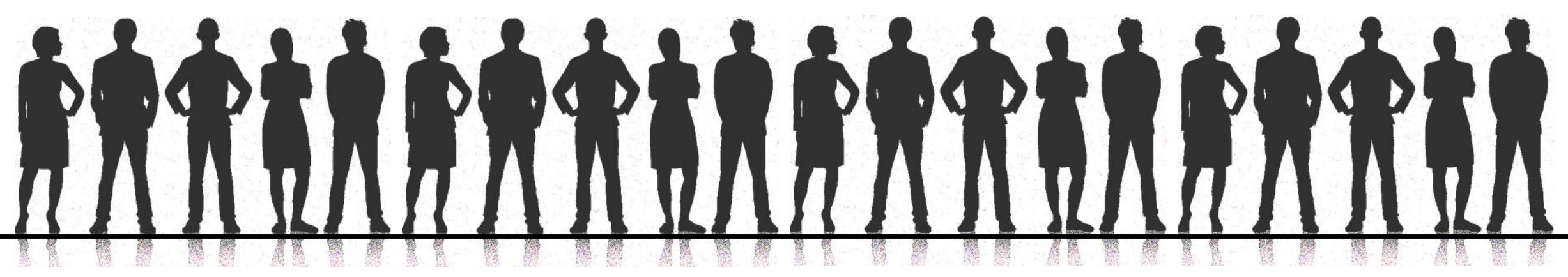
BU Department of Computer Science



NAACP Media Research

Yufeng Chen, Jiaqi Sun, Ruotian Liu | CS 506 - Spark! Project

Problem Statement

The National Association for the Advancement of Colored People (NAACP) is a civil rights organization in the United States, formed in 1909 as a bi-racial endeavor to advance justice for African Americans. Their goal is to ensure the political, educational, social, and economic equality of rights of all persons and to eliminate race-based discrimination.

NATIONAL ASSOCIATION FOR THE ADVANCEMENT OF COLORED PEOPLE · NAACP FOUNDED 1909

The Boston Globe

89.7 WGBH Boston's Local NPR®

90.9 wbur BOSTON'S NPR® NEWS STATION

In this project, we are going to help our client, NAACP Boston, to understand the coverage of Boston Media in covering Boston's Black people and Black neighborhoods. We are going to assess the volume of coverage, the general sentiment of all reports, the topics covered, and how this has changed over time. To do so, we need to collect data from newspaper and radio websites (Boston Globe, as well as WGBH and WBUR), design algorithms to understand the data, then draw some conclusions. We think the results we get can provide a good view on how the media in Boston did in the past 5 years (from 2014 to 2018). Hopefully, these results will lead our client to some possible ways that can help the media to do better on the elimination of racial hatred and race-based discrimination.

Methodology

Collecting Data: To collect articles from website, we first tried a python module named "BeautifulSoup". It could work correctly for us to get the articles after adding some restriction to it, but it's just too slow and not acceptable because we need to get the news in recent 5 years from 3 different websites. Then we tried Scrapy. After got the structure of the websites we want to scrape and find out which class the article part lies in, we successfully got the data we want.

Filtering: After collecting all data we need to see which articles are about black people or black neighborhoods, so wrote an algorithm that can traverse all text data we've got and judge whether it was related to black neighborhood or not using a keyword list we've collected before (including black neighborhood names and some other words about black people).

Sentiment Analysis: We tried to train a model just like what we did in the midterm to judge an article's sentiment but couldn't find a very good train dataset, so the results are not very satisfying. Then we planned to use some existing sentiment analysis tools to do the job. Finally we decided to use 'VADER-Sentiment-Analysis' to help us determine if a piece of news is positive or negative after testing on some sample articles.

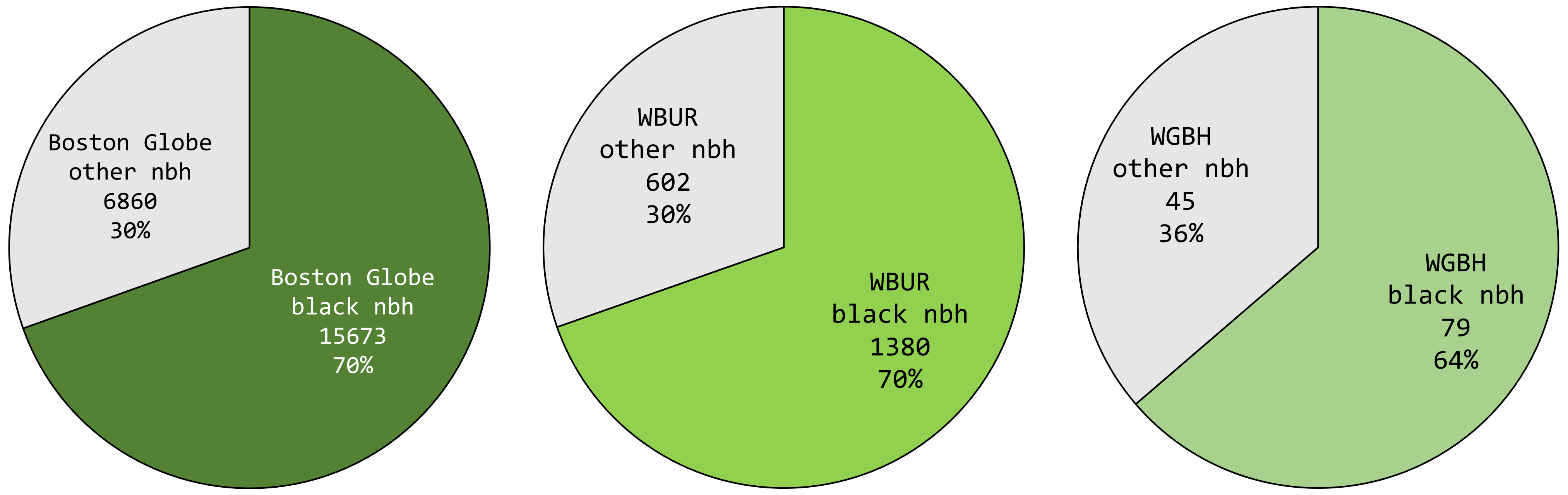
Getting Topics: The next step is to find out what topics are popular. In a word, our analysis was based on word count. We vectorized the article and tried to get rid of the influence of meaningless words for drawing conclusion like "said", "Boston" or "Monday" from two aspects. The first one was to remove all the words in our keywords used for filtering and in stop-word list since those were a common part of relevant articles. The second one was tf-idf algorithm. As we learnt in class, tf-idf was born to deal with this problem. The following step was a classification problem. We tried different models we used before and choose LDA (Latent Dirichlet Allocation) to do the classification job because it worked pretty well with text classification and it could provide a nested dimensionality reduction API so that we can make our result visible easily, which was quite clear to evaluate and show our results.

Data & Results

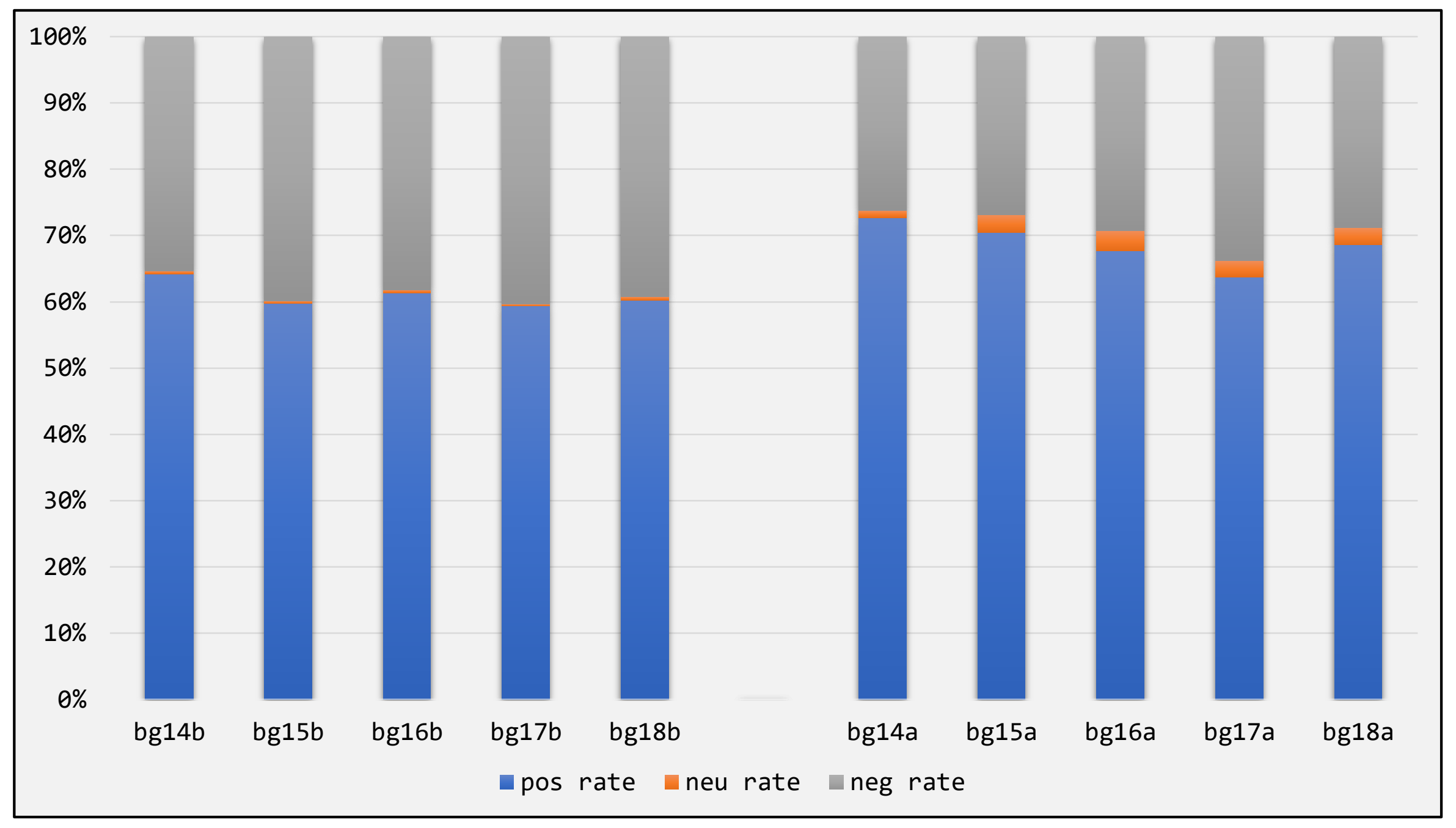
```
Total data Size: ~1GB
  Boston Globe: ~200000 pieces of news, ~24000 about black
     WBUR: ~30000 pieces of news and articles, ~2800 about black
       WGBH: ~1500 pieces of news and articles, ~150 about black
```

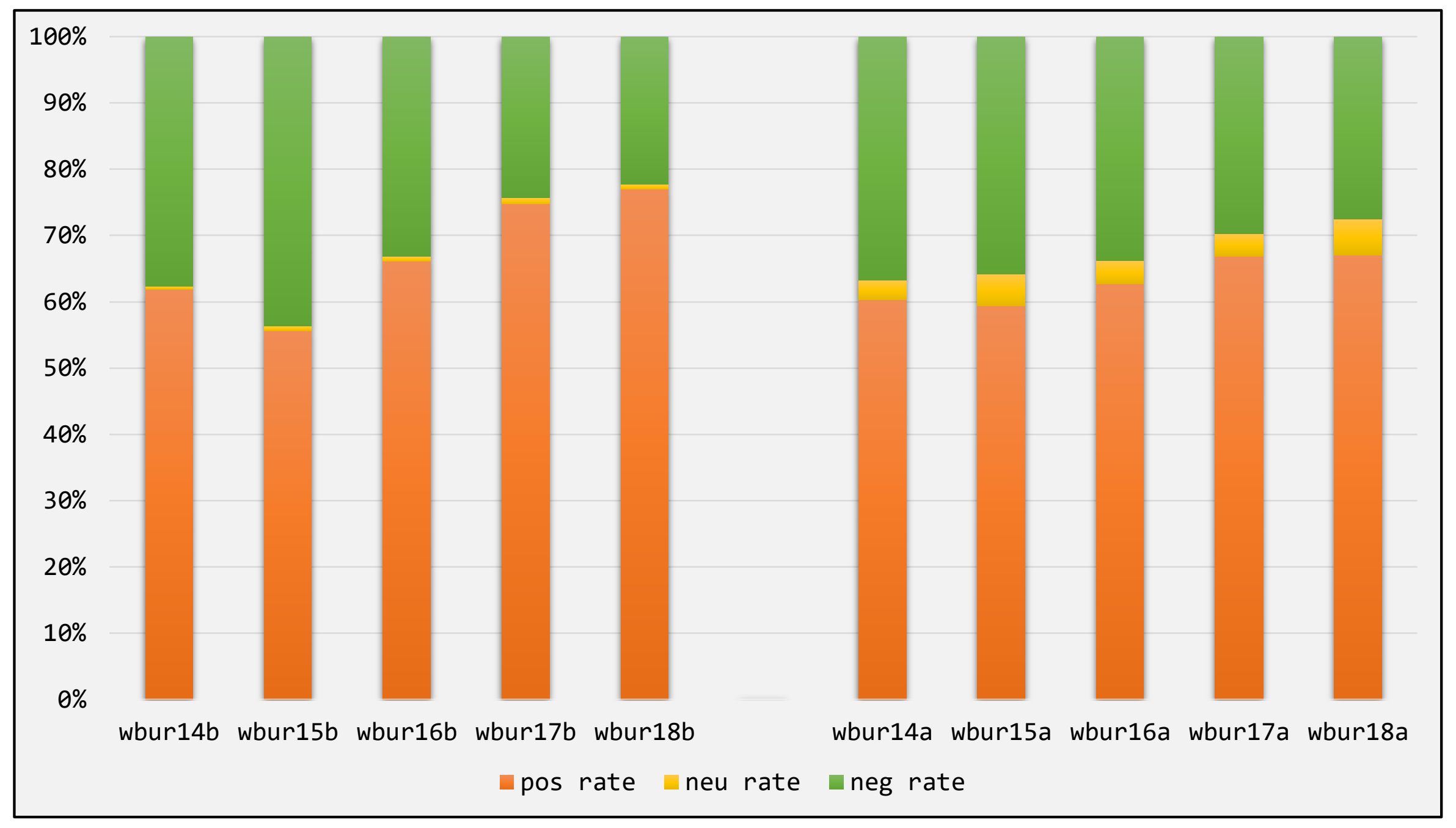
What We Have Achieved

Coverage



Sentiment

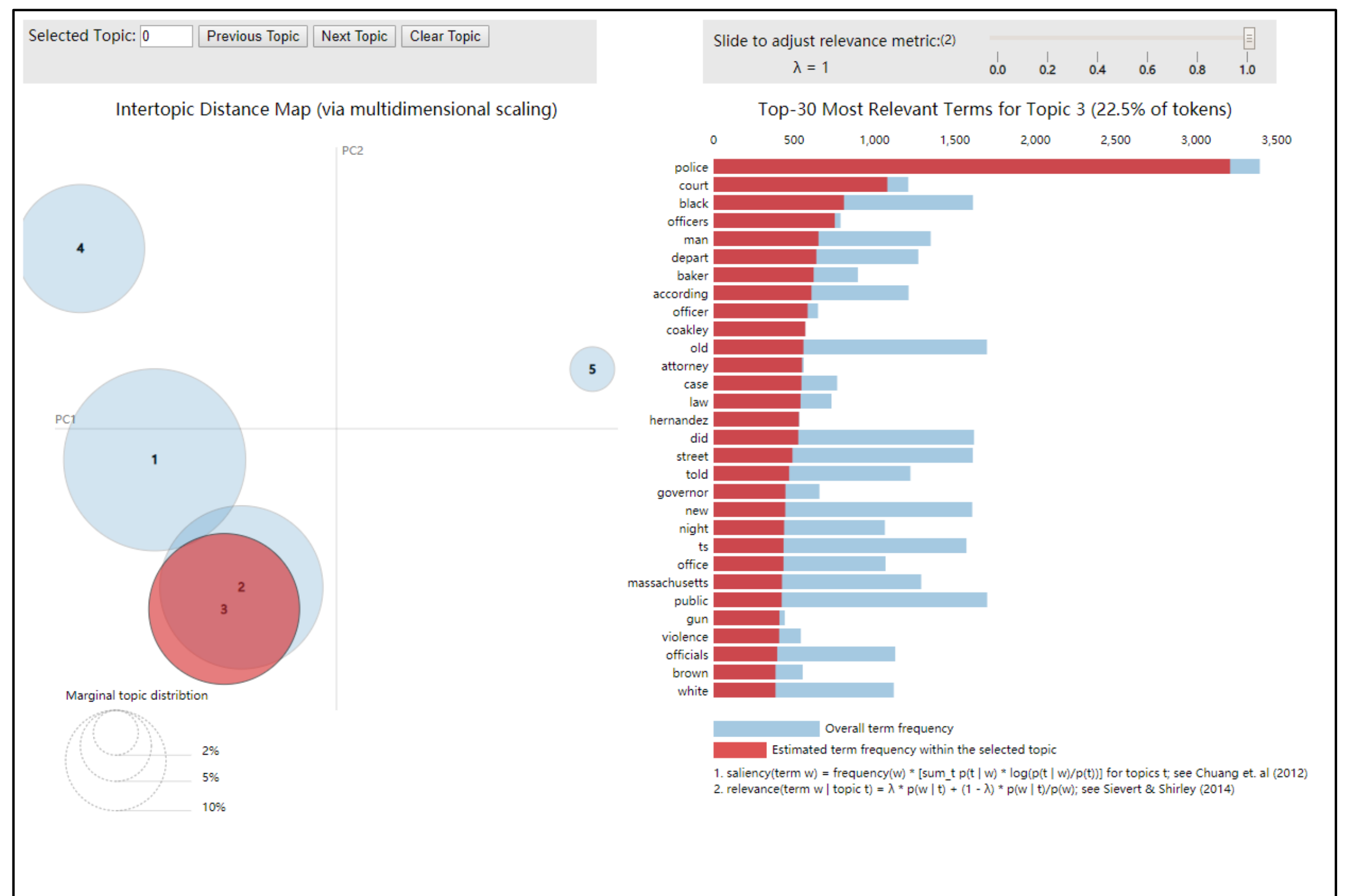




Topics

Boston Globe's top topics in news about black

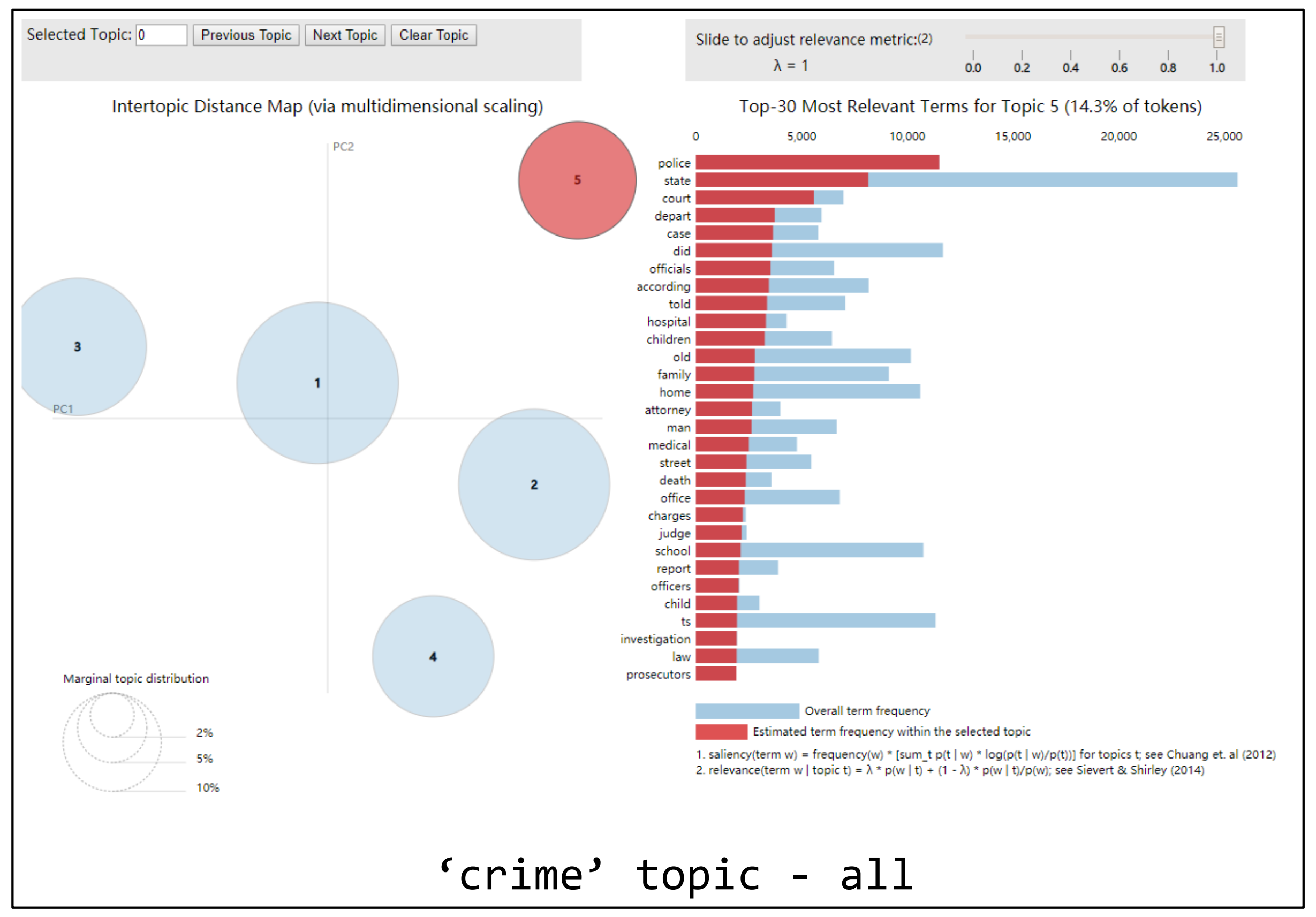
2014 top 5 topics	2015 top 5 topics	2016 top 5 topics	2017 top 5 topics	2018 top 5 topics
police court black officers man	snow 2024 olympic line south	school students black schools community	xa0 street team sox day	black white trump president american
percent 000 public health students	school students percent schools public	police officers man old street	trump president white jackson black	police street officers man court
council wall keolis contract councilor	police officers street man court	massachusetts 000 public housing million	life children family father mother	school students schools public percent
world black new american african	black life white hernandez day	day world new way old	school students percent million schools	massachusetts health depart public office
school menino mayor day family	work new building food home	trump clinton campaign president voters	police depart court officers according	life day world way xa0


Selected Topic: 0 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

'crime' topic - black


Selected Topic: 0 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

'crime' topic - all