Predicting Race and Ethnicity From Sequence of Characters in a Name[*]

Gaurav Sood[†] Suriyan Laohaprapanon[‡]

Abstract

To answer questions about racial inequality, we often need the ability to reliably infer the race of a person based on their name. The census bureau provides common last names and proportion of people belonging to different races who have the last name. But there is more information in the first names. To estimate the relationship between full names and race, we exploit the Florida voting registration data, and the Wikipedia data collected by Skiena and colleagues, to predict race and ethnicity using Long Short Term Memory Networks. Our out of sample precision and recall is .83 and .84 respectively. This compares to OOS recall of .78 and .79 for last name only models. We also provide a Python package to easily predict the race and ethnicity of names. We apply our method to the campaign finance data to estimate the share of donations made by people of various racial groups and investigate whether people are more likely to contribute to co-ethnics conditional on ideology.

[*] Data and scripts behind the analysis presented here can be downloaded at: http://github.com/appeler/ethnicolr.

[†] Gaurav can be reached at gsood07@gmail.com

[‡] Suriyan can be reached at: suriyant@gmail.com

How often are people of different races and ethnicites covered in the news? How often do African Americans contribute to political campaigns? To answer these questions and questions like these, we often need a way to infer race and ethnicity from names. Given the important questions that could be answered if we had a reliable way to do such mapping, a variety of attempts have been made to infer race from names. For instance, recently Imai and Khanna (2016) presented a way to infer ethnicity from last names by using geographic data along with the census data. In this paper, we contribute to this broad literature.

We exploit the US census data, the Florida voting registration data, and the Wikipedia data collected by Skiena and colleagues (Ambekar et al. 2009) to learn a model between sequence of characters in a name and race and ethnicity. We use Long Short Term Memory to learn this association. The granularity at which we predict the race depends on the dataset. For instance, in Ambekar et al. (2009) data, the race and ethnicity is coded fairly granularly at a geogrpahic ethnic group level, while the census data we use in the model, the we only categorize between Non-Hispanic Whites, Non-Hispanic Blacks, Asians, and Hispanics. Our out of sample precision and recall is .83 and .84 respectively. This compares to OOS recall of .78 and .79 for last name only models. We also provide a Python package to easily predict the race and ethnicity of names. Lastly, we apply our method to the campaign finance data to estimate the share of donations made by people of various racial groups and investigate whether people are more likely to contribute to co-ethnics conditional on ideology.

1 Method

If you picked a random individual with last name Smith from the US in 2010 and asked us to guess this person's race (measured as crudely as by the census), the best guess would be based on what is available from the aggregated Census file. It is the Bayes Optimal Solution.

So what good are last name only predictive models for? A few things. If you want to impute ethnicity at a more granular level, guess the race of people in different years (than when the census was conducted if some assumptions hold), guess the race of people in different countries (again if some assumptions hold), when names are slightly different (again with some assumptions), etc. The big benefit comes from when both the first name and last name is known.

2 Data, Model, and Validation

The Census Bureau provides frequency of all surnames occurring 100 or more times for the 2000 and 2010 census. Technical details of how the 2000 and 2010 data were collected can be found on the census website. The Wikipedia data were originally collected by a team lead by Steven Skiena as part of the project to build a classifier for race and ethnicity based on names. The team scraped Wikipedia to produce a novel database of over 140k name–race associations. For details of the how the data was collected, see Ambekar et al. (2009). The third dataset is the Florida Voting Registration data for the year 2017. The Florida Voting Registration data has information about voter's race.

Race	Count
asian	253808
hispanic	2179106
nh_black	1853690
nh_white	8757268

We can use the Wikipedia data and the Florida Voting Registration data as is but the Census data needs to be transformed before being used. The dataset that the Census Bureau issues aggregates data for each last name and provides percentage of people with the last name who are Black, White, Asian, Hispanic, etc. Given some names are more common than others (Smith is the last name of 2,376,206 Americans), and given our interest in modeling the population distribution, we take a weighted random sample from this data with weight equal to how common the last name is in the population. Next, we assign race to name roughly in proportion to how the name is distributed across the racial groups. We assign floor of pctwhite as whites, floor of pctblacks as blacks etc. And we lose the one or two or few observations as we are using floor. We use this as the final dataset.

For our full name model, we concatenate the last name and first name and capitalize the first character of all the words. We next split the name into two character chunks (bi-chars). For instance, Smith becomes Sm, mi, it, and th. We then remove infrequent bi-chars (occurring less than 3 times in the data) and very frequent bi-chars (occurring more than 30% in the dataset). We use the remaining bi-chars as our vocabulary. We pad the sequences so that they are the same size. And we use 20 as the window size for the last name only model and 25 for the full name model. Lastly, we split the data randomly into train (80%) and test (20%). We next train a LSTM model.

The table 1 presents some metrics that shed light on how well we did with the last name only model in predicting race OOS. The table 2 presents some metrics that shed light on how well we did with the full name only model in predicting race OOS.

race	precision	recall	f1-score	support
asian	0.77	0.35	0.48	50762
hispanic	0.80	0.81	0.80	435821
nh_black	0.68	0.13	0.22	370738
nh_white	0.80	0.94	0.86	1751454
avg / total	0.78	0.79	0.75	2608775

Table 1: Performance of the Last Name LSTM Model on the Florida Voter Registration Data.

race	precision	recall	f1-score	support
asian	0.77	0.45	0.56	50762
hispanic	0.83	0.84	0.83	435821
nh_black	0.74	0.45	0.56	370738
nh_white	0.86	0.93	0.89	1751454
avg / total	0.83	0.84	0.83	2608775

Table 2: Performance of the Full Name LSTM Model on the Florida Voter Registration Data.

3 Application

To illustrate how the package can be used, we impute the race of the campaign contributors recorded by FEC for the years 2000 and 2010 and tally campaign contributions by race.

References

Ambekar, Anurag, Charles Ward, Jahangir Mohammed, Swapna Male and Steven Skiena. 2009. Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining.* ACM pp. 49–58.

Imai, Kosuke and Kabir Khanna. 2016. “Improving ecological inference by predicting individual ethnicity from voter registration records.” *Political Analysis* 24(2):263–272.