Similar Language Detection

Group Members:

Naveen Bhookya(140050034), Chanukya Vardhan(140050043),
Uday Kusupati(140050048), Raghuveer Uppara(140050050).

Stage 1 Report:

- From the training data, for each language we first parse through the sentences and find the frequencies of all the words appearing in all the sentences.
- Among all the words and their corresponding frequencies generated for each language, we choose the top 1000 words based on their frequency values.
- Define x^l for each language l as vector $\in R^{1000}$, x^l = 1000-tuple of frequency of the chosen words.
- Now given an input statement s we define $x^{l}(s)$ for language l as follows:
  $x^l(s) \in R^{1000}$ first intialised to all zeros.
  $x^l_i(s) = 1$ if the ith word of the lth language is present in the input sentence.
- Label for statement s is
  $label(s) = argmax_{l \text{ in } \{1,2,...13\}}$(dotproduct of $x^l(s)$, x^l).
  The intuition is clear.

Observations:

- The confusion Matrix generated after running the code on training data(18000 samples for each language) is as follows:

$C =$

–	bg	bs	cz	esar	eses	hr	id	mk	my	ptbr	ptbt	sk	sr
bg	17379	0	0	0	0	0	0	620	0	0	0	0	1
bs	0	1232	34	0	48	288	0	0	0	51	22	25	16299
cz	0	396	12486	0	199	128	0	0	1	1115	411	1843	1421
esar	1	0	0	0	17999	0	0	0	0	0	0	0	0
eses	2	0	0	0	17998	0	0	0	0	0	0	0	0
hr	0	1872	57	0	73	1146	0	0	0	92	31	35	14694
id	2	0	1	0	29	0	14879	0	3077	7	4	0	1
mk	36	1	1	0	0	0	0	17961	0	0	0	0	0
my	0	0	9	0	25	0	2227	0	15715	1	1	2	2
ptbr	1	0	0	0	9819	0	0	0	0	7557	622	0	1
ptbt	1	0	0	0	10879	0	0	0	0	5822	1297	0	1
sk	0	75	151	0	62	49	0	0	0	943	349	15769	602
sr	0	214	5	0	59	38	0	0	0	38	12	29	17605

- This will be used for dividing the languages into groups.