Nationality Classification Using Name Embeddings (Notes)

1. Existing name-based nationality classifiers use **name sub-strings** that are trained on small unrepresentative sets of labeled names – therefore, they are inefficient
   1. These use decision trees and HMM (Hidden Markov Model) for their algorithms
2. We can turn towards **name embeddings** instead, which encodes gender, ethnicity, and nationality
   1. F1 Score (test regarding precision/accuracy) is 0.795
   2. Closest competitor F1 score is 0.580
3. **NamePrism** – a name nationality and ethnicity classifier
   1. relies on the phenomenon of homophily in communication rather than substring features
      1. “birds of a feather flock together”
   2. Uses word embedding methods as classifiers for nationality and ethnicity
   3. Uses a Naïve Bayes approach (family of Bayesian Probability)
   4. <https://www.name-prism.com/> (Free web-based API, need to request API token however)
      1. **\*\*\* Gender classification not available on web service**
   5. Substrings are limited to phonograms, while name embeddings with logograms as well
      1. Name embeddings are derived from word embeddings, which are used in NLP (Natural Language Processing)
   6. 39 leaf nationalities and 6 U.S. ethnicities
4. Name embedding is a variation of word embedding
   1. Vector representations for two words co-occurring frequently in their contexts
      1. In articles, the context would be the words around it
      2. Using a homophily principle, context is generated in contact lists of the names being analyzed
         1. Contacts are weighed by frequency of communications
         2. Names with large weights tend to have the same nationalities
            1. Ordering of contact lists is not useful because account holders (names) are mutually independent
   2. 1990 U.S. census data is used for labeling popular first names
   3. 2000 U.S. census data is used for labeling popular last names
5. **It is possible to request a dataset of Twitter Name Embeddings** (3 million embeddings for Twitter data)
6. Nationality Classification Methodology
   1. Naïve Bayes Model is uses for its effectiveness / interpretability
   2. Name nationalities depend on both first and last name (last names can be similar in many countries)
7. Nationality Taxonomy Construction
   1. Based on the Cultural, Ethic and Linguist (CEL) taxonomy
   2. Similarities between countries are computed using name parts distributions
      1. Ex: Hispanic countries are divided into Spanish, Portuguese, and Philippines
8. Datasets
   1. 68 million full name and nationality pairs from email sources and 6 million from Twitter
9. Performance
   1. NamePrism performs the better on large data sets from the web, as opposed to other competitors whose data sets rely on labeled names from Wikipedia entries
      1. For our research, it is better that NamePrism is more accurate on more realistic data sets
10. Ethnicity Classification
    1. Ethnicity data is gained from U.S. Census Bureau data

Other Resources

* Ioan Voicu, Using First Name Information to Improve Race and Ethnicity Classification, Statistics and Public Policy, 5:1, 1-13, DOI: [10.1080/2330443X.2018.1427012](https://doi.org/10.1080/2330443X.2018.1427012) (2018)
  + <https://www.tandfonline.com/action/showCitFormats?doi=10.1080%2F2330443X.2018.1427012>
* Tzioumis, K. Demographic aspects of first names. *Sci Data* **5,**180025 (2018).
  + <https://www.nature.com/articles/sdata201825#Sec16>
* Treeratpituk, Pucktada and C. Lee Giles. “Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching.” AAAI (2012).
  + <https://clgiles.ist.psu.edu/pubs/AAAI2012-name-ethnicity.pdf>

Note that these sources above offer research and information about the classifications of gender, nationality, and ethnicity but do not provide a method for testing with our data.

Public APIs

* NamSor API
  + <https://v2.namsor.com/NamSorAPIv2/index.html>
* BehindTheName API
  + <https://www.behindthename.com/api/>