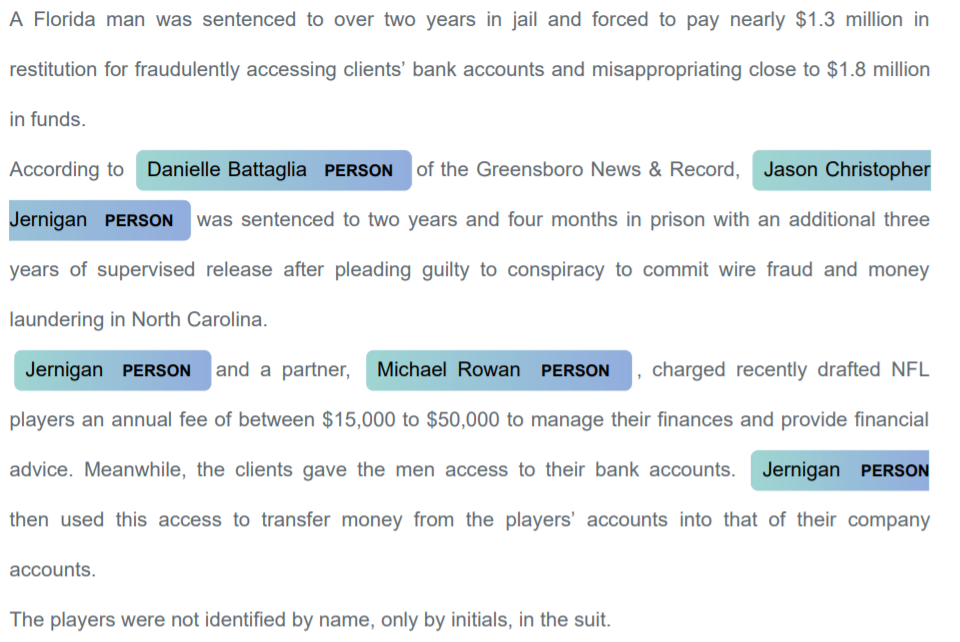
**Entity Extraction**

For each article we need to extract the individuals that have been mentioned in that article so, the aim is to extract the entities using Named Entity Recognition and also link all mentions/references of that individual to that person using Coreference Resolution (linking pronouns to the individual they refer to).

**Visualization**



We use these to get sub-article for each individual , where each sub-article is set of sentences from article that mention the particular individual. The end result is as follows :-



so for each individual we get sub-article that deals with him/her.

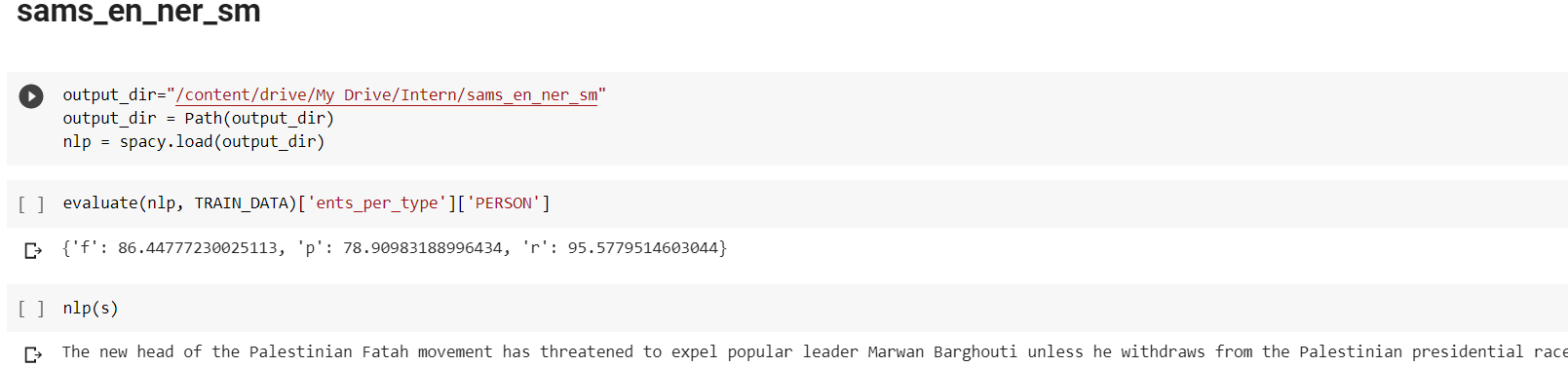
Current best NER pre-trained models in the Industry :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MODEL NAME | P | R | F1 | TECHNIQUE | EX. TIME  (/1000words) |
| SPACY(web\_en\_core\_lg) | 36.3 | 67.1 | 46.4 | Single layer NN | 0.05s |
| Deepavlov(CONELL) | 84.2 | 85.6 | 84.89 | BERT(conell-2003) | 0.5s |
| StanfordNLP | - | - | - | CRF | 5.2s |
| BERT(not-pretrained) | 92.9 | 93.2 | 93.5 | BERT | 1.2s |
| SPACY(sams\_en\_core\_lg) | 78.9 | 95.5 | 86.4 | Single Layer NN | 0.05s |

When tested on another news article corpus

CURRENT MODEL:

SPACY(sams\_en\_core)- Trained on our corpus with the help of Deepavlov for tagging the entities. So in order to train a spacy model we needed a large amount of entity-tagged dataset, and as our dataset was a mass of textual data on a specific person…we used Deepavlov(CONELL.BERT) model to tag our dataset with entities as PERSON and NPERSON. Then we used this dataset to train a spacy model.



Sams\_en\_ner\_sm evaluated on Annotated Corpus for Named Entity Recognition