Ethnic Fractionalization Using Names

1 Introduction

2 Data

3 Methods

Features

- Split names into n-grams using characters
- Test using all lowercase letters and marking which letters begin and end a name (relevant for naive bayes and SVM, which are order agnostic). Idea here is to basically give algorithm information as to whether n-gram of characters starts a name, ends a name, or is in the middle of a name.

Algorithms

- Naive Bayes
- SVM
- RNN

4 Results

Table 1: Religion Results

Accuracy		Parameters			
NB	SVM	Case	ngrams	trim min	trim max
0.9438	0.9462	lower	2	0.001	0.9
0.9438	0.9438	lower	3	0.001	0.9
0.91	0.9224	lower	1	0.001	0.9
0.8924	0.91	lower	1	0.02	0.9
0.8224	0.8876	lower	3	0.02	0.9
0.8435	0.8815	lower	2	0.02	0.9

Table 2: D3 Results

Accuracy		Avg Diff Herf		Avg % Diff Herf		Parameters			
NB	SVM	NB	SVM	NB	SVM	Case	ngrams	trim min	trim max
0.6785	0.9885	0.0021	1e-04	-0.9779	-0.98	lower	3	0.001	0.9
0.6682	0.9879	0.0024	1e-04	-0.9775	-0.98	lower	1	0.001	0.9
0.1209	0.9879	0.0054	1e-04	-0.9745	-0.98	lower	2	0.02	0.9
0.1424	0.9879	0.0049	1e-04	-0.975	-0.98	lower	3	0.02	0.9
0.3812	0.9876	0.0047	1e-04	-0.9752	-0.98	lower	1	0.02	0.9
0.6879	0.9876	0.0021	1e-04	-0.9778	-0.98	lower	2	0.001	0.9

Table 3: D10 Results

Accuracy		Avg Diff Herf		Avg % Diff Herf		Parameters			
NB	SVM	NB	SVM	NB	SVM	Case	ngrams	trim min	trim max
0.2338	0.4815	0.0039	0.0015	-0.9725	-0.9778	lower	2	0.001	0.9
0.24	0.4797	0.0038	0.001	-0.9726	-0.9772	lower	3	0.001	0.9
0.1388	0.4382	0.0035	0.0021	-0.9728	-0.9784	lower	1	0.001	0.9
0.0897	0.4259	0.003	0.0025	-0.9734	-0.9789	lower	1	0.02	0.9
0.0674	0.4171	0.0082	0.0029	-0.9682	-0.9793	lower	3	0.02	0.9
0.0653	0.4147	0.0055	0.003	-0.9708	-0.9793	lower	2	0.02	0.9