NAME ANALYSIS

PRANAY SHASHANK ADAVI (pa8546@rit.edu) | ADVISOR: PROF. CAROL J. ROMANOWSKI

ROCHESTER INSTITUTE OF TECHNOLOGY

PROBLEM STATEMENT

Name Analysis is the task of analyzing the given set of first names and last names and determining the country of origin.

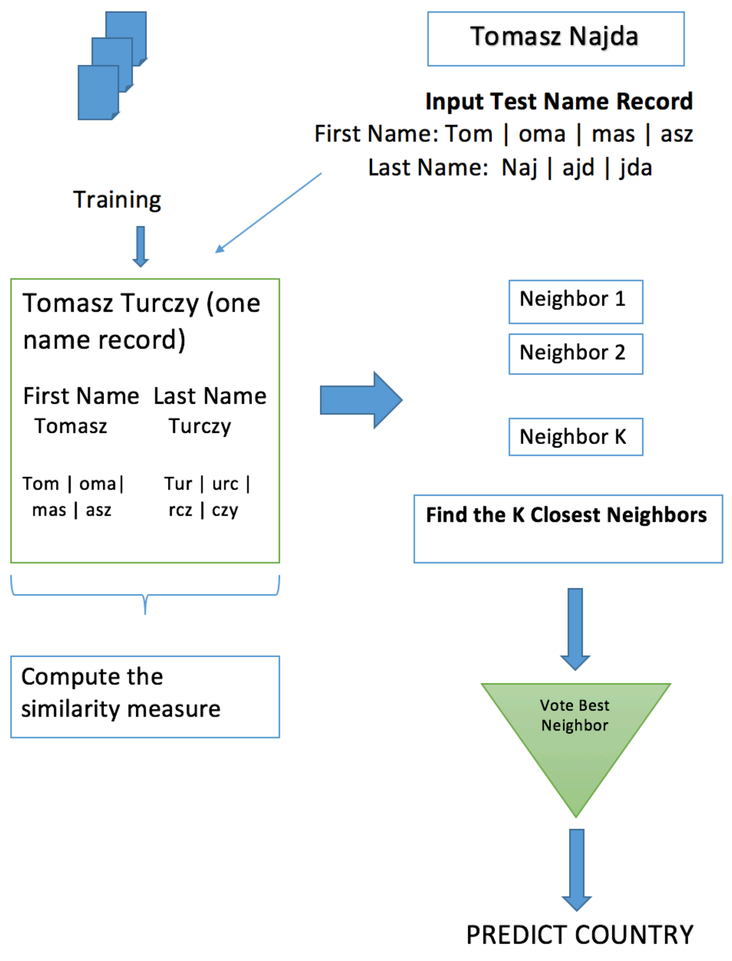
INTRODUCTION

- Names are given to any tangible object, place, person, music files etc.
- Names considered vary depending on things like the country of location, culture, origin language etc.
- Countries of origin considered: Netherlands, Ireland and Poland

BACKGROUND

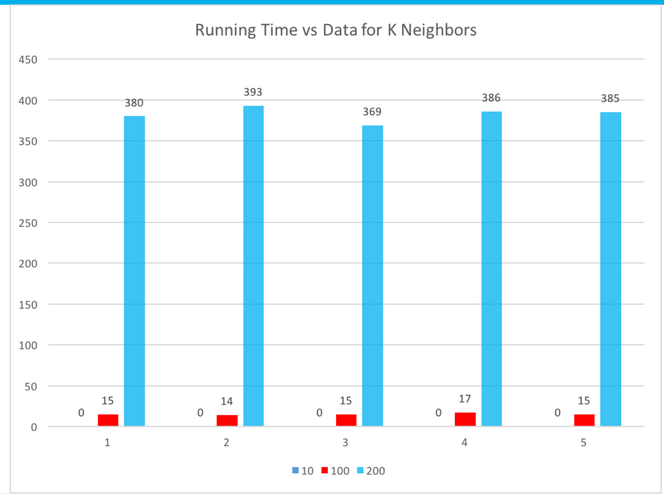
- Pachet et al analyze the syntax and semantics of file names and create heuristics on the basis of the structural analysis.
- Kridharan et al determine the ethnicity from names by dividing the names into substrings of length 3 and find the ethnicity using the naive Bayesian algorithm.
- The approach taken for analyzing the names involves parsing the names and segregating them into first names and last names.
- Each name is predicted by obtaining the set of trigrams associated with it and determining the country of origin by using the classification algorithm: k-nearest neighbors

APPROACH



- Data source: FamilySearch.org.
- Attributes included Name, Gender, Date of Birth , Address, Mother's Name, Father's Name.
- Data Cleaning and Preparation: Extracting First Name and Last Name, Gender, Country of each row.
- Modeling: Using k-nearest neighbors algorithm with trigrams.

RESULTS & OBSERVATIONS



- Accuracy & Running Time is directly proportional to data.
- Running Time is directly proportional to the value k.

CONCLUSION & FUTURE WORK

- Prediction of country of origin is possible by dividing the names into sets of tri-grams and using the k-nearest neighbors classification algorithm.
- Dice's coefficient is an adequate measure for calculating similarity between two names.
- Computation time can be significantly decreased by using the Map Reduce framework.

REFERENCES

- A Naturalist Approach to Music File Name Analysis. Pachet, F. Laigre, D. ISMIR 01, USA. (2001)
- A "Roziah" by Any Other Name: A Simple Bayesian Method for Determining Ethnicity From Names. Kridaraan Komahan and Daniel Reidpath. American Journal of Epidemiology (2014)