

# NLP Project #4

## Stopword Detection

– by Ali Mohammad Pur Fard

### Top-level Explanation

0. The documents are first tokenised normally
1. The DF|TF matrix is generated for all the word-types
2. Each word-type is evaluated with the (below) given **score** function
3. The top 300 words in order of RSD are picked as stopwords (somewhat arbitrary)

### Evaluation Measure

The **score** function is defined as such

$$score_{(x)} = \mathcal{CV}_x = \frac{\sigma_x}{\bar{x}}$$

and it is applied to the DF|TF matrix in such a manner:

$$scoredVector(DF|TF) = \sum_{rowsofTF} score(DF \cdot i)$$

with *scoredVector* being the resulting evaluated vector

It is quite apparent that neither the TF matrix, nor the DF vector are normalised. This is simply due to the fact that our chosen evaluation method works on absolute units (quite simple, it is self-normalising anyway)

### Pros and Cons of this evaluation method

#### Pros

1. no prior normalisation is necessary
2. can be calculated in exactly one pass over each vector (see `Array::'relativeStandardDeviation'`)

#### Cons

1. when a word-type's TF matrix is really sparse, tends to generate very big results (>1000% RSD%)