BT

How the Sequence of Characters in a Name Can Predict Race and Ethnicity

Gaurav Sood, principal data scientist at Microsoft, recently **spoke** at the **AnacondaCon 2019 Conference** on how the sequence of characters in a person's name can be used to predict that person's race and ethnicity, using machine learning techniques.

Learning about names helps with real-world use cases like fairness in lending of loans and user personalization. Sood talked about how **precision and recall** model evaluation metrics are enhanced by modeling the relationship between race/ethnicity and the sequence of characters in a name, using Long Short Term Memory (**LSTM**) Networks. Some of the data used for this analysis came from **US Census** Last Name Dataset, **Florida voting registration data** and **Wikipedia** data.

He also talked about how to capitalize the opportunities by using the following techniques:

Patterns in names
Use Bi-chars instead of Phonemes
Patterns in communication networks
Use a large corpus and learn context well
Preserve a few hundred vectors and pass it to a model

InfoQ spoke with Sood about the conference talk and what we can learn from names using prediction modeling techniques.

InfoQ: Can you discuss how we can learn from names? What ML/DL algorithms can we use?

> ***Gaurav Sood:*** *Learning more about a person from their name is no different from tackling any other supervised ML problem. It all starts with getting (or creating) a large labeled corpus. For instance, one key innovation in* ***ethnicolr*** *is the training data---we use voting registration files to get a large labeled corpus. In* ***another project*** *on learning from names, I scraped Google Image Search results to build the training data for inferring the gender from a name.*
>
> *Once you have the data, find ways to exploit patterns in the data to learn a model. The early ventures exploited the fact that names of different kinds of people began/ended differently. For instance, female names in India often end with an 'a,' and you can exploit that pattern to infer gender from Indian names. In ethnicolr, we generalize this intuition and use patterns in sequences of characters. (I am also working on exploiting sequences of sounds.) Like* ***Ye et al.****, you could also rely on the fact that we correspond more frequently with co-ethnics and exploit email networks for building your models.*
>
> *To exploit the patterns in the data, the full-range of DL/ML tools is available to you. Use what works best.*

InfoQ: Can you talk about phonemes and bi-chars and how they helped in the data analysis?

> ***Sood:*** *A phoneme is a unit of speech in a language. (There are 39 phonemes in English. If you are looking to get intuition, on the* ***CMU site*** *you can decompose a large list of words into constituent phonemes.) Before the era of end-to-end DL, for audio transcription, we would first get phonemes from audio waves and then learn the relationship between sequences of phonemes and words.*
>
> *Bi-chars are another (noisy but easy) way to represent sounds and structure in a word. And sequences of common bi-chars can capture our intuitions reasonably. For instance, "ashian" (think Kim Kardashian) is a common sequence in the last names of Armenians. And "ashian" can be broken down into the following bi-chars (as, sh, hi, ia, an) and we can learn to associate that sequence with the relevant race/ethnicity.*

InfoQ: What all technologies did you use in this case study?

> ***Sood:*** *We used **Python** as the programming language, **scikit-learn** to split our names into bi-chars, and **Tensorflow** with **Keras** interface to learn the embeddings and to apply LSTM.*

InfoQ: How can our readers learn more about your project and try it in their development environments?

> ***Sood:*** *The code for doing a plain vanilla version of the **project** and **models** are available on the website. If you want to get the intuition, the **paper and presentation** are likely to be of help. If you want to use the open-source package to answer a question, the **code** for our campaign contribution application may prove useful.*

BT