Nationality Prediction of an Individuals Using their First Name

Sakil Sarker
2017-2-60-079
2017-2-60-079@std.ewubd.edu

Sayeda Fatema Tuj Zohura
2017-1-60-011
2017-1-60-011@std.ewubd.edu

Md.Abuhorayra
2017-2-60-077
2017-2-60-077@std.ewubd.edu

September 26, 2020

Contents

Abstract

Is it possible to predict nationality? To answer questions about 'Ethnicity or Nationality', we often look at individuals names first. The majority works until now has been conducted considering the last name. But in most of the cases, the first name holds much information and really matters a lot. To make a relationship between first name and personality, we collected some names from internet sources and labeled them with their desired nationality. In this project, we developed a machine learning model to deploy a system that can predict the Nationality of individuals from their first names. Here we developed different types of models and then tested them with machine learning techniques to identify the best model among them.

1 Introduction

Nationality is, a people sharing a common origin, culture, and/or language, and possibly constituting a nation-state (Your Dictionary). Nationality is also a legal identification of a person in international law, establishing the person as a subject, a national, of a sovereign state. It affords the state jurisdiction over the person and affords the person the protection of the state against other states (Wekipedia). Name analysis is often the only practical way to gather nationality annotations because there are also privacy concerns. Yet, nationality remains one of the fundamental attributes that can not be easily masked like age and gender even in disguise. In this project, we collected data contain first names and corresponding nationalities through which we will try to predict an individual's nationality which was not in the data set. To do this, we used supervised learning which actually infers a function from labeled training data consisting of a set of training examples. We tried to make nationality prediction work much accurate so it can accurately label the nationality of an individual.

1.1 Objective

1. To collect an average numbers of data from different nationality. 2. To create a model and fit those data to the model. 3. Run machine learning algorithm. 4. Check whether our model can predict nationality or not.

1.2 Motivation

1. Nationality prediction can be a key component that can be deployed in various systems at security checkpoints where they have the list of names [1]. 2. Nationality identification may unlocks important demographic information, with many applications in biomedical and sociological research [1]. 3. From social media posts and comments, by applying nationality prediction may demonstrate stark differences in the nationality of the followers of public figure such as Trump and Obama, and in the sports and entertainments favored by different groups. 4. By applying this, e-commerce businessman's can make targeted add.

1.3 Existing Works

Names data is easy to use and as a result, analyzing names data is a lucrative opportunity for researchers to get insights and make some prediction on it. The problem of nationality identification from names has a variety of important applications, including biomedical research, demographic studies, and marketing [2]. A group of researcher did a study where they proposed, name classification models can be also used for classifying users in social networks [3]. Another group of researchers proposed that instead of utilizing user profiles, used clusters of users' first names, last names, and locations to identify users [4]. There was a research related to our work where they use 74M labeled names from 118 major countries and these 118 countries covered up 90% of the world's total population (Nationality Classification Using Name Embeddings). Later on they published a web app called t NamePrism based on their work (Name Prism). Another group of researcher proposed a novel name-Nationality classifier based on the multinomial logistic regression. Their name-Nationality classifier is trained and evaluated on Wikipedia data, achieving around 85% accuracy (Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching) .

1.4 Necessity

As the world is so much concern about security issues, nationality prediction can be a key component in a sense because we have seen in different contexts, crimes is interrelated to nationality. We have seen several news article titled like, "This was planned by a group of African team" or something like this. So this can be deployed in various systems at security checkpoints where they have the list of names and also have the system to classify their nationality. Just like the above mentioned scenario, nationality prediction may unlocks important demographic information about any individuals, which will help in many applications in biomedical and sociological research. Now a days, a big part of business is targeted add. Businessman's always tries to do targeted add and that makes the different for sure. Based on different nationality they can make also targeted add and that's why it's also necessary to know consumers nationality to make this happen.

2 Methodology

In our dataset, we got initially 62 nationality types. We noticed that after applying Naïve Bayes model is not giving us good result for several reasons and three of those are, very small dataset, dataset contains non English terms and have less than 80 names in some particular nationality type. Later we work on our dataset and finally make it to 25 nationality type with no non English terms and having more than 80 names in a particular nationality type. Then we used Neural Network (NN) model because of its effectiveness and interpretability. NN has a huge reputation for predicting sequential data too good. NN also has reputation for language modeling. In our dataset, we also have sequential data and also natural language data. There was another reason to use NN and that was the effectiveness of it's in small dataset also.

3 Implementation

After importing the data set, first task was to preprocess the dataset. Here, non-English terms were removed from the data set then we discard those nationality types which has less than 80 names. Then we did the feature selection part where we actually make all the features to lowercase letter and make a list of labels. After that we split the dataset into train and test set. Then we apply create tokenizer to the features and made a vocabulary. Then we applied one hot vector which is a 1 N matrix (vector) used to distinguish each word in a vocabulary from every other word in the vocabulary. Having these steps done, the data set was then prepared for fitting to the required models. Initially Naive Bayes and Logistic Regression were used. They showed accuracy less than 10 percent. To overcome this issue, Neural Network is used with 100 epochs.

3.1 Data Collection

We initially got a dataset from a github source (github) and then we add more data to the dataset from internet sources by searching as names related to 'X' nationality.

3.2 Data Processing

The data came somewhat preprocessed as there was no missing value but some adjustments had to be made to make the data compatible with the classifiers. Removing non-English words, removing nationality types containing less than 80 names, making all the features to lowercase letter, convert all the labels to a list, tokenizing and making a vector with one hot vector.

3.3 Model Development

To accomplish the goal, so far, we have used a machine learning model called naïve bayes from sklearn library. We have also used MultinomialNB from the same library. We used Logistic Regression also. From the dataset, 70 percent of the data is used to train the model, and the rest 30 percent is for testing the model. For better accuracy, later we

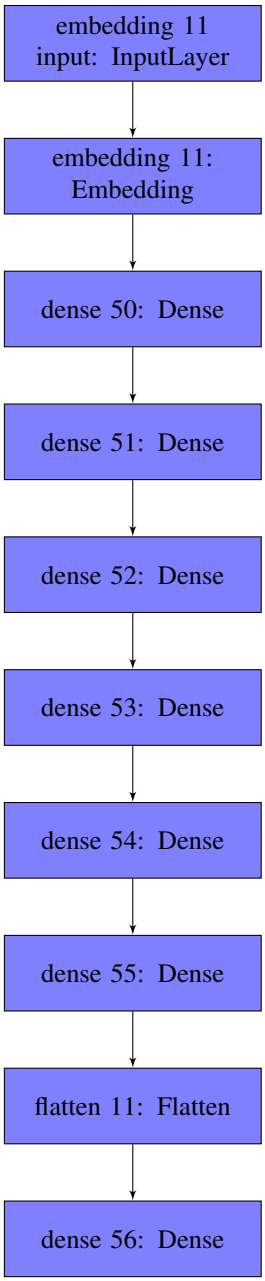
used Word2Vec, Cosine Similarity and Neural Network model. Finally got some results from Neural Network model. Python is used for coding.

We can see our Neural Network model's summary in the following:

Model: "sequential 11"

Layer (type)	Output Shape	Param
embedding 11 (Embedding)	(None, 12, 8)	216
dense 50 (Dense)	(None, 12, 16)	144
dense 51 (Dense)	(None, 12, 24)	408
dense 52 (Dense)	(None, 12, 32)	800
dense 53 (Dense)	(None, 12, 24)	792
dense 54 (Dense)	(None, 12, 12)	300
dense 55 (Dense)	(None, 12, 8)	104
flatten 11 (Flatten)	(None, 96)	0
dense 56 (Dense)	(None, 25)	2425
Total params: 5,189		
Trainable params: 5,189		
Non-trainable params: 0		
None		

We can see our Neural Network model's flow chart here:



3.4 Results

Initially we got very low accuracy of 7% using Naïve Bayes model and made some prediction. The result was following:

Names	Predicted Class	Actual Class
Kofi	african-american	african
Zaide	Filipino	yiddish
Virsila	Filipino	czechoslovakian

Later we used Neural Network on our very small dataset containing only 25 nationality types and got the highest

accuracy of 18% in the test set and 31% in the train set. We are hoping to get higher accuracy from our model if we can make our dataset bigger than that.

4 Conclusion

Predicting Nationality through names is not an easy task where there we have only one feature to train. For doing this study, we must need a bigger dataset to get a good result. Although we have a very small dataset, we must say that we got a pretty good accuracy through our model.

4.1 Challenges

It was somewhat hard for us at some point because the model we built using Naive Bayes did not show good result. Later on, we had to try few other things such as Logistic Regression, Word2Vec, NLP and Neural Network. It was totally a new things to try for us. We had to spend a lot of time to learn a simple things also.

4.2 Limitations

Because of lower dataset, our model can't predict that much good. So we have to wait for building a bigger dataset to work this model properly.

4.3 Future Prospect

A Recurrent Neural Network can be used for further research if anyone have a bigger dataset and also, a hierarchical recurrent neural networks can be applied to extract higher and more complex representation of personal names.

References

[1] J. Ye, S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, and S. Skiena, "Nationality classification using name embeddings," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1897–1906, 2017.

[2] A. Ambekar, C. Ward, J. Mohammed, S. Male, and S. Skiena, "Name-ethnicity classification from open sources," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 49–58, 2009.

[3] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow, "epluribus: Ethnicity on social networks.," *ICWSM*, vol. 10, pp. 18–25, 2010.

[4] S. Bergsma, M. Dredze, B. Van Durme, T. Wilson, and D. Yarowsky, "Broadly improving user classification via communication-based name and location clustering on twitter," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1010–1019, 2013.