github repo: <https://github.com/AllenChenGH/NAACP_MEDIA_RESEARCH> * Team Libby Github repo (Boston Herald): <https://github.com/etjames/NAACP-Media-Research> |
| **Relevant articles and studies** | <https://arxiv.org/pdf/1801.05802.pdf> |

**Notes from 2/14**

Goals for 2/21:

Step one:

* Identify neighborhoods and sub-neighborhoods that are black (based on census)
* Use exact block for Longitude/ Latitude and collect census information for that block to determine demographics (census tract)
* Calibrate with Singer (i.e. determine which neighborhoods are considered as “Black” vs. “non-Black”)

Step two: additional data collection

* Review Herald to make sure the data is high quality/ complete and that the dates are comprehensive for January 1, 2014-December 31, 2018
* Collect additional articles as desired (e.g. from WBUR, WGBH, etc.)

Step three:

* Analyze existing media corpus of media articles to filter articles about “black people” - see keywords
  + - Black
    - Afro Latino
    - Afro Latina
    - Cape Verdean
    - Haitian
    - Haitian-American
    - African
    - African-American
    - Caribbean
    - Jamaican
    - Haitian
    - Dominican
    - West Indian
    - + people, male, female, man, men, woman, women, child, kid, youth, community, neighborhood, business, company (alongside any of the keywords above)

**Notes from 2/21**

* Teams to scrape the race and demographic breakdown by census tract in Suffolk County, MA. First half of census tracts assigned to Shubhangi’s team, second half assigned to
* Next Week: cross reference with [geonames data](https://geonames.usgs.gov/apex/f?p=138:2:0::NO:RP::) (sort for MA, Suffolk County, filter map column for Boston)