What's in a name?

Data matching and linkage with ethnic data

Tatiana Nobels Lee + Phan Quốc Thái

Tatiana

Tatiana (she/her) is currently the Data Manager at VOCAL-NY where her primary responsibilities involve database management, training, and working with the organizing team to improve & develop data systems. She previously held several administrative and operations roles at the organization from 2012-2016. Tatiana is a native of The Bronx, NY and now lives near Seattle, WA with her husband and recently adopted dog.

Phan Quốc Thái

Tim Phan is a data analyst from Los Angeles who works in political campaigns. He most recently managed a city council race in Los Angeles in 2022, and has also worked in Democratic / advocacy organizations in previous cycles. He's quite passionate about house music and activist photography.

What is the problem?

- Algorithmic biases against non-Anglicized names makes it more difficult for ethnic names to be properly matched against the voter file.

- There are many ethnic communities who do not conform to a traditional “first name / last name” convention, possess diacritical marks like accents, or are rooted in non-Latin alphabets like Arabic or Chinese, which necessitates correct usage of tones.

- While many people in those communities may have an Anglicized or shorthand nickname of their given name for official records, this is less true in foreign-born and immigrant communities.

Voter File Data

- Lack of a unified national voter file (unlike Canada)
- Role of private organizations like NGP-VAN, Catalist and relationship to the Democratic Party
- “Funnel” of progressive orgs and campaigns

What is Data Matching?

- A process for matching records in one set of data with records in a separate set of data
  - Which two records from each data set represent the same **entity**

- Records can represent:
  - Voters
  - Customers
  - Patients
  - Inventory (items)

Anna Thorson, 2022
Based on *Data Matching* by Peter Christen, Springer, 2012

Problems with Data Matching

- It’s hard to get right!
- Aiming for the “least wrong” result
- Lack of unique identifier
- Computational complexity

Christen, P. (2014). Aims and Challenges of Data Matching. In *Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection* (pp. 3–6). essay, Springer Berlin.

Examples of Data Matching

Are they a match?

Record 1:

first	middle	last	street	city	state	age
chi ho		park	123 main st	new york	NY	47

Record 2:

first	middle	last	street	city	state	dob
chi	ho	park	123 main st	new york	NY	3/24/1975

Are they a match?

Record 1:

first	middle	last	street	city	state	age
chi ho		park	123 main st	new york	NY	47

Record 2:

first	middle	last	street	city	state	dob
park		chi ho	123 main st	new york	NY	

Are they a match?

Record 1:

first	middle	last	street	city	state	age
chi ho		park	123 main st	new york	NY	47

Record 2:

first	middle	last	street	city	state	dob
chi	h	park	879 broadway	new york	NY	

Pew Research Center's Ruth Igielnik + Senior Survey Advisor Scott Keeter

"You found that the five voter files collectively were more complete and accurate in matching to our American Trends Panel data than any individual file. What lessons should researchers and other users of voter files draw from that?"

There's a trade-off between matching accuracy (i.e., do you believe the voter file match received is the correct person?) and coverage (i.e., the percent of people able to be matched).

- Individually the files match rate is 50-79%
    - Lower match rates appear more accurate, but disproportionately exclude younger and mobile people
- The 5 voter files had a combined match rate of 91%
    - High match rates produced more representative samples but may have included more inaccurate matches.

DeSilver, D. (2020, August 27). Q&A: The growing use of 'voter files' in studying the U.S. electorate. Pew Research Center. Retrieved January 24, 2023, from https://www.pewresearch.org/fact-tank/2018/02/15/voter-files-study-qa/

Solution

- In the immediate short-term, we've created an a "data dictionary" of common cultural names can be matched that will live on GitHub.

- The dictionary has two parts:
  - Given names and nicknames
  - Anglicized versions of given names

- Add to & share the dictionary: rebrand.ly/NameForm

 **onyxrev** initial batch of names

Latest commit `098ddb6` on Apr 12, 2013 **History**

1 contributor

1432 lines (1432 sloc) | 27.3 KB

Raw Blame

Search this file...

	id	name	nickname
2	1	Aaron	Erin
3	2	Aaron	Ron
4	3	Aaron	Ronnie
5	4	Abel	Ab
6	5	Abel	Abe
7	6	Abel	Eb
8	7	Abel	Ebbie
9	8	Abiel	Ab
10	9	Abigail	Abby
11	10	Abigail	Gail
12	12	Abigail	Nabby
13	13	Abner	Ab
14	14	Abraham	Ab

Next Steps & Additional Considerations

- An API where organizations can connect their databases to an open-source service in which they can match up
- Impact on ethnicity modeling with machine learning
- Cultural competency “fact sheet” for folx working with this data
- Role of Secretary of States / government admin data
- What is the role of VAN / Catalist / the DNC, etc