# NLP Class Project (CWI)

## Features:

The following additional features were added : Pos tags, named enitity and log probabilities (unigram) of the target phrase for the English language only. The results are in the table below.

## Observations:

1. Macro F1 is always better when original baseline features of length of characters and length of tokens are added than when they are not.

2. Named entity gives the poorest scores (Macro f1: 0.38) and so is abandoned and the results are not displayed in the table below.

3. The highest macro F1 score (0.71) is achieved when all three features of PoS tags, unigram probabilities and baseline were used.

4. Precision for class 1 was imporved by adding unigram probabilities to baseline features and PoS tags which seems to have contributed to slightly better F1 score.

5. The metric that shows the lowest scores and needs more improvement is the recall for label 1 (complex word). The recall does not improve beyond 52% which means the classifier has primarily learnt to predict the majority class i.e. the label 0 (non complex word). Perhaps this can be imporved by tuning the classifier or finding better features or both.

## Next Steps

1. Experiment with features of word vectors and number of syllables

2. Tune classifier to make it less biased towards to majority class.

3. Try classifiers like decision trees and SVMs.

| **Features: POS+baseline** | **Accuracy:72.89%** | **Macro f1: 0.70** | |
|---|---|---|---|
| *Label* | *Precision* | *Recall* | *F1* |
| 0 | 0.72 | 0.88 | 0.79 |
| 1 | 0.75 | 0.52 | 0.68 |
| **Features: Baseline+unigram probabilities** | **Accuracy: 73.1%** | **Macro f1: 0.70** | |
| 0 | 0.72 | 0.89 | 0.79 |
| 1 | 0.76 | 0.52 | 0.62 |
| **Features: POS+unigram probabilities** | **Accuracy: 65.8%** | **Macro f1: 0.57** | |
| 0 | 0.64 | 0.95 | 0.76 |
| 1 | 0.79 | 0.24 | 0.37 |
| **Features: POS+baseline+unigram probabilities** | **Accuracy:73.73%** | **Macro f1: 0.71** | |
| 0 | 0.72 | 0.89 | 0.80 |
| 1 | 0.78 | 0.52 | 0.62 |