learning from names

gaurav and suriyan

Why Learn From Names?

Why Learn From Names?

- Because sometimes all you have is names

Why Learn From Names?

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

Why Learn From Names?

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

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

Why Learn From Names?

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

Why Learn From Names?

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

Why Learn From Names?

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

Why Learn From Names?

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

Why Learn From Names?

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

Why Learn From Names?

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

Why Learn From Names?

- 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:

Why Learn From Names?

- 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?, ...

Why Learn From Names?

- 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

Why Learn From Names?

- 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?, . . .

Why Learn From Names?

- 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)

Why Learn From Names?

Why Not Learn From Names?

Why Not Learn From Names?

- Instrument for Discrimination

Why Not Learn From Names?

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

Why Not Learn From Names?

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

Why Not Learn From Names?

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

But if we were to learn from names, how would we?

But if we were to learn from names, how would we?

- p (ethnicity | name)?

But if we were to learn from names, how would we?

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

But if we were to learn from names, how would we?

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

But if we were to learn from names, how would we?

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

But if we were to learn from names, how would we?

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

But if were to learn from names, how would we?

- 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

But if were to learn from names, how would we?

- 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

Problems

Opportunities

Opportunities

- Classifying the other 160k+ last names

Opportunities

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

Opportunities

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

Opportunities

- 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

Opportunities

- 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?

How to capitalize on the opportunities?

- Patterns in names

How to capitalize on the opportunities?

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

How to capitalize on the opportunities?

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

How to capitalize on the opportunities?

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

How to capitalize on the opportunities?

- 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`

How to capitalize on the opportunities?

- 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 $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.

How to Classify Text

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

How to Classify Text

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

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ 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

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ 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:

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ 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

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ 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

But what do you mean by race and ethnicity?

But what do you mean by race and ethnicity?

- How would I describe myself?

But what do you mean by race and ethnicity?

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

But what do you mean by race and ethnicity?

- 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`?

But what do you mean by race and ethnicity?

- 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?

But what do you mean by race and ethnicity?

- 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

Data

- Voting Registration data from Florida

Data

- Voting Registration data from Florida
- Race/Ethnicity = Asian or Pacific Islander, Hispanic, NH Black, NH White

Data

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

Success Using Florida Voter Registration Data

Success Using Florida Voter Registration Data

		F1-Score	
		Last Name	Full 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	Full 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

	Census		Florida	
	2000	2010	2000	2010
asian	2.2%	2.7%	2.0%	2.3%
black	11.0%	10.2%	8.9%	7.9%
hispanic	3.2%	4.3%	3.2%	3.3%
white	83.5%	82.7%	85.8%	86.5%

	Census		Florida	
	2000	2010	2000	2010
asian	2.2%	2.7%	2.0%	2.3%
black	11.0%	10.2%	8.9%	7.9%
hispanic	3.2%	4.3%	3.2%	3.3%
white	83.5%	82.7%	85.8%	86.5%

Python Package

```
> import pandas as pd
> from ethnicolr import census_ln, pred_census_ln
Using TensorFlow backend.
> names = [‘name’: ‘smith’,
... ‘name’: ‘zhang’,
... ‘name’: ‘jackson’]
> df = pd.DataFrame(names)
> census_ln(df, ‘name’, 2010)
```

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