COGS9: Introduction to Data Science
Final Project
Due date: Wednesday 2019 December 11 23:59:59
Grading: 10% of overall course grade. 40 points total.
Completed as a group. One submission per group on Gradescope.

Group Member Information:

Please read COGS 9 team policies to best understand how to approach group work and to understand what the expectations are of you in COGS 9.

First Name	Last Name	PID
Hongping	Lu	A15551610
Xing	Hong	A15867895
Yutong	Luo	A15467566
Daimeng	Sun	A15562248

Question

What relationship we can find about the occupation changing of Asian to the text of a 100 years based on historical text and the actual proportion of occupation in America taken by Asians changed over the past 100 years? (1910-2010)

Hypothesis

We hypothesize that there would be a steady slow increase in the bias score (the sum of numerized association between each occupation and Asian) as well as the proportion of occupation in America taken by Asian during the first 40 years (1910-1950), because although during which the mass Asian migration happened and took place in the labor market, Asian immigrant workers were discriminated against. After the Immigration and Nationality Act in the 1960s, there should be a faster increase, and around 1980 as the second generation Asians had grown up, we expect a rapid increase in the bias score and in the proportion. Then the two curves would slowly flatten as the bias score and proportion reaches equilibrium.

Background Information

We come up with the idea to analyze the change in occupation of Asian workers because how Asian Americans developed in the labor market and how they changed occupations over time in the 20th century is a dynamic story. When they first came to the United States, Asians worked as an important part in the labor market because they were cheap and productive. However, after the number of Asian laborers began to grow in the United States, the dominant Americans developed hatred toward Asians because they thought

Asians deprived their job opportunities. For example, the first Chinese immigrant worked on railroad. The employers liked them because they are "silent, cheap but productive"(UCSD Hild 7A, lecture). As the industry wanted Asians as new and cheap labor, the white was angry at Asians for their race and job competition.(Wikipedia) Therefore, dominant Americans further developed discrimination against Asians, and their hatred developed to a series of movements to anti Asian. As a result, many Asians were pushed out of the professions and they became self-employed. Later on, Asians kept immigrating into the United States and they united as a community. The dominant Americans slowly changed their attitudes towards them. During the 21st century, when the public raised their awareness of racial equality, discrimination towards Asians in work is decreasing. So, again, the Asians have more opportunity to work in the United States.

We use word embedding to find the occupation taken by the Asians from historical text because the "language analysis is a standard tool used to discover, understand, and demonstrate"(Hamilton and Trolier, 1986; Basow, 1992; Wetherell and Potter, 1992; Holmes and Meyerhoff, 2008; Coates, 2015). Previous literature broadly establishes that language both reflects and perpetuates cultural stereotypes. As a result, it is reasonable to infer the relationship between words and Asian occupation descriptions.

Data

The perfect occupation census we need would include individuals' names, races, occupations, and would be the most representative of the population in the time period, meaning it would include all individuals, and all types of occupations. The perfect text analysis model we want would be trained by unbiased datasets and uses unbiased algorithms.

We need data about the proportion of Asian occupations in the dominant American labor market. We also need to extract the bias of Asians in the labor market from historical text through word embedding. In this case, at least 10000 observations that each observation represents the Asian occupation proportion from 1910 - 2010. Year, race, occupation name, all occupations in the labor market so that we can calculate Asian proportion in the labor market.

Because the occupation census doesn't include race of individuals, we try to distinguish Asian and White individuals by having typical Asian and White last name list. Thus, we find some representative historical datasets and trained models: Occupations, American Census data, Typical Asian & White last name list, trained text analysis model which captures the relative strength of association two groups.

The data we found:

A last names' list that the most common and well representing Asian & White. We will use this name list to limit the range of our target, which are only Asian or White throughout our project.

Whats The Most Common Name In America, Chalabi, M., and Flowers, A. (Filtered 20 top last names of Asian and by Google Books/COHA vectors) **20 names/each**

Occupation names list: **161 words**

What we want are lists of raw text occupations, clean enough for searching and grouping, and without phrase, combined words. We can use these words to find their vectors in embedding models and also retrieve them from the census.

1.Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes **115 words**

2.Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. **46 words**

Occupation Census Data:

Through the average proportion of Asians by 10 years bins, we can clearly see the changes in Asian occupation over the years. Ultimately, we will use the proportion to plot a line chart and compare it to the average bias scores (Asian vs White) over time for occupations in word embedding models.

The expected result is that the bias reflected in the text is consistent with the actual occupation census data.

Integrated Public Use Microdata Series.(IPUM) **9 variables**
72810 observations

Word Embedding models:

To express and analyze the correlation between different occupations and Asians, and how the correlation changes over time, we use the word embedding method with relative norm difference:

1. Genre-Balanced American English (1830s-2000s), SGNS and SVD (Google Books/COHA vectors)

   A trained embedding model from 1910-1990 on a combined corpus of genre-balanced Google books and the Corpus of Historical American English (COHA), for each decade, a separate embedding is trained from the corpus data corresponding to that decade.

2. Google News, word2vec

   (Since the COHA has only modeled from 1910-1990, we will use Google News embedding to rank the occupation of 1990-2010)

   To conclude, we will use the IPUM Census datasets to see how the occupation proportion of Asian changes, also using word embedding to find out the changes of bias in the historical text of the Asians occupation(compared to whites) over time.

Limitations:

1. The word lists may not be sufficient enough.
2. Adjectives list from the websites are messy, some of them are sentences, phrases, with other symbols.
3. IPUMS occupation list has a column of occupations with long, combined names, for example, Electrical-Engineers and Industrial-Engineers, we may have to hand-map it into separated, specific single words. For example, group as Engineer.
4. All the embeddings used are fully black-box, where the dimensions have no inherent meaning.

5. We need functions to compare the occupation associational strength between two groups rather than the raw percentages in IPUMS.
6. There is a large gap between the early and late data on observation quantity.

Ethical Considerations

Data Collection includes informed consent, Collection Bias, and Limit PII exposure. The occupation census we use might not cover all the individuals in America and it might miss some occupations, mainly the minorities. The occupation word list, similarly, will not include every occupation in America. The lack of representation might generate some degree of bias. To generate the word list, we use occupation words for which we have gender and ethnic subgroup information over time. Occupation percentages are obtained from the Integrated Public Use Microdata Series (IPUMS). We avoid collection bias by gathering a large amount of data from different reliable sources. Since the sources are public, and we do not collect individual data personally, the need for informed consent and limit PII exposure could be minimized.

Data Storage covers data security, right to be forgotten and data retention plan. We protect and secure data by keeping the data on our personal device and only shared within our group. For the right to be forgotten, since our data is from news, literature, and census, it will not be related to the individual having their information. If the data is no longer needed, we will delete the data from our own personal devices and the shared document within our group online.

Analysis comprises information on Missing perspectives, Dataset bias, Honest representation, Privacy in analysis and Auditability. Our analysis in very simple: only the proportion of occupation taken by Asian, calculated from the occupation census. Since we gathered a large amount of data, we could use data from other sources if there are missing perspectives. Since we are studying on the bias, the dataset bias helps us to learn the trend of level on the bias. As we are using public published data, we assume the representation is honest, and privacy in analysis and auditability has been addressed by the publishers.

Modeling, which we used to define the relative association between occupation and Asian by transforming text into vectors, lacks transparency. There might be biases in the model training parameter, or algorithms. The algorithms the model uses are COHA (Corpus of Historical American English)and SVD(singular value decomposition). However, the embeddings we use are among the most commonly used English embeddings, vary in the datasets on which they were trained, and between them cover the best-known algorithms to construct. They are all publicly available online; refer to the respective sources for in-depth discussion of their training parameters. The goal of our work will be showing the bias among time, which variables, observations will be sufficient to show all the aspects without bias. Also, we will be careful to avoid any kind of potential discriminatory interpretations.

Deployment involved Redress, Roll back, Concept drift, and Unintended use. If there is any discrimination or unfairness occurs in the study, we will remove the part involving these problems and try to find different data that could make up the missing part.

1.Requirements for assuring compliance by research institutions
UCSD Data Science Institute

2.Requirements for researchers' obtaining and documenting informed consent

All the data are public and widely used without concern

3. Requirements for Institutional Review Board (IRB) membership, function, operations, review of research, and record keeping.

The experimental conception has a sense of social responsibility and basic professional ethics. There is no experimental risk but may have some benefits to society.

4. Protections for certain vulnerable research subjects.

The samples are all historical texts from the United States that are not relevant to a particular individual and therefore do no harm to the subject.

Analysis Proposal

Data Collection

1. We want to get data from website by using HTTP or API. If they are not in nice spreadsheet format, we need to extract variables we need such as year, occupation, and race. Here, We could use the online tool provided by Integrated Public Use Microdata Series(IPUMS) to primarily select and download our full dataset of the occupation census which contains harmonized census and American Community Survey (ACS) data from 1790 to the present. Allow us to customize our variables and year range as well as file size, format.

Data Wrangling (Select, Clean, Missingness)

1. From the census data(IPUM), we clean data through picking the observations by our occupation name list and then built a new table with only the columns of “Occupation”, “Years” and “Race”. Our project requires data points that can apply for our calculation of Asian proportion among the sum of Asian and White. Hence, we first need to group the dataset by Race, in order to do the proportion. Then, we will split all the observations by bins of 10 years and labeled them by decades.
2. We need consistent data structure in our spreadsheet. For example, we will write the date in the format YYYY-MM-DD. We want to make sure there is no empty cell and put one thing in each cell. We will choose good variable names such as occupation, race and year to make it clean. The Occupation word list we are going to collect is unformatted, for better fitting it into the word embedding model and retrieve in our IPUM dataset, we will do word segmentation manually. The original occupation list contains phrases, specific symbols, and different formats. Through data cleaning and subtraction, we have a new list joint by our two original lists, without combined words, only separated, representative occupation names. Format and organized in one column. Hence, we will get tidy data. (The IPUM provided value codes for each occupation ,which allow us to manually allocate. Also the occupation of IPUM has different basis, we chose 1950s basis.)
3. Some of the observations are lacking variables, like occupation name or race. We will mark missingness in our census table to make sure no empty cell there.
4. After clean our data and transfer them into a table, we will save it in plain text files such as CSV.

Descriptive & Exploratory Data Analysis (Size, Shape, Trend, Oversampling, Plot, Distribution)

1. Due to the enormous size of file, most of the times we have to random sample then explore.
2. We will find the size of our data, looking at observations and variables. By grouping this 9 variable dataset by occupation and year to see the different proportion of occupations in our dataset, we will make interactive horizontal bar charts to show the shifts of occupations over the past 100 years.
3. Randomly pick a few words in our occupation word list and retrieve in IPUMS dataset, then plot the change of the single occupation proportion of Asians by line plot, we can see the relationship between years and specific occupations of Asians. For example, how the Asian engineer proportion change over the years. In this case, besides exploring our data, we can also explicitly show the trend of different occupations in favor of Asians.
4. Grouping the data by “Race” and “Years” in bins of 10 years is one of the ways to show the relationship between Race and Years. By doing so, the change of the proportion structure in different races can be shown on our interactive horizontal bar charts.
5. We need to find the shape of our data, to see if it is uniform, skewed, bimodal, bell-shaped or random so we could conclude a relationship between proportions of Asian labor from text and in reality. We also need to pay attention to any outliers that could influence our understanding.
6. We will find the mean and median the proportion of Asian labor. We will group the data by “Years” and have a table of two columns: Years and Count. By the calculation of mean and median and drawing the histogram, boxplot, normal possibilities plot of Count, we might discover that our census data is **skewed left**, which implies that the data is overall more dense by years. (Due to the large quantity gap between the early and late data, we selected the oversampling datasets of IPUM at the very beginning and oversampled it appropriately ourselves)
7. We also will look for the range, variance and standard deviation of our data to see if the proportion is changing greatly all the time. We can conclude from that to see if the change in occupation is dynamic.
8. For word embedding models, we will try out our average name vectors (group vectors) and see what kind of words are most relative to our groups. Also, we will compare occupation names with Asian group vectors to see how the relative strength of association changes in specific occupations. (Approaches will be discussed below.)

Statistical & Predictive Analysis (Word Embedding, Census proportion, Correlation, Regression)

1. **Word Embedding**: We want to discover the relationship between historical text data and Asian occupation changes over the years. In order to do this, we need to choose not to go through the usual methods of text analysis (like TF-IDF). Instead, we should select a state-of-the-art method in machine learning to approach, word embedding, as a new framework to measure, quantify, and compare trends in language over time. In word embedding models, each word in a given language is assigned to a high-dimensional vector such that the geometry of the vectors captures semantic relations between the words. The closer the vectors are, the more similar words are.

   Our project is about finding the association between Asian and our occupation names. To the light of word embeddings, we can easily find our corresponding vectors.
   Our approach is below:

Given two vectors, their similarity can be measured either by their negative difference norm.

$$normdiff\ (u, v) =-\ ||u - v||2$$

Bias in the embeddings, between two groups with respect to a neutral word list, is quantified by the relative norm difference, which captures the relative strength of association of a set of neutral words with respect to two groups.

$$\sum_{vj \in J} ||vj - v1||2 - ||vj - v2||2$$

Where **J** is the set of our Corresponding Occupation vectors in word embedding model.
v1 is the average vector for the Asian, which are the Last name vectors in the word embedding.
v2 is the average vector for white Last name vectors. We called **v1**, **v2** as their group vectors that can represent each group in word embedding space respectively.
Bias Score: Through the relative norm difference, we are able to separately calculate a total bias score of Asian to White in each year. The total bias score will reflect real-world Asian occupation in time. Because the model is trained on an annual historical text, which is one model per year, we pick the model every ten years, thus, we can have a bias score for Asians vs. white every 10 years.

2. **Census proportion:** The Census data can provide us the distribution of Asian occupation over time and the portion in the total occupation. We define the total occupation as Asian plus White, so the **proportion** of Asians is under the sum of Asian and White occupations.
3. **Correlation:** We hypothesize that the embedding could reflect social stereotypes that can be explained by occupation participation. We hope to see a high correlation between the bias score of Asian occupation vs white and the proportion of Asians in the Census. More positive indicates more Asians associated, for both the proportion of occupations and for relative distance. Each point is one occupation name. Therefore, we did Pearson's correlation test and got $p<0.05$ with r=some positive value. We can conclude that bias scores are significantly correlated with the Asian proportion of the occupation.
4. **Regression:** If the bias score and proportion have a positive correlation, we will find whether linear regression fits the relationship between them. We will conduct the regressions with occupation proportion as the independent covariate and the embedding bias scores as an outcome. By drawing the regression line, we can have an effect size of a positive value and a small standard error(SE) with $p<0.05$.

Data Visualization (Top 10, Compare, Scatter & Correlation, Word Cloud)

1. **Top 10:** By searching our occupation names in word embedding models, we will generate different occupation ranks which strongly associated with Asians over time. In order to show how the occupation stereotypes of Asian changes, we can make a complete list of Top 10 most Asian associated jobs in each decade.
2. **Compare:** We are going to make the double y-axis line plot of average bias score (Asian vs White in 10 years each)over time for occupations in word embedding models vs the average proportion. We suppose they will have similar patterns and fold points (the lines may also indicate some historical events).
3. **Scatter & Correlation:** We will make a scatter plot of our occupation names, which is Asian occupation proportion vs embedding bias scores. More positive would indicate more Asians based on both axes. Then we will do the correlation on this plot, with SE in gray indicated p-value and Pearson's r, and perform some highly/rarely associated occupation names beside the points. Where most of the points gathered will have less gray area.

4. **Word Cloud:** We will also make word clouds of each 10 years based on the proportion of Asians.

Discussion

Interpretation:

1. If the top 10 occupations changes, we can clearly see how Asian occupation changes, our prediction is that they will shift from labor positions to intellectual positions.
2. We predict that Census proportion line graph and word embedding bias score line graph will have a same pattern, which responses to specific historical events, they should reflect the same trends. Also, the correlation and p-value should confirm the referential relationship between these two. Then, we can say that historical text can reflect reality social changes, especially, occupation.
3. Similarly, by showing the most relevant occupation names in word embeddings and compare them with our Census Top 10, we can discover the differences between Asian occupation stereotypes and actuality.

Limitations: In the analysis, we only compare Asians with the whites, which could result in missing the proportion of other groups. However, we are planning to do this because there is a large difference between the proportion of the white and Asians in occupations so that we can see the trend of change. Also, the population of the white is relatively stable, which makes it an ideal comparing group. For the early text, there are few samples available for the wording embedding to collect data, which may not completely reflect the popular social attitude of that time. Therefore, the word embedding is less accurate as it has fewer samples to generate the model. Thus, the number of Asians in different occupations from the early historical text might be inaccurate that some Asians will not be included. One possible way to minimize the bias is to use a larger sample so that the effect of underestimation could be smaller. Another concern is that some Asians may use the name that the whites usually use, and it is hard to distinguish other races, which could lead to bias against since some Asians are included in the white. As the embeddings used are fully black-box, where we do not know exactly how it works, it is hard to fix the problem from the root. The only approach we could access is to use a big sample to decrease the bias.

Addressing ethical consideration: The results we get depend on the data sources and the metrics/algorithms we choose to represent bias or association. To make sure we have the data sources and metrics/algorithms to get the unbiased results, we look at other word-embedding papers and choose the ones we use for its reliability and simplicity. Another concern is the dependency of our result on the specific word lists we choose. It is difficult to generate the most representative word list. We go through occupation census from several sources and combine them to make sure our word list misses least.

Group Participation

Xing joined the discussion about project direction, data fetching, clean up the data set, search algorithms, some writing work. Hongping did the ethical consideration part and the discussion part about the limitations of the analysis proposal. She also modified the analysis proposal and the background to fix some minor problems. Yutong did the background part about why we came up with the hypothesis and analysis on our topic. She took part in the data part and contributed to the analysis proposal part. Daimeng came up with the hypothesis draft, edited the draft addressing the problems in the project proposal.