A Supplemental Material

A.1 Item weights in weighted least-squares regression

We used linear or non-linear weighted least squares regression routines to fit the parameters a, b and c of the extrapolation models described in section 3. The ℓ data items $(n_j, e_j), j = 1, \ldots, \ell$, for these regressions are the error rates e_j of the document classifier system when trained on subsets of size n_j of the training data. The regression routines minimise the weighted sum S of the squares of the residuals:

$$S = \sum_{j=1}^{\ell} w_j (e_j - f(n_j; a, b, c))^2$$

Here f is the extrapolation model, and w_j is the *weight* placed on item (n_j, e_j). Theoretically, the optimal weight w_j is the inverse of the variance of e_j, but we don't know this variance.

We investigated three different weighting functions in this paper, which correspond to different assumptions about the variance of e_j:

- constant weights ($w_j = 1$),
- linear weights ($w_j = n_j$), and
- binomial weights ($w_j = {}^{n_j}/_{e_j(1 - e_j)}$)

We experimented with constant weights because these are the default weights provided by the regression routines.

Linear weights are motivated by the Central Limit Theorem, which implies that the variance of the mean of i.i.d. variables decreases as $O({}^1/_n)$ as $n \to \infty$.

Binomial weights are motivated by the assumption that the success or failure of each document classification can be modelled by a draw from a binomial distribution, so the variance of e_j should be the variance of the estimate of the success probability p given a sample of size n drawn from a binomial distribution:

$$\frac{p(1 - p)}{n}$$

Binomial weights are obtained by assuming that $p \approxeq e_j$. Binomial weights place more weight on data items with larger n_j than constant or linear weights. This seems reasonable, especially since our goal is to extrapolate to even larger values of n.

Clearly these three weighting functions only scratch the surface of possible weighting functions. Because our task is extrapolation, weighting functions that place even more weight on n might do well. It might also be possible to use methods such as the bootstrap to provide more accurate estimates of e_j and its variance.