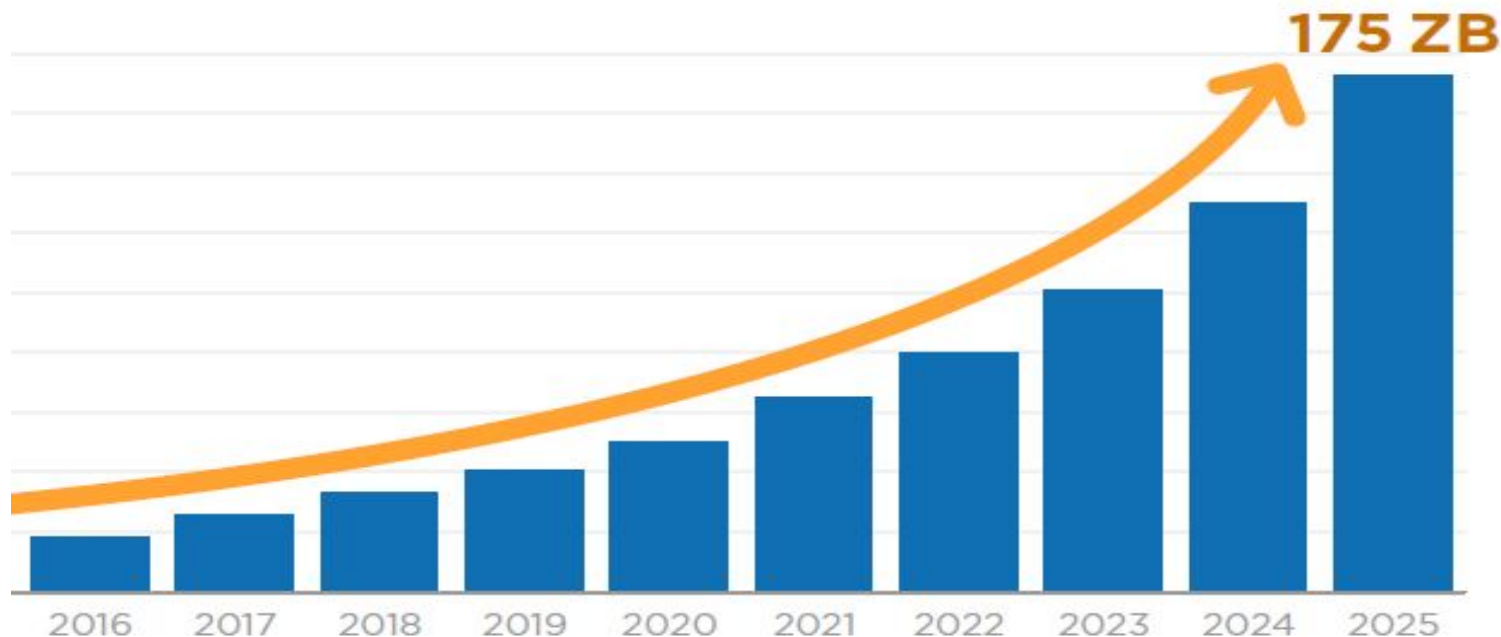


ML NER Custom Model and Network Visualization

17.12.2023
Serhii Khilimon

Data created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025



*<https://www.statista.com/statistics/871513/worldwide-data-created/>

Can you imagine **TRUE** detective?



Not So True



Sad but True

Movies Vs Reality



Every single day Law Enforcement Agency encounter with paper storm of structured and unstructured documents of criminal cases

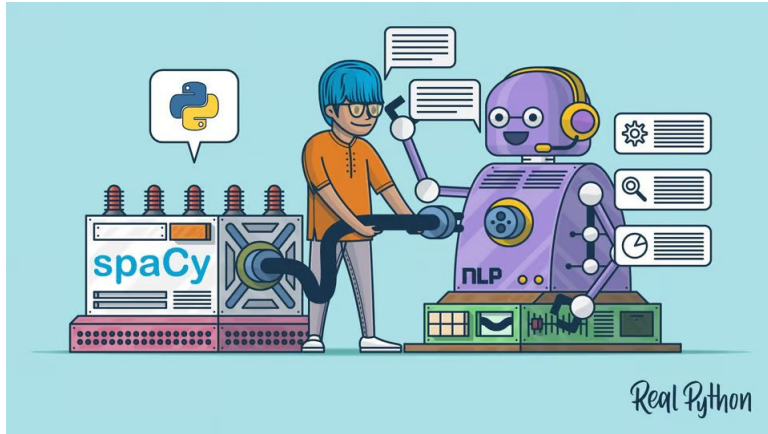
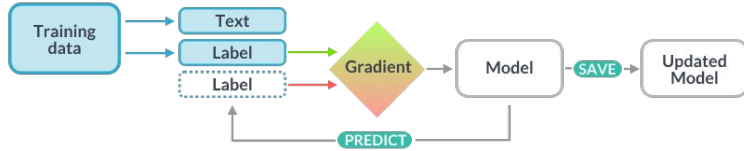


What we need is:

- Reduce time spent on processing unstructured documents by 40 %;
- Increase a quality of NER by 25 %;
- Expand the field of research by 50 % (by including new types of entities)

*picture is generated by <https://gencraft.com/>

Here is a Plan of Attack:



- to build **NER** system based on **spaCy's** pre-trained model to process unstructured documents (PDFs, Doc, HTML etc.);
- - to enrich default list of model's Named Entities with custom types of Entities, based on needs of stakeholder (for example - "Drugs", "Weapon", "Crime" etc.);
- - to retrain model with custom types of Entities;
- - to bring marked **NE** into Link Analysis (Social Network Analysis) tool for visualization;
- - to accumulate detected **NE** into local or cloud store;

(**CNN** **ORG**) - Amy Schneider **PERSON** , an engineering manager from **Oakland** **GPE** , **California** **GPE** , became the first **ORDINAL** woman and the fourth **ORDINAL** person on " **Jeopardy** **WORK_OF_ART** !" to earn more than \$1 million **MONEY** in winnings on Friday **DATE** 's episode

spaCy's `en_core_web_lg`

Feature	Description
Name	<code>en_core_web_lg</code>
Version	<code>3.7.1</code>
spaCy	<code>>=3.7.2,<3.8.0</code>
Default Pipeline	<code>tok2vec</code> , <code>tagger</code> , <code>parser</code> , <code>attribute_ruler</code> , <code>lemmatizer</code> , <code>ner</code>
Components	<code>tok2vec</code> , <code>tagger</code> , <code>parser</code> , <code>senter</code> , <code>attribute_ruler</code> , <code>lemmatizer</code> , <code>ner</code>
Vectors	514157 keys, 514157 unique vectors (300 dimensions)



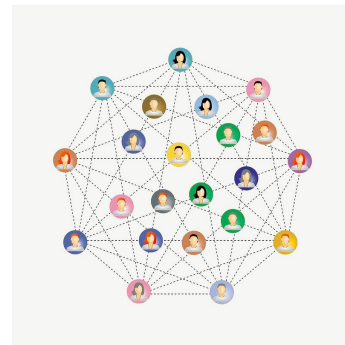
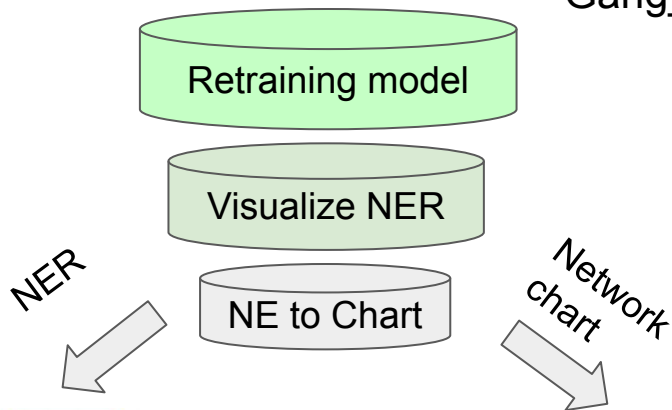
Custom Entities types (via `ruler.add_patterns`)



“Gang_Org”



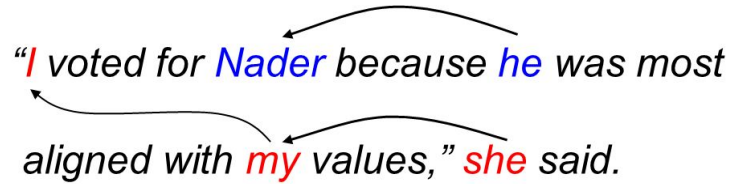
“Drug”



(`CNN` `ORG`) - Amy Schneider `PERSON` , an engineering manager from `Oakland` `GPE` , `California` `GPE` , became the first `ORDINAL` woman and the fourth `ORDINAL` person on " Jeopardy `WORK_OF_ART` !" to earn more than \$1 million `MONEY` in winnings on `Friday` `DATE` 's episode

What is left to do?

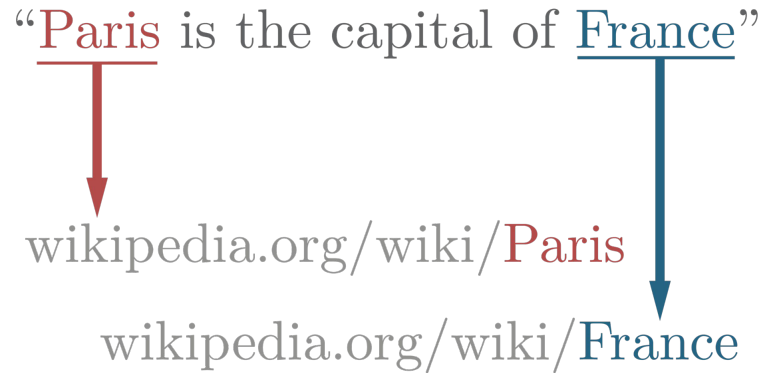
"I voted for Nader because he was most aligned with my values," she said.



A diagram illustrating coreference resolution. In the sentence "I voted for Nader because he was most aligned with my values," she said.", the words "I", "he", "my", and "she" are highlighted in red, blue, red, and red respectively. Curved arrows connect "I" to "she", "he" to "my", and "my" to "she", indicating that these words refer to the same entities.

- To implement NER coreference system
- To implement Entity Linker based on semantic links in sentences and the whole text
- To implement storing results of NER to the local or cloud-based Database

Paris is the capital of France



A diagram illustrating entity linking. The sentence "Paris is the capital of France" is shown. The words "Paris" and "France" are underlined. A red arrow points from "Paris" to the URL "wikipedia.org/wiki/Paris". A blue arrow points from "France" to the URL "wikipedia.org/wiki/France".

wikipedia.org/wiki/Paris

wikipedia.org/wiki/France