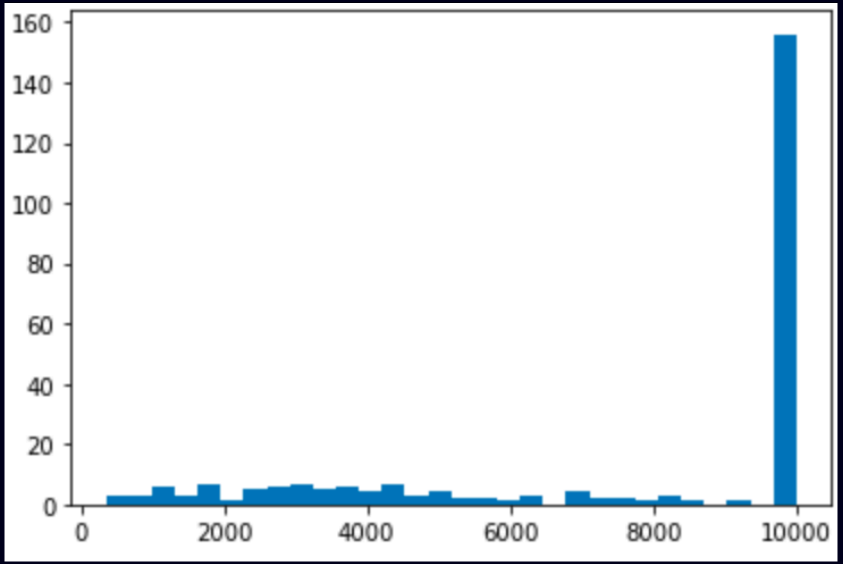
Automatic Language Identification and Relatedness Mapping

Sonia Cromp

Dataset

- 248 languages’ Wikipedia article texts
- Cleaning: Remove punctuation, non-text characters
- Result: Random 500-character “chunks”



Chunks per Language

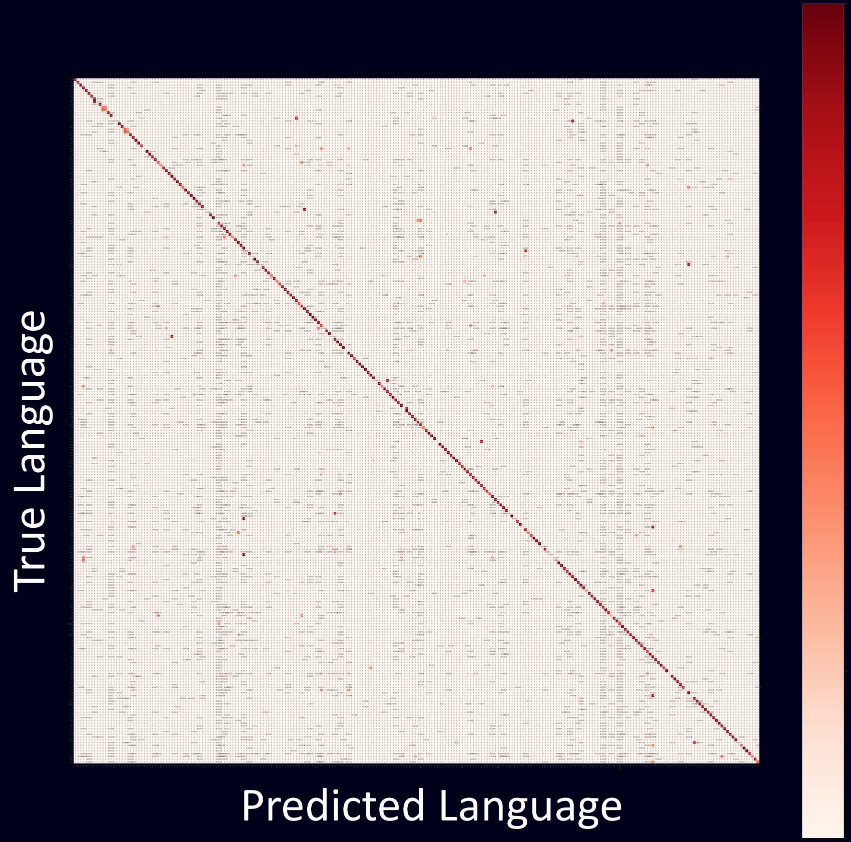
Acknowledgement: Compute resources from University of Pittsburgh

Center for Research Computing (CRC) offered through LING 1340 coursework

Classification, Clustering

Language Identification

- Naïve Bayes on character bigrams
- 90.6% accuracy



Mapping Relatedness

- *k*-means clustering, *k=150*

Future Work

- Anonymize writing systems
- Fine-tune Naïve Bayes, implement Convolutional Neural Net

Clusters

- Swahili, Tsonga, Xhosa, Chewa, Kinyarwanda (Bantu)
- French, Norman, Picard (Oïl)
- Persian, Gilaki, Mazanderani (Western Iranian)
- Central Bikol, Cebuano, Javanese, Pangasinan, Tagalog (Malayo-Polynesian), Tok Pisin (English creole)
- Alemannic, Ripuarian, Pennsylvania German, Palatine German (High German)
- Romanian (Balkan Romance), Silesian (West Slavic), Kotava, Lojban (constructed)