Project
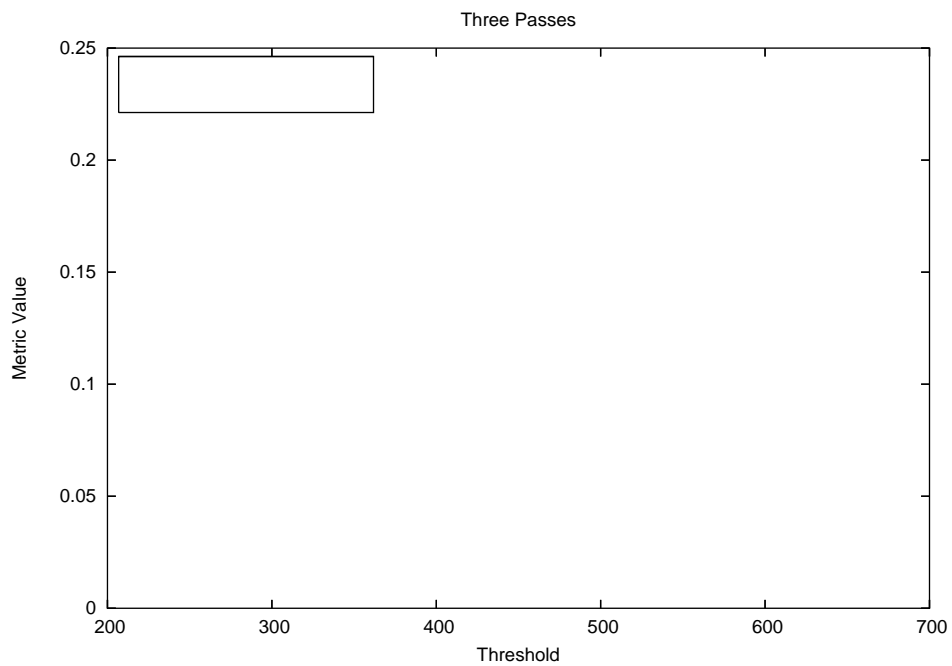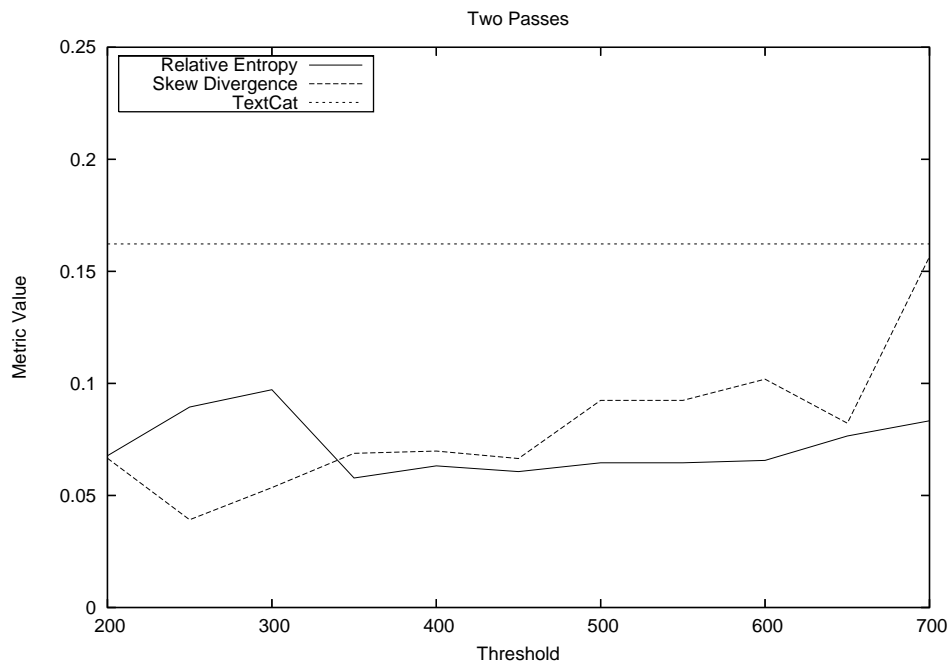
ns per sed L n ẽ den t c ton

Jeremy Nicholson

We chose a well-known information theoretic measure, relative entropy, which has been applied to language identification before [SR96]. Relative entropy, also known as Kullback-Liebler divergence, is a comparison of probability distributions p and q with lexicons P and Q. The usual conventions of $0 \cdot \}$ $\frac{0}{q}$ $\cdot\}$ p $\frac{p}{0}$ ∞ are assumed. D

$$\sum_{x} p(x) \cdot \} P(x) \quad (x)$$

Evaluation

We presented an instance of the language identification task using, like Sibun and Reynar

Two Passes

Legend:
- Relative Entropy ———
- Skew Divergence − − −
- TextCat ·········

X-axis: Threshold (200 – 700)
Y-axis: Metric Value (0 – 0.25)

Three Passes

X-axis: Threshold (200 – 700)
Y-axis: Metric Value (0 – 0.25)

Skew Divergence
TextCat