Data Science & Analytics

White-List & Black- List: Enhanced NER Model

www.geninvo.com

North America | UK | Europe | India

Business Need

- A spacy Named Entity Recognition(NER) model has been trained to detect named entities from the documents. But due to lack of enough training data this model sometimes misidentifies or ignores some of the entities. These entities fall into the categories of false positive or false negative.
- False positive entities are those entities which are not into any of the NER categories but falsely identified as named entities on the other hand false negatives are those which were named entities but somehow not detected by the model as named entities.
- So, in order to correctly identify these entities and to make the model learn from these mistake there was a need to build the Blacklist and whitelist.

Solution Methodology

- Simple algorithm was developed which uses a python dictionary object which contains filename as first parameter and the second parameter is a Boolean value indicating whether the dictionary should be written as false negative or false positive. The default value is false negative.
- There are some other helper functions which helps to keep track of the items to be added in the list. whenever a new document is opened it first checks from the stored blacklist- whitelist whether any item is already present in those lists it won't update the list but if the item is not already present in the lists and is identified as black or white- list item it is the added to the corresponding lists.
- This way the list grows every time the item is added to the list and the list is stored in the database. This list is then used to retrain the spacy NER model. So, whenever the model is run again it checks into blacklist and whitelist and accordingly it include or exclude the entities in the document.

Solution Methodology

Whitelist & Blacklist

- Create a list for all false positive as Blacklist
- Create a list for all false negative as Whitelist

How users identify something as False Positive/False Negative

- For end user the FP/FN information is available in WLBL Tables. The user can correct or enhance the behavior of strategy application and are able to indicate what level of hierarchy are most applicable for WBL processing and they submit approvals to the Deld Lead and Model Administrator.

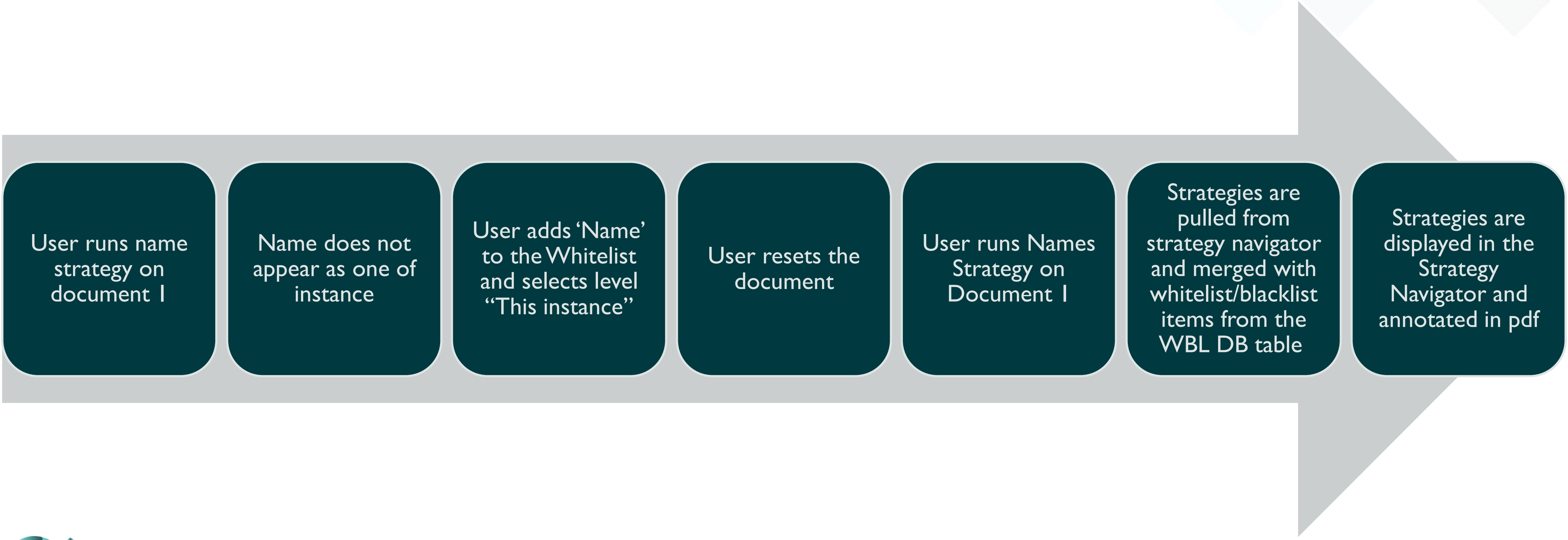
How do we know when to remove items from BL/WL? When to reduce the size ?

- Solution?

Position changed when strategies were applied

- After applying strategy, the new name will be cut off to fit old space. So, position of other words in the document will not change.

Scenario



User runs name strategy on document 1

Name does not appear as one of instance

User adds 'Name' to the Whitelist and selects level "This instance"

User resets the document

User runs Names Strategy on Document 1

Strategies are pulled from strategy navigator and merged with whitelist/blacklist items from the WBL DB table

Strategies are displayed in the Strategy Navigator and annotated in pdf

Business Impact

- It has enhanced the model capability to recognize named entities .
- User's also have the capability to tag the text as false positive or false negative so they can also tag the entities if something is missed or wrongly tagged. This way more useful data is collected from the users and stored as blacklist and whitelist for retraining the model.
- Enhances user engagements and improve their retention.

GENINVO™

Ideas
Delivered
Intuitively

Thank You!

www.geninvo.com | Reach out to us info@geninvo.com