Named Entity Recognition (NER)

NER is the process of identifying and classifying named entities. So, the accuracy of a NER model is dependent of the number of entities identifies and accuracy of the tags.

In NLP, two major go-to options for NER are spaCy and NLTK. One could also adopt some of the transformer-based models on HuggingFace for named entity recognition.

SPACY

SpaCy is an open-source NLP library that has a number of pipelines with different model components across major languages like English. These components are for:

- Token2vec
- Part of Speech Tagging
- Dependency Parser
- Attribute ruler
- Lemmatizer
- Named Entity Recognition.

While implementing any of the pipelines, one can also select which component they want to perform with. For example, NER.

SpaCy pipeline variants for English can be found here. They were all trained on texts gotten from the web like news, blogs, etc. **en_core_web_sm** has the smallest size of 12MB unlike **en_core_web_trf,** which has a size of 438MB. The trade-off between the both is that the former is most efficient, while the latter is most accurate.

The pipelines are trained on CNNs, except for **en_core_web_trf** which was combined with HuggingFace transformer for better accuracy.

SpaCy models were trained using the Thinc Model, which is a wrapper for PyTorch, TensorFlow and MXNET. In this case, the wrapper was used on PyTorch to run the CNN architectures.

The entities available in spaCy are 18 in number: **'CARDINAL', 'DATE', 'EVENT', 'FAC', 'GPE', 'LANGUAGE', 'LAW', 'LOC', 'MONEY', 'NORP', 'ORDINAL', 'ORG', 'PERCENT', 'PERSON', 'PRODUCT', 'QUANTITY', 'TIME', 'WORK_OF_ART'.**

NLTK

There is no much public information on the model behind NLTK NER. However, NLTK also has a wrapper for the StanfordNER that was built on Java [source] but is said to be computationally expensive. The tags available could be between 3 to 7 depending on choice and they are l**ocations, persons, organizations, times, money, percents, and dates.**

QUALITATIVE MODEL COMPARISONS:

The following sentences are annotated using different models, and their results are displayed in the table below.

(Pink highlights represent wrong labels)

"WASHINGTON -- In the wake of a string of abuses by New York police officers in the 1990s, Loretta E. Lynch, the top federal prosecutor in Brooklyn, spoke forcefully about the pain of a broken trust that African-Americans felt and said the responsibility for repairing generations of miscommunication and mistrust fell to law enforcement.",

"Nigerian President Muhammadu Buhari has condemned the killing of one of two Catholic priests by kidnappers in the northern state of Kaduna.",

'He said the armed groups behind the spate of kidnappings of Christian clerics "seem to be bent on creating chaos and disorder in the country"',

"Emmanuel Onuarah, the president of the Premium Bread Makers Association, told the BBC that some bakeries had had to fire their staff as they were not able to pay salaries.",

"He suggested that the government should stop charging a 15% tax on imported wheat - the price of which has already shot up this year because of the war in Ukraine.",

"The youths had attended a wedding in Imo state and were on their way home when they were shot, Amnesty said.",

"The famous Nigerian author Chimamanda Ngozi Adichie has thrown her support behind the Labour Party's presidential candidate, Peter Obi, ahead of the country's February 2023 elections.",

"He was dubbed Super Mario for his handling of the eurozone crisis as head of the European Central Bank. In February last year, he was given the task of guiding Italy through the Covid pandemic and economic recovery, bolstered by a big EU package conditional on major reforms."

spaCy: en_core_web_sm **(Model size 12mb)**	**spaCy: en_core_web_md** **(Model size: 40mb)**	**NLTK** **(N/A)**
WASHINGTON GPE New York GPE the 1990s DATE Loretta E. Lynch PERSON Brooklyn GPE African-Americans NORP	WASHINGTON GPE New York GPE the 1990s DATE Loretta E. Lynch PERSON Brooklyn GPE African-Americans NORP	GPE WASHINGTON GPE New York PERSON Loretta E. Lynch GPE Brooklyn
Nigerian NORP Muhammadu Buhari PERSON one CARDINAL two CARDINAL Catholic NORP Kaduna GPE	Nigerian NORP Muhammadu Buhari PERSON one CARDINAL two CARDINAL Catholic NORP Kaduna GPE	GPE Nigerian PERSON Muhammadu Buhari ORGANIZATION Catholic GPE Kaduna

Christian NORP	Christian NORP	GPE Christian
Emmanuel Onuarah PERSON the Premium Bread Makers Association ORG BBC ORG	Emmanuel Onuarah PERSON the Premium Bread Makers Association ORG BBC ORG	PERSON Emmanuel ORGANIZATION Onuarah ORGANIZATION BBC
15% PERCENT this year DATE Ukraine GPE	15% PERCENT this year DATE Ukraine GPE	GPE Ukraine
Amnesty ORG	Amnesty ORG	GPE Imo PERSON Amnesty
Nigerian NORP Chimamanda Ngozi Adichie PERSON the Labour Party's ORG Peter Obi PERSON February 2023 DATE	Nigerian NORP Chimamanda Ngozi Adichie PERSON the Labour Party's ORG Peter Obi PERSON February 2023 DATE	GPE Nigerian ORGANIZATION Chimamanda Ngozi Adichie ORGANIZATION Labour Party PERSON Peter Obi
the European Central Bank ORG February last year DATE Italy GPE EU GPE	the European Central Bank ORG February last year DATE Italy GPE EU ORG	PERSON Super Mario ORGANIZATION European Central Bank GPE Italy GPE Covid GPE EU