**How to work on word document**

Using docx library – we can fetch the data in different parts like header, footer, paragraph,run.

But while replacing the data in the word file we lost the format of the data.

So for this we need to replace the text in the XML format of the docx by using zipfile library.

We can also use win32 client to read and write document nut this words only in window not for ubantu.

Different techniques to detect NER and their limitations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Core NLP** | **NLTK** | **spaCy** | **Rasa NLU** | **Syntaxnet** |
| **Time** | Takes more time as comparison to other libraries. | Provide better speed than Core NLP. | Better speed than any other library. | It provides only pipeline to process the text data, basically it uses certain model to extract the NER like spacy, CRF. | Syntaxnet does not explicitly offer any Named Entity Recognition functionality but using regex we can do POS tagging and then can use this algo to extract the entities. |
| **Result** | Accuracy is approx. 78 % | Accuracy is approx. 51 % | Accuracy is approx. 72 % |
| **Custom entities** | Using regex we can add custom entity but have certain limitation for different kinds of data. | Using regex we can add custom entity but have certain limitation for different kinds of data. | We can add custom entity by using training data or by using regex. |
| **Implementation** | Hard to implement as we need to write the entire code for different entity and process each part like POS tagging. | Medium but results are not accurate as desired. | Easy to implement as we need to call entity detection method. | it depend on model, which model we are going to use. | Hard to implement |

We can use below algorithms also but for these we need training data to train our model and predict the NER.

* CRF
* LTSM
* BERT
* Syntaxnet

**Anonymization:**

1. Faker
2. We can write a Function which can encrypt or remove the NER from the original text.