learning from names
gaurav and suriyan

Why Learn From Names?

- Because sometimes all you have is names

- Media, most voter lists, . . .

Recause cometimes all you have is names

- Because sometimes all you have is names

- Media, most voter lists, ...

- Passusa samatimas all vau bava is namas
 - Decause sometimes an you have is names

WASHINGTON -- Some of Robert S. Mueller III's investigators have told associates that Attorney General William P. Barr failed to adequately portray the findings of their inquiry and that they were more troubling for President Trump than Mr. Barr indicated, according to government officials and others familiar with their simmering frustrations.

Source: The New York Times

WASHINGTON -- Some of Robert S. Mueller III's investigators have told associates that Attorney General William P. Barr failed to adequately portray the findings of their inquiry and that they were more troubling for President Trump than Mr. Barr indicated, according to government officials and others familiar with their simmering frustrations.

Source: The New York Times

- Because sometimes all you have are names
- Media, most voter lists, . .

- Because sometimes all you have are names
- Media, most voter lists, . . .,
- But you want to:

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases
 - Prevent biases (Regress Out)

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases
 - Prevent biases (Regress Out)
 - Personalize

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases
 - Prevent biases (Regress Out)
 - Personalize
- Concretely:

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases
 - Prevent biases (Regress Out)
 - Personalize
- Concretely:
 - Biases in media coverage: who is (covered) cited as a source?, . . .

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases
 - Prevent biases (Regress Out)
 - Personalize

- Concretely:

- Biases in media coverage: who is (covered) cited as a source?....
- Fairness in lending

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases
 - Prevent biases (Regress Out)
 - Personalize

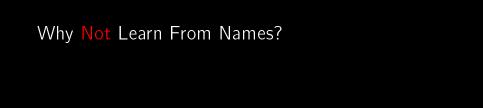
- Concretely:

- Biases in media coverage: who is (covered) cited as a source?....
- Fairness in lending
- Political accountability: whose emails are read?, ...

- Because sometimes all you have are names
- Media, most voter lists, . . .
- But you want to:
 - Highlight biases
 - Fight biases
 - Prevent biases (Regress Out)
 - Personalize

- Concretely:

- Biases in media coverage: who is (covered) cited as a source?....
- Fairness in lending
- Political accountability: whose emails are read?, ...
- Personalization—recommending same race doctor (Alsan et al. 2018)



- Instrument for Discrimination

- Instrument for Discrimination
- But that risk exists for credit risk scoring too.

- Instrument for Discrimination
- But that risk exists for credit risk scoring too.
- To enable predatory lending

- Instrument for Discrimination
- But that risk exists for credit risk scoring too.
- To enable predatory lending
- But still a concern . . .

we?

we?

- p (ethnicity | name)?

- p (ethnicity | name)?
- Classify to the majority class

- p (ethnicity | name)?
- Classify to the majority class
- Makes the least mistakes—Bayes Optimal Error

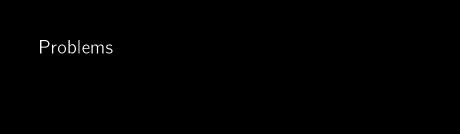
- p (ethnicity | name)?
- Classify to the majority class
- Makes the least mistakes—Bayes Optimal Error
- Census Last Name Dataset

- p (ethnicity | name)?
- Classify to the majority class
- Makes the least mistakes—Bayes Optimal Frror
- Census Last Name Dataset
 - Split by race for 1,000 most common last names

SURNAME	RANK	FREQUENCY (COUNT)	PROPORTION PER 100,000 POPULATION	CUMULATIVE PROPORTION	PERCENT NON- HISPANIC OR LATINO WHITE ALONE	PERCENT NON- HISPANIC OR LATINO BLACK OR AFRICAN AMERICAN ALONE
SMITH	1	2,442,977	828.2	828.2	70.9	23.1
JOHNSON	2	1,932,812	655.2	1,483.4	59.0	34.6
WILLIAMS	3	1,625,252	551.0	2,034.4	45.8	47.7
BROWN	4	1,437,026	487.2	2,521.6	58.0	35.6
JONES	5	1,425,470	483.2	3,004.8	55.2	38.5

- p (ethnicity | name)?
- Classify to the majority class
- Makes the least mistakes if that is all you have
- Census Last Name Dataset
 - Split by race for 1,000 most common last names

- p (ethnicity | name)?
- Classify to the majority class
- Makes the least mistakes if that is all you have
- Census Last Name Dataset
 - Split by race for 1,000 most common last names



Problems

Opportunities

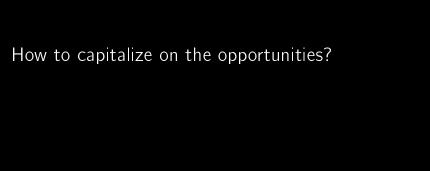
Opportunities - Classifying the other 160k+ last names

- Classifying the other 160k+ last names
- When we have more than the last name

- Classifying the other 160k+ last names
- When we have more than the last name e.g., African-Americans have more distinctive first names

- Classifying the other 160k+ last names
- When we have more than the last name e.g., African-Americans have more distinctive first names
- When we have minor variations

- Classifying the other 160k+ last names
- When we have more than the last name e.g., African-Americans have more distinctive first names
- When we have minor variations
- Geographic variation



How to capitalize on the opportunities? - Patterns in names

- Patterns in names
- Capturing sounds, sequence of sounds, ...

- Patterns in names
- Capturing sounds, sequence of sounds, . . .
- Phonemes are dead. Long live bi-chars!

- Patterns in names
- Capturing sounds, sequence of sounds, ...
- Phonemes are dead. Long live bi-chars! da_as_sh_ia_an,

- Patterns in names
- Capturing sounds, sequence of sounds, ...
- Phonemes are dead. Long live bi-chars! da_as_sh_ia_an, jo_oh_ha_an_ns_so_on

- Patterns in names
- Capturing sounds, sequence of sounds, ...
- Phonemes are dead. Long live bi-chars! da_as_sh_ia_an, jo_oh_ha_an_ns_so_on
- Patterns in communication networks, e.g., Skiena et al. 2018.

How to Classify Text - Text → Embeddings → Classifier

- Text → Embeddings → Classifier
 - Embeddings leverage the adage:
 You are the company you keep.

- Text → Embeddings → Classifier
 - Embeddings leverage the adage: You are the company you keep.
 - Use a large corpus

- Text → Embeddings → Classifier
 - Embeddings leverage the adage: You are the company you keep.
 - Use a large corpus
 - Learn the context well

- Text → Embeddings → Classifier
 - Embeddings leverage the adage: You are the company you keep.
 - Use a large corpus
 - Learn the context well
 - Preserve tens of hundreds of vectors and pass them to a model

- Text → Embeddings → Classifier
 - Embeddings leverage the adage You are the company you keep.
 - Use a large corpus
 - Learn the context well
 - Preserve tens of hundreds of vectors and pass them to a model
- In Our Case:

- Text → Embeddings → Classifier
 - Embeddings leverage the adage You are the company you keep.
 - Use a large corpus
 - Learn the context well
 - Preserve tens of hundreds of vectors and pass them to a model
- In Our Case:
 - Embeddings of common bi-chars

- Text → Embeddings → Classifier
 - Embeddings leverage the adage: You are the company you keep.
 - Use a large corpus
 - Learn the context well
 - Preserve tens of hundreds of vectors and pass them to a model
- In Our Case:
 - Embeddings of common bi-chars
 - LSTM

- How would I describe myself?

- How would I describe myself?
Indian? But there is systematic variation across names across linguistic groups within India.

- How would I describe myself?
 Indian? But there is systematic variation across names across linguistic groups within India.
- Are you willing to fight for the group?

- How would I describe myself?
 Indian? But there is systematic variation across names across linguistic groups within India.
- Are you willing to fight for the group?
- How people describe themselves in free text?

- How would I describe myself?
 Indian? But there is systematic variation across names across linguistic groups within India.
- Are you willing to fight for the group?
- How people describe themselves in free text?
- What do people choose when asked to force-fit into administrative categories?

Data

- Voting Registration data from Florida

Data

Data

- Race/Ethnicity = Asian or Pacific Islander,

- Voting Registration data from Florida

Hispanic, NH Black, NH White

Data

coded race

- Voting Registration data from Florida
- Race/Ethnicity = Asian or Pacific Islander,
 Hispanic, NH Black, NH White
- Hispanic, NH Black, NH White
 Wikipedia ∼ Famous people but more finely

Success Using Florida Voter Registration Data

Success Using Florida Voter Registration Data

		F1-Score		
		Last Name	Fu∥ Name	
Race	Asian	.54	.60	
	Hispanic	.72	.75	
	NH Black	.32	.55	
	NH White	.88	.90	

Success Using Florida Voter Registration Data

		F1-Score		
		Last Name	Fu∥ Name	
Race	Asian	.54	.60	
	Hispanic	.72	.75	
	NH Black	.32	.55	
	NH White	.88	.90	

Donations to Political Campaigns in 2000 and 2010 by Race and Ethnicity

	2000	2010	2000	2010
asian	2.2%	2.7%	2.0%	2.3%
black	11.0%	10.2%	8 9%	7 9%

83.5% 82.7% 85.8% 86.5%

Census

hispanic

white

Florida

	2000	2010	2000	2010
asian	2.2%	2.7%	2.0%	2.3%
black	11.0%	10.2%	8.9%	7.9%

Census

83.5% 82.7%

hispanic

white

Florida

85.8% 86.5%

Python Package

```
> import pandas as pd
> from ethnicolr import census_In, pred_census_In
Using TensorFlow backend.
> names = ['name': 'smith',
... 'name': 'zhang',
... 'name': 'jackson']
> df = pd.DataFrame(names)
> census_In(df, 'name', 2010)
```

name	race	pctwhite	pctblack	pctap
smith	white	70.9	23.11	0.5
zhang	арі	0.99	0.16	98.06
jackson	black	39.89	53.04	0.39