Quantitative Error Analysis
Joshua Shapiro | 18 October 2016

Accuracy:

Experimental Condition	Overall Accuracy %
BigramLetterLangId	60.67%
BigramWordLangId-AO	62.00%
BigramWordLangId-GT	88.67%
TrigramWordLangId-KBO	98.67%

Confusion Matrix:

BigramLetterLangId

	ENGLISH	FRENCH	GERMAN
ENGLISH	N/A	16	41
FRENCH	2	N/A	0
GERMAN	0	0	N/A

BigramWOrdLangId-AO

	ENGLISH	FRENCH	GERMAN
ENGLISH	N/A	12	30
FRENCH	6	N/A	9
GERMAN	0	0	N/A

BigramWordLangID-GT

	ENGLISH	FRENCH	GERMAN
ENGLISH	N/A	1	5
FRENCH	6	N/A	5
GERMAN	0	0	N/A

TrigramWordLangId-KBO

	ENGLISH	FRENCH	GERMAN
ENGLISH	N/A	0	0
FRENCH	2	N/A	0
GERMAN	0	0	N/A

Perplexity Measure:

Experimental Condition	ENGLISH	FRENCH	GERMAN
BigramLetterLangId	INFINITY	INFINITY	INFINITY
BIgramWordLangId-AO	5590.6	6409.6	7278.7
BigramWordLangId-GT	31419045.9	18154781.0	51987043.7
TrigramWordLangId-KBO	116511557.1	6263714.8	139670097.7

It is not surprising that the perplexity for BigramLetterLangId is infinity, since the some of the characters in the test set have not been seen in the training set. When this happens, the probability for the bigram with the unseen letter in it = 0, and the probability for the entire test data is 0. Since we can't divide 1/0, we call the perplexity infinity.