Named Entity Recognition : Using Naive Bayes

- **Abhishek Vijayan**
B110158CS

Features Used :

- ifPresentInDict – indicates if word is present in comWord.txt
- ifAllCaps
- ifAllSmall
- ifNumber
- ifAlphaNum
- containsHyphen
- containsSubstr – 325 features
  - 1 feature corresponding to each string present in prefix_suffix list.txt

Modifications to prefix_suffix list.txt

- Added the following words
  - cell
  - line
  - myeloid
  - lymphocyte
  - mononuclear
  - factor-kappa
  - codon
  - upstream
  - enhancer
- Modified 'mrna' to 'mRNA'
- Added a space in the first line before the word

Modifications to comWord.txt

- Added a space in the first line before the word

Results

Training Data

Metric	DNA	RNA	protein	cell_type	cell_line	other
Precision	0.720151	0.755682	0.518759	0.750581	0.328679	0.872786
Recall	0.142744	0.321644	0.557343	0.146080	0.077643	0.956799
F-Measure	0.238262	0.451230	0.537360	0.244562	0.125613	0.912863

Macro Average

Precision = 0.657773

Recall = 0.367042

F-Measure = 0.418315

Test Data

Metric	DNA	RNA	protein	cell_type	cell_line	other
Precision	0.636496	0.584507	0.447874	0.830369	0.238434	0.887671
Recall	0.152501	0.272131	0.522679	0.123124	0.089572	0.947670
F-Measure	0.246050	0.371365	0.482394	0.214450	0.130224	0.916690

Macro Average

Precision = 0.604225

Recall = 0.351279

F-Measure = 0.393529