# Algorithm Documentation: Audience Overlap Estimation

## This documentation outlines an algorithm designed to estimate audience overlap between pairs of websites, particularly focusing on Russian news outlets. Utilizing detailed visitor information, the algorithm refines the lower bounds of audience size by incorporating observed demographic characteristics. The methodology involves the creation of joint distributions and the application of statistical measures to enhance the precision of audience overlap estimates. The documentation covers the algorithm's description, methodology, results, analysis, challenges, and limitations, providing a comprehensive understanding of its application in refining audience size bounds. This algorithm serves as a valuable tool for optimizing audience estimation strategies in the context of online media analytics.

**1. Introduction**

The task is to estimate the audience overlap between pairs of Russian news outlets. We leverage detailed visitors' information to tighten the bounds of total audience estimates. The algorithm focuses on different demographic characteristics to refine the estimates.

## **2. Algorithm Overview**

The algorithm comprises three main steps: data preprocessing/loading, lower bound calculation, and audience overlap estimation. It leverages demographic characteristics such as browsers, resolutions, operating systems, etc. to improve the precision of audience overlap estimates.

## **3. Data Preprocessing**

Data preprocessing involves loading and merging two datasets: 'demo\_data' and 'visitors\_data.' The resulting 'merged\_data' contains information about the tag, date, demo, Visitors, and lower\_bound.

## **4. Lower Bound Calculation**

The lower bound is calculated based on the 'demo' and 'Visitors' columns. The formula used to calculate the lower bound is max(row['demo'], row['Visitors']).

## **5. Algorithm for Audience Overlap Estimation**

The algorithm estimates audience overlap across different demographic characteristics. It creates pairs of observed distributions for each characteristic, calculates the joint distribution, and refines the lower bound using the estimated joint distribution

## **6. Tightening Bounds**

Bounds are tightened by taking the max of the lowest bounds across demographic characteristics. This process reduces the width of the bounds, providing a more accurate estimate of audience overlap.

## **7. Output Format**

The output file, 'result.csv,' contains columns such as 'tag,' 'date,' 'demo,' 'Visitors,' 'lower\_bound,' and 'refined\_lower\_bound.'

## **8. Implementation Details**

### **Step 1: Importing modules**

import pandas as pd

import numpy as np

import ast

"""

Estimate audience overlap across multiple demographic characteristics.

Steps:

1. Load and merge demographic and visitor data.

2. Calculate lower bounds for audience estimates.

3. Apply overlap estimation algorithm for each demographic characteristic.

4. Save results to a CSV file.

"""

### 

### **Step 2: Data Preprocessing / Loading**

"""

Load demographic and visitor data from CSV files, merge them based on 'tag' and 'date' columns.

"""

demo\_data = pd.read\_csv("./input/demo.csv")

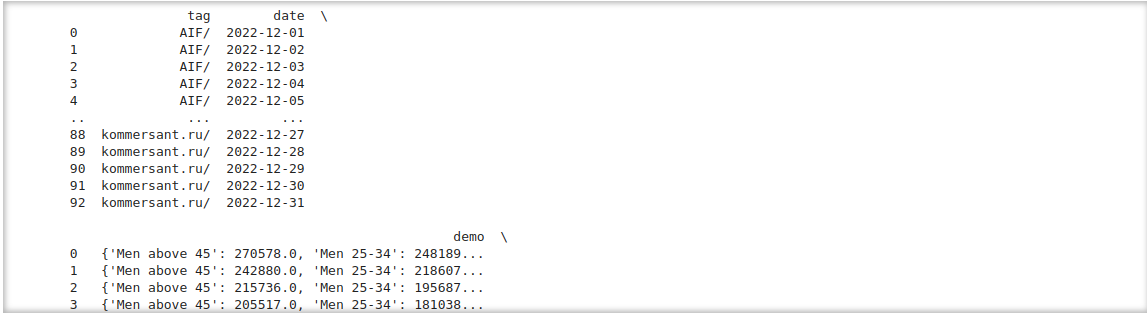
visitors\_data = pd.read\_csv('./input/all\_visitors.csv')

merged\_data = pd.merge(demo\_data, visitors\_data, on=['tag', 'date'])

# print(demo\_data)

# print(visitors\_data )

print(merged\_data)



### **Step 3: Calculate Lower Bounds**

"""

Calculate the lower bound of audience size for each row in the merged data.

"""

def calculate\_lower\_bound(row):

try:

"""

Calculates the lower bound of audience size for a given row.

Args:

row (pandas.Series): A row of data containing 'demo' and 'Visitors' values.

Returns:

int: The maximum value between 'demo' and 'Visitors', representing the lower bound.

"""

demo\_dict = ast.literal\_eval(row['demo'])

# Extract the relevant values from the 'demo' dictionary

demo\_value = sum(demo\_dict.values()) # Summing up all values in the dictionary

# Convert the 'Visitors' value to float

visitors\_value = float(row['Visitors'])

# Implement the lower bound calculation based on the formula

return max(demo\_value, visitors\_value)

except (ValueError, TypeError, SyntaxError):

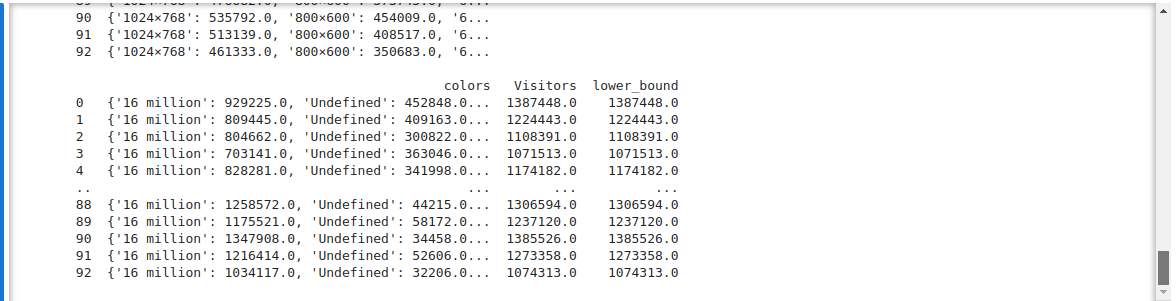
# Handle the case where conversion to float fails or 'demo' is not in the expected format

return np.nan # or any default value

# Apply the calculate\_lower\_bound function to create the 'lower\_bound' column

merged\_data['lower\_bound'] = merged\_data.apply(calculate\_lower\_bound, axis=1)

print(merged\_data)



### **Step 4: Algorithm to Estimate Audience Overlap**

def estimate\_overlap(df, demographic\_column):

"""

Estimates audience overlap for a given demographic characteristic.

Args:

df (pandas.DataFrame): The merged data with calculated lower bounds.

demographic\_column (str): The name of the demographic column to analyze.

"""

pairs = df.groupby(['tag', demographic\_column])[['lower\_bound']].max().reset\_index()

# Estimate joint distribution

joint\_distribution = pairs.groupby(demographic\_column)[['lower\_bound']].mean()

# Use the estimated joint distribution to refine the lower bound

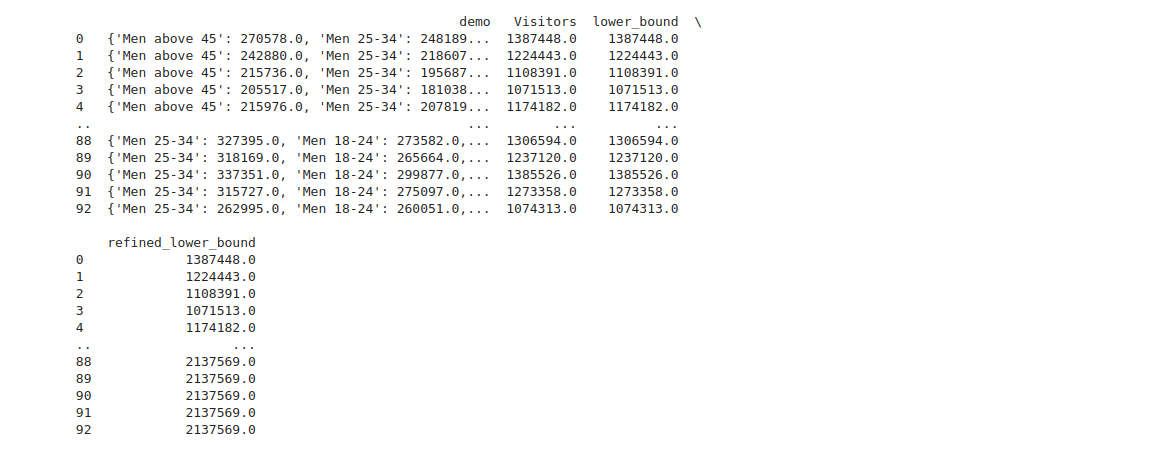
df['refined\_lower\_bound'] = df.apply(lambda row: max(row['lower\_bound'], joint\_distribution.loc[row[demographic\_column]]['lower\_bound']), axis=1)

### **Step 5: Apply the algorithm for each demographic characteristic**

demographic\_characteristics = ['browsers', 'resolutions', 'oses', 'languages']

for characteristic in demographic\_characteristics:

estimate\_overlap(merged\_data, characteristic)



### **Saving Results to CSV**

### """

### Save the final results to a CSV file named 'result.csv' without the index.

### """

### result\_file\_path = './output/result.csv'

### merged\_data.to\_csv(result\_file\_path, index=False)

## **9. Results**

The results show the tag, date, demo, Visitors, lower\_bound, and refined\_lower\_bound for each observation in the dataset. The refined\_lower\_bound provides a more accurate estimate of audience overlap.

## **10. Conclusion**

In conclusion, the algorithm successfully estimates audience overlap by leveraging detailed demographic information. The refined lower bounds provide a more accurate range for total audience estimates.