Name Analysis

Anonymous

ABSTRACT

Name Analysis is the task of analyzing a given set of names. The first names and last names are considered independently to focus on determine the country of origin. The country of origin is being predicted independently to evaluate the predictions.

1. INTRODUCTION

In the modern world, immigration of people belonging to one country to another country is extremely common. This affects the demographics of the population of the country where the relocation occurs. One of the key areas affected is the distribution of ethnicities of the population. Ethnicity refers to the percentage of population belonging to a particular country or type of countries. Example of ethnicities included are Hispanic, Asian etc. Hispanic refers to people belonging to Spanish speaking countries or maybe people belonging to Latin America. Asian refers to all the people belonging to countries of south Asia, southeast Asia and east Asia collectively. One of the distinct features of people belonging to different countries is the different languages that are spoken in each of these countries. This change is fairly discernible when we look at individuals and their corresponding names. Naming is a practice that is associated with communication and is not limited to using with just human beings but also tangible objects, places, file names etc.

Naming conventions belonging to human beings are influenced by the language, origin and culture. Analyzing the names of individuals belonging to a subset of the countries is the primary motive of the project. The rate of immigration is usually affected by economic, political and social factors. Individuals are motivated by a combination of these factors. One of the countries of the world which has seen a great amount of immigration since the start of the twentieth century is the United States of America. There was a significant increase during the World War-I era that led to people belonging to from the European countries like Poland leaving their home countries and immigrating to the United States seeking refuge. There have been a great number of immigrants from Ireland to the United States stretching from the pre-American revolution era to the modern era.

The countries of the world considered for the task of determining the origin in this project ar Netherlands, Poland and Ireland. The genders of the individuals considered include a mixture of both males and females. This task can be used to study the health of the population residing in a country and can be expanded to applications that work on the demographics of an individual.

This capstone project report is organized as follows. Section 2 describes the background of work done by others around the project and lays down the context of their work in this project. Section 3 describes the data mining algorithm that was used for this project. Section 4 describes the steps taken for the implementation of the project. Section 5 discusses the results obtained and the analysis done on these results. Section 6 discusses the conclusions drawn from the implementation and results and Section 7 presents the future extensions and directions that can be taken to solve the problem.

2. BACKGROUND

Pachet et al. [6]analyze the syntax and semantics of file names and create heuristics on the basis of the structural analysis. The growth of internet has led to the increase in the digital presence of the entertainment industry. One of the key components of the entertainment industry is music. The days of music being stored in physical vinyl records have been replaced with discs to digital music files. There is no particular standard for identification of the source that is followed throughout the music files despite ISO and other organizations developing their identification standards. The differences in standards arise particularly when there is a need to access local user systems and there is a lack of a global database consisting all the available music files. The approach suggested tackles this issue by focusing on the context of popular file names and running them against the local files present. The extraction of artist name and title is considered to be an important piece of information from popular file names. Data sources considered are local music files, audio files obtained from radio broadcasts and metadata pertaining to these files.

A key factor when it comes to local files on individuals is the simplicity in the syntax for the nomenclature for these files. Usually the nomenclature is devised in an easy to understand manner which includes the artist information and

the title of the song. Another differentiating factor between local files of different users is the directory structure used for a popular file. This would suggest there would be several abnormalities between sets of music files and there is an absence of a common template for representing a file name. This thing is acknowledged and there is usually similarities found when the files are handled as clusters at the hierarchical level. One of the most noticeable attributes of music file names is the segregation by using a common delimiter. The next attribute considered is the structure of the file name and there are a limited number of possibilities in the arrangements of the pair, Artist and Title Name. One of the naive attributes to consider is the presence of common terms within a set or subset of files. Examples of these terms include the names of the albums, album launch or file saving date and repetition of key phrases like feat when there is a collaboration between several artists. In handling files in groups, it is often seen that there is a repetitive usage of a common syntax within the groups. The authors consider these set of attributes and develop a hypothesis for each of them for the data considered by them.

The hypothesis developed is tested against the files present and evaluated with statistical results developed. The evaluation phase is essential in developing a system that is able to work on the basis of a correct hypothesis. The built system here incorporates several of the attributes tested before by implementing them like grouping the file names with similar syntax or by considering any of the special cases that arise like the different delimiters that were present leading to different syntaxes amongst all the file names. Once the file names are segregated down into groups or sub-groups, a set of heuristics that include looking at popular artists are applied for the identification of the artist names, album names. The authors design a system that is able to navigate through file names using heuristics without having the access to the syntax of the music file.

Mateos et al. [5] divide the UK population on the basis of forenames and surnames into common origins by evaluating several techniques. The classification is done on the basis of a concept known as CEL. It is a set of characters constructed as a pattern on the basis of the language, ethnicity and culture representing a particular language, culture or language or a combination of these resulting in very specific classifications to very wide classifications. It is a list that is dynamic, the classification categories can be altered according to the needs of a particular task. These patterns represent several ranges to classify an individual as they include the geography and religion to go along with the ethnicity and language. This concept can be used to represent a mapping between the names present in a population and the characteristics associated with it. The authors compiled a mapping of the population of the UK according to the defined CELs. This enables them to classify target populations. There is a reference population that is divided according to the CELs and the role of the target population is to get the ethnicity assigned according to the reference. This assigned ethnicity is evaluated by comparing against the original ethnicity. This classification task was approached in several different ways. A trivial way was to classify the forenames on the frequency of surnames present and assign the forename with the ethnicity of the highest featuring surname. The next couple of approaches were to consider the distribution geographically. It involved classifying a forename on the basis of frequency of a surname in a particular geographic region and then looking at a geographic region in terms of concentration of ethnic groups on the basis of the common financial status associated. The techniques of text mining are applied by focusing on the morphology of the names. This is done by focusing on the endings and stems of a name along with the presence and absence of sequences of letters in a name. There was a traditional approach also considered where the hard bound books were utilized for developing a name to ethnicity ratio. Similarly, the last two approaches are focused on researching by looking for them through data available through the rest of the sources available to everyone via the world wide web. The first is based on the reference of all the previous techniques, where there was a frequency count computation, it is repeated on the global scale where the frequency of names of an ethnicity is looked across all the nations of the world. The last approach is an extremely brute force approach, which was considered if the other approaches were deemed not sufficient, it involved looking at web resources like google for each individual name. The authors use the above approaches to build a classification model which works appropriately for the data sources involving a countryâĂŹs census.

Kridharan et al. [3] determine the ethnicity from names by dividing the names into substrings of length 3 and focus on the occurrence of these substrings throughout the set of names by using the Naive Bayes algorithm. The authors suggest using substrings to overcome the difficulties of parsing an entire name and this plays a significant role if the dataset consists a huge number of names. Using substrings lead to possibilities of capturing common substrings between ethnicities and can be used to compute the ethnicity of a new name.

The authors give an example to show how capturing the similarity between three distinct names is possible with substrings. Although the names were distinct, there were many shared substrings between them. All three names belonged to one country, the authors chose the algorithm, Naive Bayes, since there was a good ratio between the matching substrings and the entire names. Naive Bayes is a probabilistic algorithm. The probabilities are calculated for mutually exclusive events or independent events by using Bayes′ theorem. Bayes′ theorem is a conditional probability theorem, if there are two events, A and B, it is possible to calculate the probability of event A occurring given event B has occurred. This is applied with the substrings and ethnicity problem and the probability of an individual being of an ethnicity is calculated. The authors implemented this approach on the data consisting of the following countries: Malaysia, India and China. The dataset was split into training data and test data. They built a grid for the training data to evaluate the performance with the test data.

The approach taken for analyzing the names involves parsing the names and segregating them into first names and last names. Pachet et al. [6] suggest to analyze the syntax and morphology of the file names. Mateo et al. [5] also had techniques that focused on the morphology of forenames and surnames. Lastly, Kridharan et al. [3] handle the structure by breaking it into substrings. This approach is considered as there are sequence of characters that are common in names from a specific country of origin. Each name is considered by obtaining the set of trigrams associated with it and determining the country of origin by using the clas-

sification algorithm: k-nearest neighbors. Kridharan et al. use the Naive Bayes algorithm to predict but instead the classification task can also be computed using an algorithm that is able to predict by calculating the similarity between strings and not just their occurrence.

3. ALGORITHM

The data mining algorithm chosen for the task of predicting the country of origin is k-Nearest Neighbors [1]. This algorithm is a classification task that enables to choose the label based on a decision made among the several chosen attributes and their corresponding labels. Classification is known as a supervised learning task, the class of the data instance is obtained on the basis of the known class labels. This process is carried out by dividing the datasets into two sets, known as the training dataset and known as the test dataset. The split ratio for the division is usually 67:33 or 70:30 but this ratio can be altered according to the requirements of a specific problem statement. Training dataset is used to create the model and the test dataset is run on the model to obtain a measure of the correctness. There are several algorithms that are categorized under classification tasks including Naive Bayes, decision trees like J48, Random Forest etc [2]. All these techniques differ in how they handle the given set of data, the attributes and there is an effect on the evaluation criteria, unlike an unsupervised learning task. In an unsupervised learning technique, there are no prior labels known so the technique is used to find hidden patterns of the data.
k-Nearest Neighbors, as stated before, is a classification task. The division of training data and test data is similar to other techniques but unlike other techniques, there is not any prior modeling in the training phase. Instead, the learning is done once the test instance arrives, the test instance is evaluated with all the training samples. This step is iteratively repeated for all the test samples against the test instances.

- Split dataset into training and test dataset
- Compute the distance between test instance with a training instance
- Search for the k-most nearest neighbors from the test instance
- Predict the class label from the nearest neighbors
- Compute the accuracy for the prediction

Distance between two points is usually computed in two dimensional space by using the Euclidean distance. Euclidean distance gives us the result as a metric to measure the straight line that will be formed between those two points. It is a very popular measure and it is used very commonly in mathematical computations.

$$(a, b) = sqrt((a1 - b1)^2 + (a2 - b2)^2) \quad (1)$$

Equation (1) is used to calculate the euclidean distance between two points a(a1,a2) and b(b1,b2) and sqrt represents the square root of a number.

Euclidean distance is an appropriate measure to use when the data is present in the coordinate space. It is difficult to use this measure when in context of a computer science problem. Depending on the type of data available, there are other distances like Manhattan, Mahalanobis available but none of them present an adequate measure to compare the similarity between two strings.
Distance measure can also be called similarity measure in the context of handling two strings. One of the most prominent similarity measures for such a scenario is called the edit distance. Edit distance is a way to measure the dissimilarity between two strings and is found by computing the smallest number of operations performed on a string to convert it the other. One of the simplest variations to the edit distance is keeping track of the edits made to a single character and computing the minimum number of steps to convert it from one string to the other string. The type of edits that can be performed can be considered as primitive operations. These operations are adding a character at a specific position, remove a string from a specific position and replacing a character with another character at a specific position. Dice′s coefficient [4] is also a measure to compare two strings and find the degree of similarity between them. It is a measure of the similarity between the set of bigrams that are generated. It represents a ratio between the set of bigrams that are common and the complete number of bigrams that are generated by the two strings i.e. the sum of all the bigrams that are generated taking both the strings.

Dice's coefficient,

$$d = (2 * common\ bigrams)/(total\ bigrams) \quad (2)$$

where, 0.0<=d<=1.0
Equation (2) is used to calculate the Dice′s coefficient for a set of bigrams and it can be extended in general for a set of n-grams for calculating the similarity measure between two strings. The range of the coefficient is between 0.0 to 1.0, 0.0 is the value expected between strings not having any common n-grams and 1.0 is the value expected when they form the same n-grams. In the approach followed, we have considered trigrams. Once the distance or similarity measure has been decided, it is used for computing the list of nearest neighbors. Nearest neighbors are essentially instances from the training set when compared to the test instance that have a high similarity value or very less distance value, remaining instances of the training set are no longer considered. The number of training instances chosen as the nearest neighbors is determined by k. k can be a value that can be changed according to the desired results in terms of performance or it can be tested and found. Such an approach would lead the optimum k.
The next step after finding the list of nearest neighbors is finding an agreement between all the neighbors. The chosen neighbors all have a value for the class label and they all are suitable candidates to pass their label as the classifier value. To overcome this, there is a method called majority voting. This voting mechanism is carried out by letting all the neighbors vote for their class value and then the value with the highest count among all neighbors is considered as the majority vote and assigned as the class label for the unassigned object. If there are even number of neighbors, there

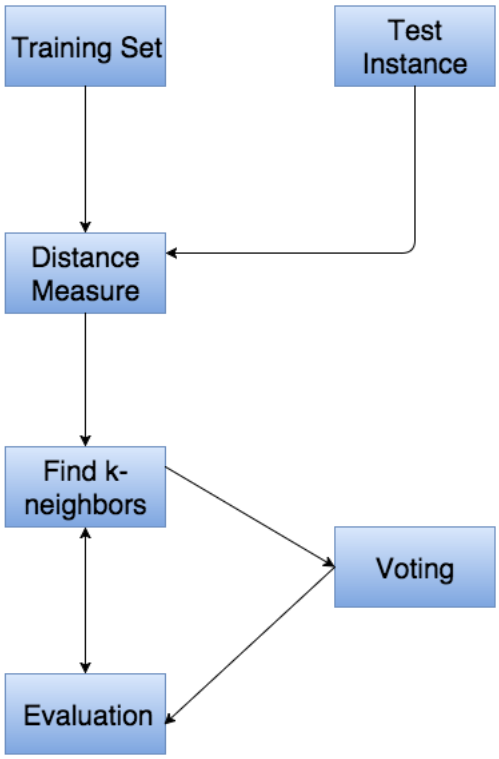

Figure 1: kNN Steps

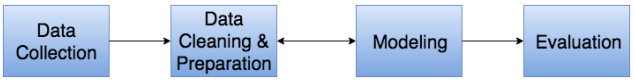

Figure 2: CRISP-DM Process

is a chance of a tie among the voting. There are alternative techniques like the weighted voting, which assigns weights to neighbors according to their proximity and as such can be used in these cases.

The evaluation can be done by computing the accuracy of the classifier. Accuracy can be calculated by comparing the assigned class label with the known class label. This is a supervised learning task and as such the test instance has a class label. The accuracy is computed by taking the ratio of the correctly classified instances to the total number of correctly classified instances.

4. IMPLEMENTATION

The prediction of country of origin was a data mining task and the implementation was based on the CRISP-DM methodology (see Figure 2). This methodology enabled dividing the process into four main components.

4.1 Data Collection

This is the first phase of the implementation which involved a great deal of searching and collecting data from several data sources. The data source chosen in the end was from a website, FamilySearch.org. It is an open source repository for collecting data. The data needed for this project had some custom-defined constraints. The chief constraint was the availability of person name records inclusive of first name and last name of people from the years of 1900-1920. The range of years chosen is impacted by the second constraint enforced, picking a period in history after which there was a steep increase in the rate of immigration. The countries of origin were the Netherlands, Poland and Ireland. These countries were chosen on the basis of sociological factors like the increase in immigration during that period. The data collected included the following attributes: Name, Gender, Date of Birth , Address, MotherâĂŹs Name, FatherâĂŹs Name. The data was available in separate files with each file having 75 records and .xls format. This process involved a lot of manual work and there was no way to automate it as the data was not released by the website as a collection.

4.2 Data Cleaning and Preparation

This phase had the final objective of merging the separate data files into a single data source file that can be fed to the model. The files were available according to the country of origin. First step involved merging all those files according to the country of origin separately. The datasets were cleaned separately, this involved handling missing values, empty values. Missing values are handled either by deletion of rows, if there are very few or by using domain knowledge. Replacing them with statistical values was not approached as there were no such attributes relevant to the model. The next step involved removing the irrelevant attributes from each file and reducing it to the following attributes: Name, Gender, Date of Birth, Address, Mother′s Name, Father′s Name. Importantly, the attributes Mother′s Name and Father′s Name provided were now removed and added as rows to the merged files. The merged country files had the following attributes: Name, Gender, Country, the attribute name was inclusive of first names and last names along with the middle names. All three country files were merged into one file which was used as the dataset for the model. The number of instances from each country before staging them into one file was as follows: Netherlands-8486, Poland-3939, and Ireland-3086. The difference in the number of records for each country is generated due to adding the attributes MotherâĂŹs Name and FatherâĂŹs Name as rows and the lack of presence in all files. The merged dataset had 15212 instances in total.

4.3 Modeling

This phase involved building the model by using the k-Nearest Neighbors algorithm for trigrams and implementation of the model was done in the programming, language. The first step involved in building the model is the division of the dataset into a training dataset and a test dataset. The rows were split randomly in the ratio of 80:20. This ratio was chosen since the total number of instances was relatively low.

Before the division of the dataset itself, there was the need to extract first names and last names separately while eliminating the middle names. This task was based on domain knowledge as there was a possibility of a discrepancy of parsing through a name. Heuristics was constructed that segregated first names, last names by considering the first word

to be the first name and last word to be the last name. The middle words were not considered a part of either. Ignoring all middle words also ignored common last names that included the substrings like van, de which are followed by a space. They were included in the heuristics as special cases that were considered a part of the last name. The dataset was divided into two files each for the first name and last name. Test instances for first name were 3096 whereas test instances for last name were 3101. Training instances for first name were 12340 whereas training Instances for last name were 12404. The model was run separately for first name and last name. The model followed the approach mentioned in the steps below and Figure 3 explains one pass of the approach.

- Each test name is run against the set of names in the training dataset.
- The test name is broken into a set of trigrams and the training instances are all broken into trigrams.
- Test name is compared with each instance of the training dataset by calculating the DiceâĂŹs coefficient between the two names using the set of trigrams.
- The desired number of neighbors, k, are chosen from the computed pairs of names.
- To decide the country from the neighbors, we take the majority vote of the countries.
- Predict the class label from the received vote.

When majority voting is used to decide the class label, there is a chance that the value of k considered is even. In such a scenario, there is a possibility that there is no majority and there is a tie and there is no decision for the country. This is handled by randomly assigning a label which is a part of the tie to the unassigned instance. k can be altered and any value of k will be accepted from the input. The class labels for the decision are the same as the class labels present in the data mining model.

4.4 Evaluation

Evaluation is done by measuring the accuracy of the classification. Accuracy is calculated by taking the ratio of correctly classified instances to the total number of classified instances, comparing the predicted country with the existing country. It is calculated for each value of k after all the test instances have been run against all instances in the training set considered for that iteration. It is important to compute the accuracy for different values of k to determine the optimum value of k. The optimum value of k for a sample is the k for which the accuracy of classification is the highest among all.
Another factor of evaluation is the running time of the model. It is important as k-Nearest Neighbors algorithm is said to be a lazy algorithm. It means there is no time consumed in the training period as the algorithm waits till the test instance to arrive. There is a lot of computation overhead added by splitting the names into trigrams. Running time can be calculated for each value of k for all data samples.

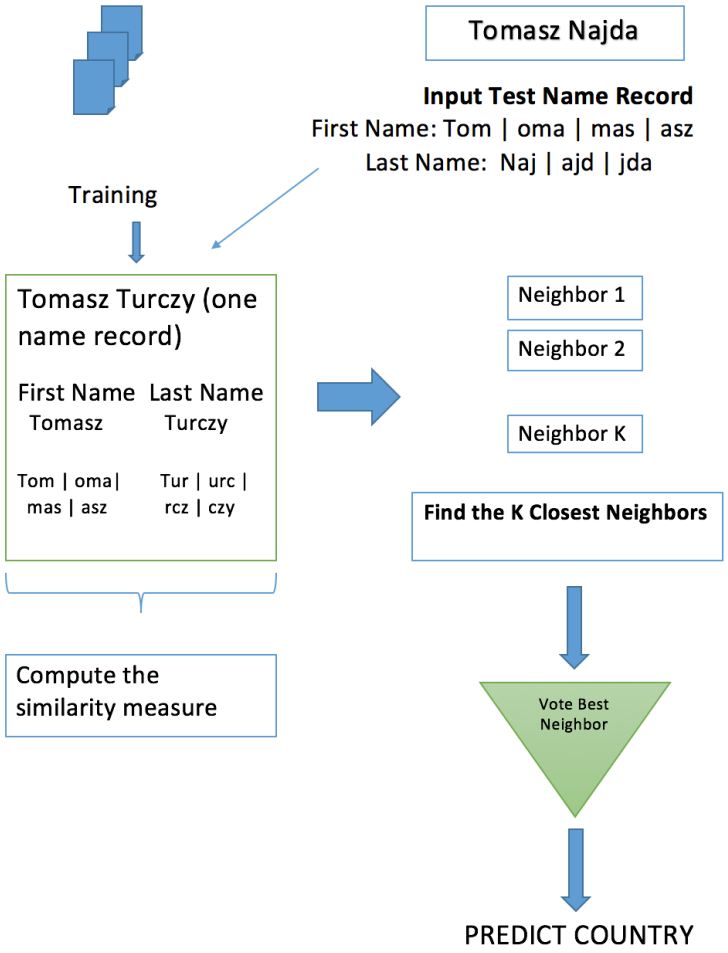

Figure 3: kNN with Trigrams

Lastname	Gender	Country
Wnużka	M	Poland
Adair	U	Ireland
Otter	F	Netherlands
Łatwińsk	F	Poland
Skrzypcz	M	Poland
Skrrekut	M	Poland
Strrmanr	M	Netherlands
Gertrudis	F	Poland
M'Cabe	U	Ireland
Akkerma	M	Netherlands
Csimiega	M	Poland
Rook	M	Netherlands
M'Adam	U	Ireland
Braber	F	Netherlands
Fase	F	Netherlands
Solis	M	Poland
DeLang	F	Netherlands
Adams	U	Ireland
Hollande	F	Netherlands
Sztrepka	F	Poland
Sars	M	Netherlands
Evertsen	F	Netherlands
Skorniew	M	Poland
Schouter	M	Netherlands
Skowron	F	Poland
Rovers	M	Netherlands
M'Carron	U	Ireland

Figure 4: Sample Test Last Name

This evaluation helps in understanding the impact of the data size on the model, the impact of the value of k on the model. If the running time is high, it affects the decision to choose k and there is a possibility that the optimum value of k is not discovered.

5. RESULTS

The division into two test datasets for first name and last names enables running two models. The number of neighbors is changed from k=1 to k=5 for a test first name and a test last name. Output of the results is compiled by writing the predicted country and actual country. This is a supervised learning task due to which the value of the actual country is accessible. Once the list of neighbors is obtained and there is a vote among the neighbors, the decided vote is written as the predicted country.

Figure 4 and 5 show the sample of the files, Test Last Name and Test First Name. The results are compiled in a .csv file with a newline delimiter. The attributes of this file are input, result and expected. Input represents the test name, result represents the predicted country of origin and expected represents the actual country of origin. Figure 6 represents a subset of the final results.

Evaluation of the prediction model is done by computing the accuracy. Accuracy is the ratio of the value of result to the value of expected. Accuracy of this process is performed to pick the appropriate value of k.The number of instances in the training dataset and test dataset results in a large computation time for prediction.The computing time involves finding the similarity between two set of trigrams each time. This process is repeated for both test datasets independently.
The model is built by running samples of the test and training dataset. This is taking into account the computation time associated with the string similarity and the reason

Firstnam	Gender	Country
Michał	M	Poland
William	U	Ireland
Klara	F	Netherlands
Jadwiga	F	Poland
Stanisław	M	Poland
Vincentiu	M	Poland
Matthijs	M	Netherlands
Gertrudis	F	Poland
Laura	U	Ireland
Klaas	M	Netherlands
Michael	M	Poland
Arie	M	Netherlands
Mary	U	Ireland
Willempj	F	Netherlands
Johanna	F	Netherlands
Stanisław	M	Poland
Elizabeth	F	Netherlands
Jane	U	Ireland
Lijntje	F	Netherlands
Carolina	F	Poland
Jacobus	M	Netherlands
Sophia	F	Netherlands
Bolesław	M	Poland
Cornelis	M	Netherlands
Josepha	F	Poland
Hendriku	M	Netherlands
John	U	Ireland

Figure 5: Sample Test First Name

input	result	expected
Michał	Ireland	Poland
William	Ireland	Ireland
Klara	Ireland	Netherlands
Jadwiga	Poland	Poland
Stanisław	Poland	Poland
Vincentiu	Ireland	Poland
Matthijs	Poland	Netherlands
Gertrudis	Poland	Poland
Laura	Poland	Ireland
Klaas	Poland	Netherlands
Michael	Ireland	Poland
Arie	Poland	Netherlands
Mary	Ireland	Ireland
Willempj	Ireland	Netherlands
Johanna	Netherla	Netherlands
Stanisław	Poland	Poland
Elizabeth	Poland	Netherlands
Jane	Poland	Ireland
Lijntje	Poland	Netherlands
Carolina	Poland	Poland
Jacobus	Poland	Netherlands
Sophia	Poland	Netherlands
Bolesław	Ireland	Poland
Cornelis	Netherla	Netherlands
Josepha	Poland	Poland
Hendriku	Netherla	Netherlands
John	Netherla	Ireland

Figure 6: Sample Result

trainingsize	time	k	accuracy
10	0	1	22.222222222222
10	0	2	22.222222222222
10	0	3	22.222222222222
10	0	4	22.222222222222
10	0	5	22.222222222222
100	15	1	11.111111111111
100	14	2	16.161616161616
100	15	3	16.161616161616
100	17	4	15.151515151515
100	15	5	18.181818181818
100	24	5	24.242424242424
100	25	5	23.232323232323
200	380	1	16.080402010050
200	393	2	19.095477386934
200	369	3	17.085427135678
200	386	4	16.582914572864
200	385	5	17.085427135678

Figure 7: Sample Analysis File Built

that kNN does not have a prior trained model. The sample size of 10, 100, 200, 1000 from the test dataset is run against the training dataset. This is followed by running the entire test dataset against the sample training dataset sizes of 5000, 10000 and the entire training dataset. The model for the last names is built before the model for the first names. This choice was influenced by the length of the last name being greater than the first name. Figure 7 shows the file and the attributes written to compare the running times, accuracies, different values of k considered for each sample sizes.

Figure 8 considers the running time for the sample sizes of 10, 100 and 200. The running time is greatly affected by the size of the dataset as it involves more sets of strings that need to be compared. Conversely, the choice of the value for k also determines the running time as more neighbors need to be chosen and there is a tradeoff when the final decision for the class variable is made. Figure 9 is plotted to compare the training size with the time taken and there is an increase in time relatively for each sample size. There is an increase in time with the increase in the value for k for almost all sample sizes that are taken into consideration. The samples are taken from the set of last names

Figure 10 plots the values of K=1,2,3,4,5 for the sample data sizes taken against the accuracy that is achieved for these values. The accuracy is low for the extremely small sample sizes of 10,100 and 200 at around 20% and it increases to an average of 33 % for the higher sample sizes. This is due to the lack of a fixed training phase when kNN is used; more

Figure 8: Running time vs Data for K Neighbors

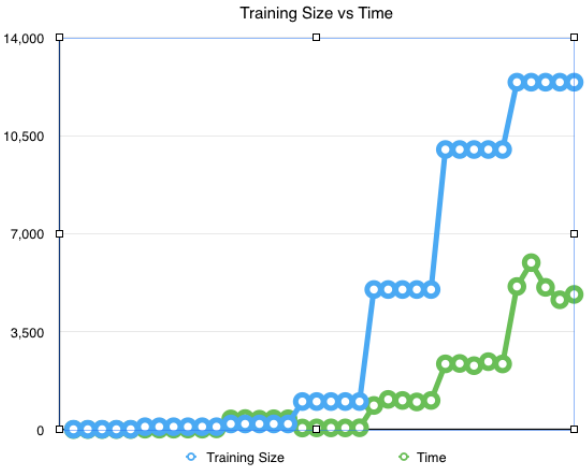

Figure 9: Training Size of Last Name against Time

number of samples in the training set enables better learning. Increase in accuracy leads to much better predictions for newer test samples. The accuracy rises moderately and does not spike suggesting that there is no bias introduced due to the training samples or there has been no overfitting so far.
Figures 11,12 are plots made for the model built when the test set of first names was run. The sample sizes chosen still remained the same. Choosing different sample sizes for the training datasets and test datasets enables analyzing the change in running time associated with the prediction task.

Comparison between figures 7 and 10 shows the running time was much quicker for the first names than the running time for the last names. One potential reason is the syntax of some of the last names leading to extremely long number of trigrams. Some examples of last names included spaces and the spaces can not be ignored as the position of spaces is relevant. First names usually are much more common and as such there are far more repeated trigrams found, which leads to less storage and quicker access. Although the run-

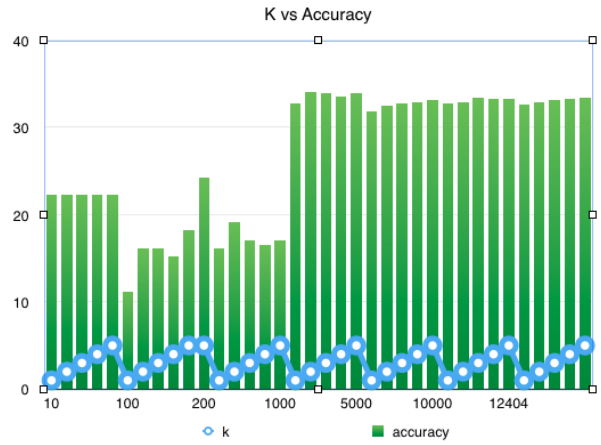

Figure 10: Values of K Last Name against accuracy achieved

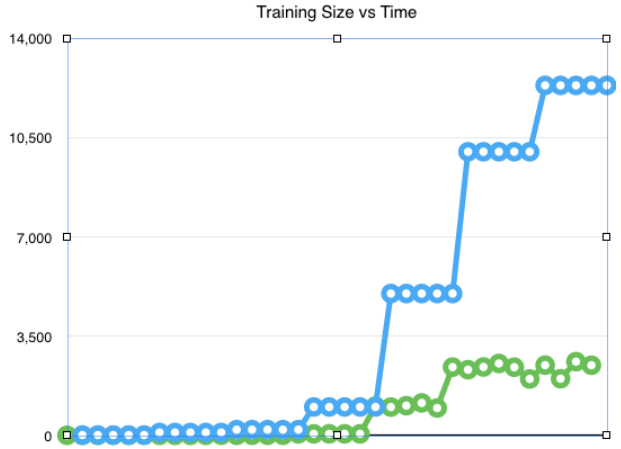

Figure 11: Training Size of First Name against Time

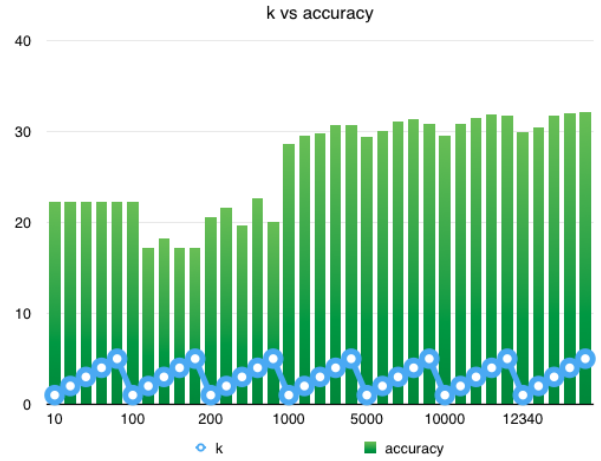

Figure 12: Values of K First Name against accuracy achieved

ning time has been affected but when there is a comparison between the accuracies achieved, there seems to be similarities between both the models. Accuracy for the smaller sizes of the sample first names dataset has remained in the range of 17-20% and the highest achieved is in the range of 30-33%. The average accuracy achieved for all values of k considered is around 33%. This similarity is due to the similarity in the number of instances in the dataset and the amount of learning that has been done prior to running against the latest sample training set size. The more the number of samples, the higher the number of trigrams that can be detected. One of the reasons for running different values of k, is to find the optimum values of k. This can be found by using the values achieved for accuracy and choosing the highest percentage of accuracy obtained for the entire training dataset. The corresponding value of k is the optimal k. The value of k for the models built for the first names and the last names with the highest accuracy is k=5 and the accuracies achieved are 32.05 and 33.45%. There is a small difference in accuracies achieved for both the models.

6. CONCLUSIONS

Name analysis was done on the names of people who were born between 1900 to 1920 from the countries of Netherlands, Poland and Ireland.Prediction of country of origin is possible by dividing the names into sets of trigrams and using the k-Nearest Neighbors classification algorithm. kNN is a classification task that can be altered to suit the properties of the problem and the dataset by using the appropriate distance or similarity measure. Comparison between strings is possible even after they are broken down into trigrams and it involved the trigrams being computed regularly. DiceâĂŹs coefficient is an adequate measure for calculating similarity between two names when we consider the trigrams of the test and training instances. kNN is able to predict the country of origin of a particular instance by using majority voting. Running time of the algorithm is impacted by the size of the data but it enhances the accuracy as more data is learnt. It is affected by the value of k as the higher values of k take more time to compute.

7. FUTURE WORK

The number of instances in the dataset is comparatively less but there is still overhead in the running time due to the comparisons between the set of trigrams. If the dataset is significantly increased, the running time will also significantly increased. In order for the computation time to be significantly decreased, the implementation can be altered by using the MapReduce framework and performing the comparisons on clusters. This will significantly reduce the running time and will enable computations for much larger values of k. The approach uses a lazy algorithm like k-Nearest Neighbors with trigrams without considering the positions at which the trigrams can be found. If names are broken into trigrams and are represented in a tabular form there is no fixed position defined for the trigram in the table and as such traditional classification algorithms like decision trees cannot be used directly. There could be changes made to the approach to improve the accuracy by using a better algorithm than kNN or the voting mechanism can be altered by using weighted voting. Additionally, the names of the persons present in the dataset, primarily first names can be used to determine the gender of a person.

8. REFERENCES

[1] K. Fukunaga and P. M. Narendra. A branch and bound algorithm for computing k-nearest neighbors. *Computers, IEEE Transactions on*, 100(7):750–753, 1975.

[2] D. Gupta, A. Malviya, and S. Singh. Performance analysis of classification tree learning algorithms. *IJCA) International Journal of Computer Applications*, 55(6), 2012.

[3] K. Komahan and D. D. Reidpath. A âĂIJroziahâĂİ by any other name: A simple bayesian method for determining ethnicity from names. *American journal of epidemiology*, page kwu129, 2014.

[4] G. Kondrak. N-gram similarity and distance. In *String processing and information retrieval*, pages 115–126. Springer, 2005.

[5] P. Mateos, R. Webber, and P. Longley. The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. 2007.

[6] F. Pachet and D. Laigre. A naturalist approach to music file name analysis. In *ISMIR*, 2001.