**Bio-molecular Even Extraction using NLP Project**

**Final Product URL:** [**http://ec2-18-215-155-13.compute-1.amazonaws.com:5000/**](http://ec2-18-215-155-13.compute-1.amazonaws.com:5000/)

**Git repository:**

Used Stanza – A python NLP package

Treebank: CRAFT & MIMIC

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Domain & Type | Dataset (NER models) | Stanza (micro-average F1 score) |
| Bio | Chemical, Disease | BC5CDR | 88.08 |
| Bio | 16 types of Cancer Genetics | BioNLP13CG | 84.34 |
| Clinical | Problem, Test, Treatment | I2b2 | 88.13 |

Ref: <https://stanfordnlp.github.io/stanza/biomed_model_performance.html>

**Experiments**

ScispaCy

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | UAS | LAS | POS | F1 | Web UAS |
| en\_core\_sci\_sm | 89.54 | 87.62 | 98.32 | 68.15 | 87.62 |

Suggestion: We can use en\_ner\_bc5cdr\_md to enhance F1 score by 84.49 in Diseases & Chemical Entity types

Ref: <https://allenai.github.io/scispacy/>

Source code: <https://github.com/allenai/scispacy>

Future Scope: Working on BioBert model

Ref: <https://github.com/naver/biobert-pretrained>

Name: Saurav Gupta

Email: sauravgupta02oct@gmail.com