## Name – mother tongue mapping

**Dataset description:**

- Dataset Characteristics = Categorical
- Total Number of Instances = 31608
- Number of attributes = 3
- Attributes :-
  - Name
  - Surname
  - Community
- No missing values
- Total 27 mother-tongues/ communities

**Data preprocessing:**

- The text data 'Name' and 'Surname' are vectorized as follows:
  - Each instance vectorized with 54 features of which 27 features are for 'Name' and 27 for 'Surname'
  - The 27 features correspond to each of the 27 communities.
  - Each of the 54 features is the conditional probability of a name/surname being in the community corresponding to that feature.
- Communities are label encoded 0 to 26.

**Model training and fitting:**

- 3 layer DNN Classifier used
- Number of nodes in 3 layers are 100, 200, 100 respectively
- Rectifier activation function used.
- Number of output classes = 27
- Batch size = 20
- Epochs = 2000

**Results:**

- Accuracy obtained ≈ 93%