NER (Named Entity Recognition) for Court Cases

KnowHax Hackathon - SCALES

By: Angelica Spratley (MSc), Daniel Burdeno (MSc), Uohna Thiessen, PhD

April 26th, 2024

KNOWHAX-SCALES

Agenda

-	INTRODUCTION
-	BUSINESS PROBLEM
-	DATA OVERVIEW
-	RULE-BASED NER MODEL
-	GLINER ZERO-SHOT NER MODEL
-	CUSTOM TUNED SPACY MODEL
-	LEARNINGS/RECOMMENDATIONS

KNOWHAX-SCALES

PROJECT REPO

Angelica “Jelly” Spratley, MSc

Daniel Burdeno, MSc

Uohna Thiessen,PhD

Business Problem

KNOWHAX

Business Problem

- U.S. Criminal Justice System Spends ~$264B/year
- Weak National Level Data Collection across state systems (hard to easily find who was involved and case outcomes)
- Proper data collection and entity tagging (to identify what entities are involved in cases) can help research efforts about resource allocation for courts in needs, evaluating safe prison conditions, and more!
- This project is focused on tagging ‘government’ vs. ‘person’ entities involved in the cases to help research efforts done by SCALES (Hackathon sponsor)

KNOWHAX

Bottom Line

Rule-Based Model

- Using finite rules can easily ‘tag’ an entity as government or person.
- A model like a Decision Tree can use the rule-based target created above to predict new data with getting a False Negative (identify an entity as a person when in reality it was government) ~0.6% of the time

Custom LLM Model (spaCy)

- Creating custom training data with the labels of ‘government’ and ‘person’ can improve the foundational spaCY small english language pipeline* this is still in the works
- Zero Shot GLiNER Model** was roughly 59% accurate in predicting the rules-based target we created

**spaCy small model documentation: Here*

***GLiNER NER Model documentation: Here*

Data Overview

KNOWHAX

DATA Overview

SPONSOR WEBSCRAPED DOCKETS FROM PACER (PUBLIC ACCESS TO COURT ELECTRONIC RECORDS)

NATIONAL CASES FROM 2002-2020

HACAKTHON TEAM SAMPLED 100K OBSERVATIONS FROM THE SPONSOR'S 3.4M CSV FILE

LLMS WORKED WITH A SMALLER 2K SAMPLE

FEATURES GIVEN:

NAME
EXTRA_INFO
SUIT SUBTYPE
CASE IDS

BINARY FEATURES CREATED FOR RULE BASED MODEL (DOES PATTERN MEET RULE OR NOT)

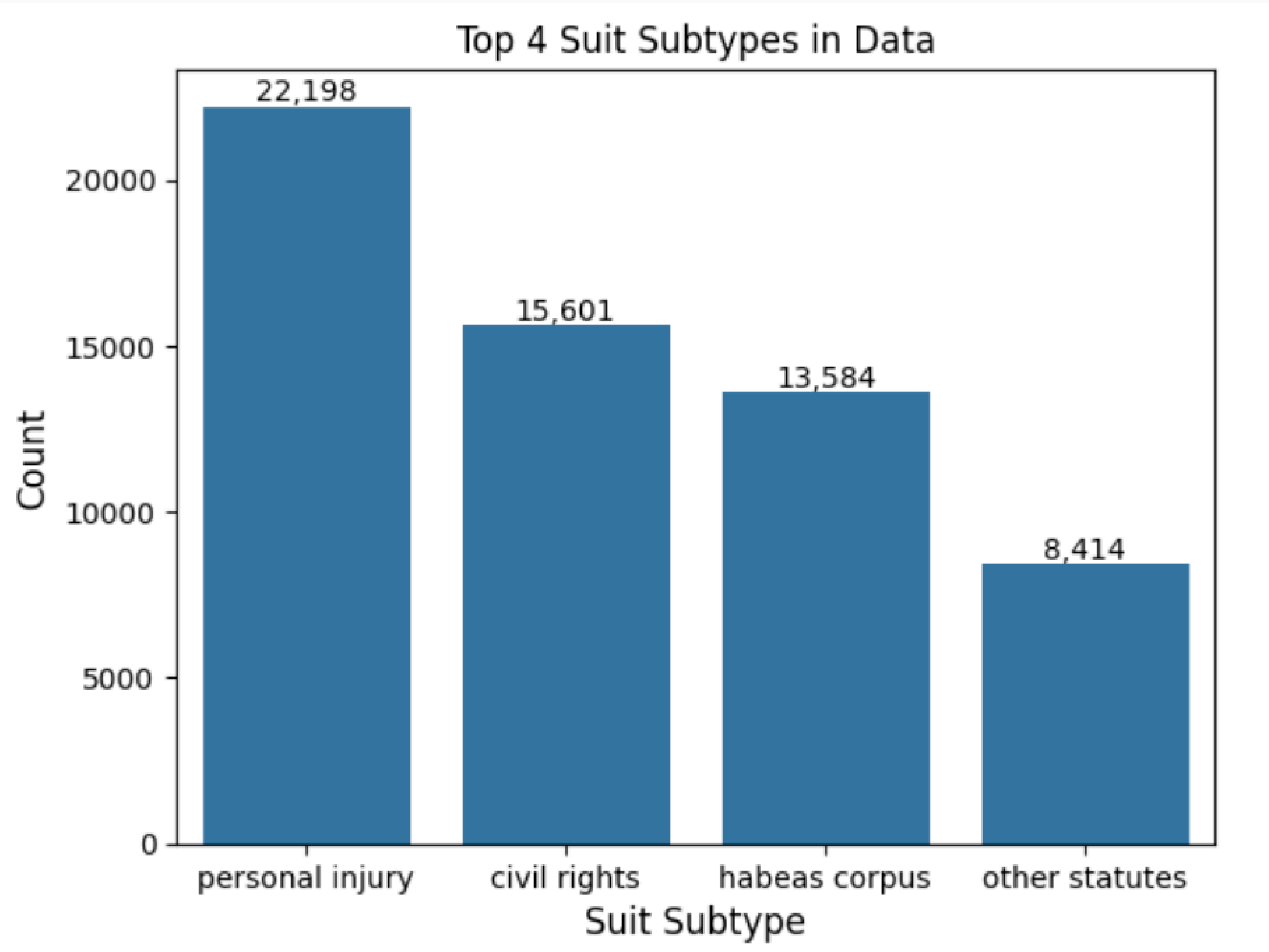
LIMITATIONS:

NON-RULE BASED MODELS APPLIED TO SMALL SAMPLE
FINITE RULES DOES NOT INCLUDE ALL POSSIBLE PATTERNS
COMPUTATIONAL RESOURCES LIMITED FOR LLMS

KNOWHAX

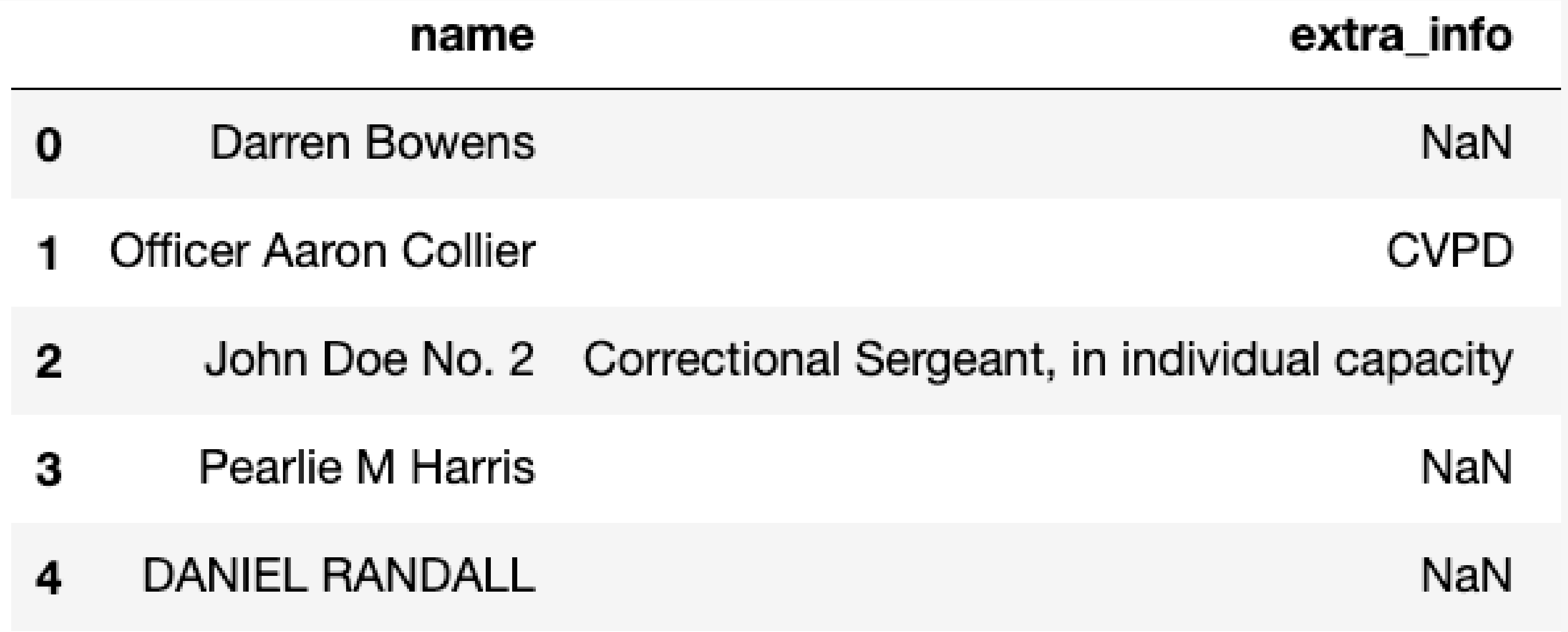
60% of the cases comprised of these 4 suit categories

- DATA CONTAINED 18 TYPES OF SUITS
- TOP 4 SHOWN TO THE RIGHT AND COMPRISED OF ~60% OF THE DATA (100K OBS)

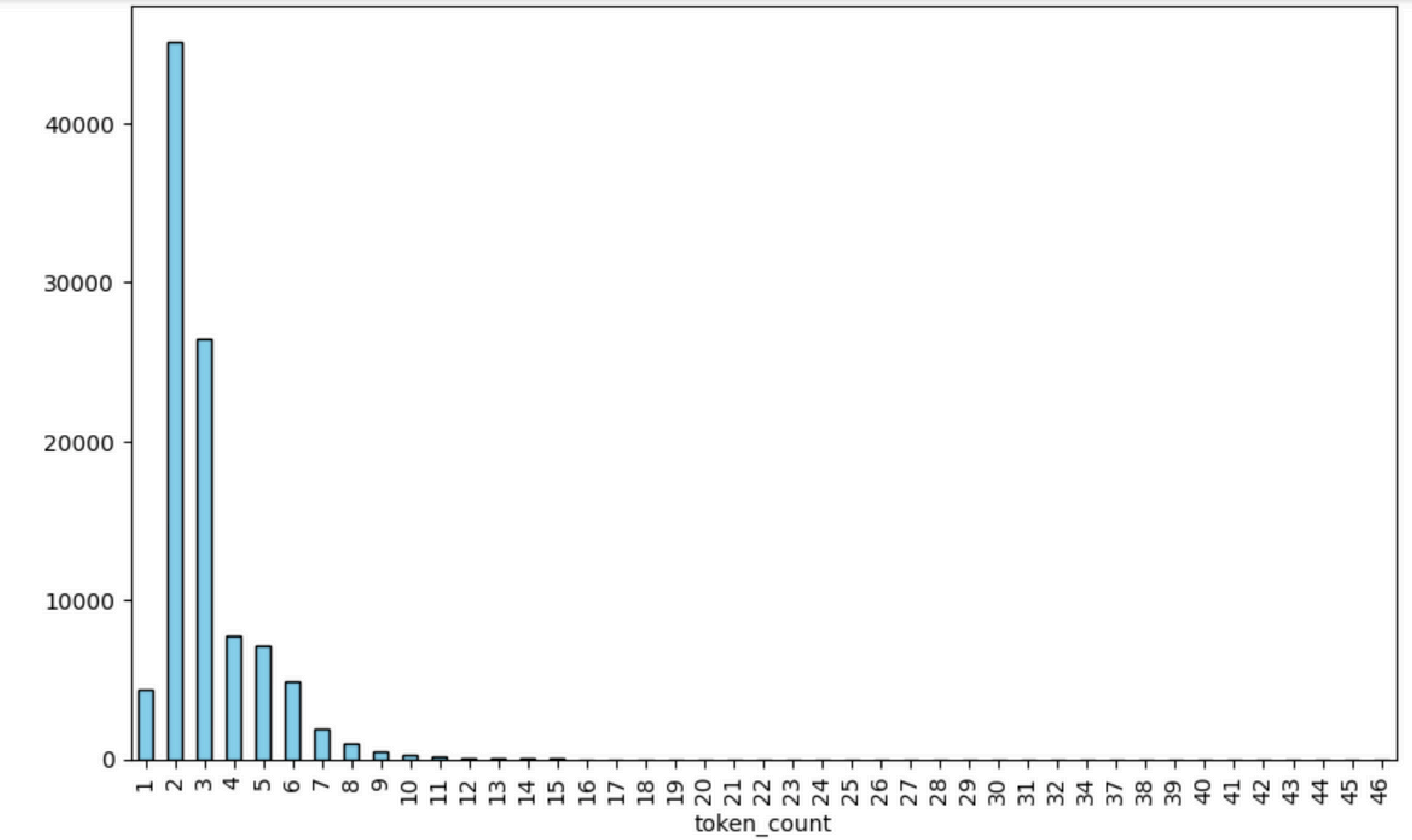


Are 2-3 token names more likely to be a person entity?

- DATA WAS TOKENIZED FOR FURTHER EXPLORATION AND FEATURE ENGINEERING
- TEST THE HYPOTHESIS: ARE 2-3 TOKEN WORDS MOST LIKELY TO BE INDIVIDUALS IF NO FORMAL TITLES EXIST?
- TOKENS CAN RANGE FROM 1-46



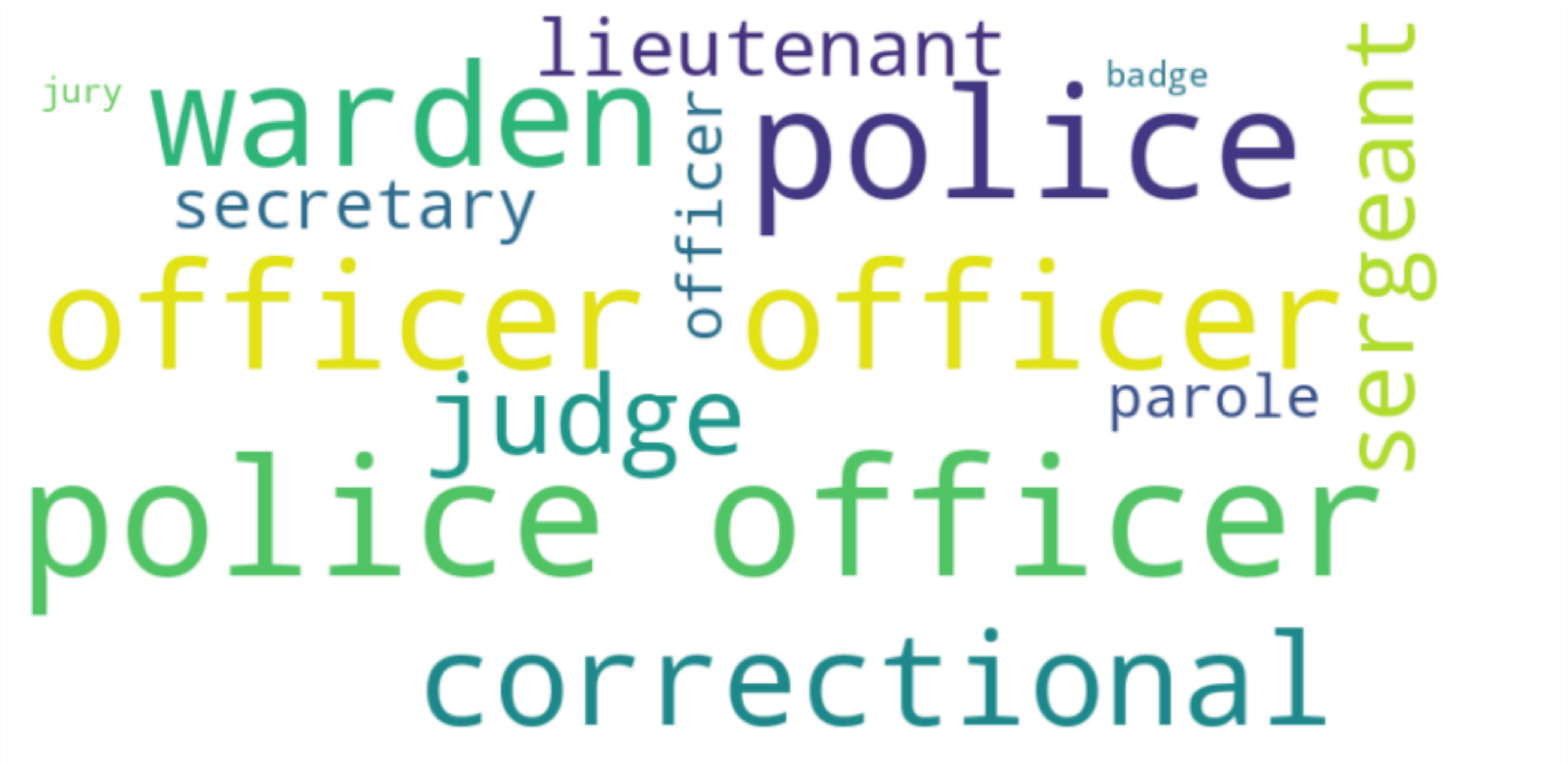
	name	extra_info
0	Darren Bowens	NaN
1	Officer Aaron Collier	CVPD
2	John Doe No. 2	Correctional Sergeant, in individual capacity
3	Pearlie M Harris	NaN
4	DANIEL RANDALL	NaN



KNOWHAX

Formal title frequency over the entire data sample

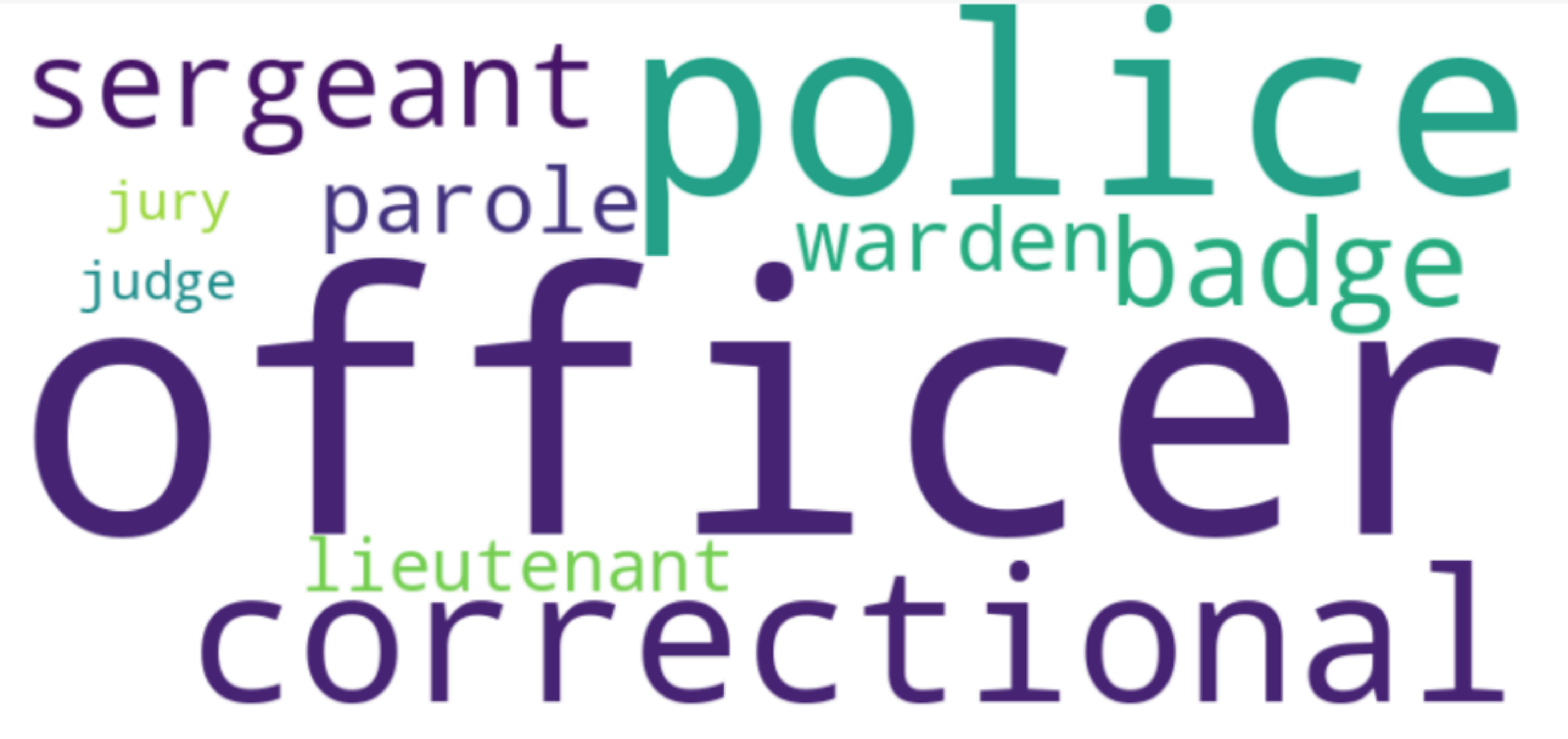
- WORD CLOUD OF THE FORMAL TITLES FOUND IN THE ENTIRE DATASET (100K OBS)
- ANY ENTITY WITH THESE FORMAL TITLES WERE LABELED 'GOVERNMENT'
- FORMAL TITLE IDENTIFICATION IS ONLY AS GOOD AS THE TITLES YOU GIVE IT



KNOWHAX

Even if name is ‘john doe’ use extra_info to classify as ‘govt’

- JOHN DOES EXISTED IN ABOUT 0.6% (LESS THAN 1%) OF THE DATA
- WHEN USING THE ‘EXTRA INFO’ COLUMN WE CAN CLASSIFY JOHN DOES WITH FORMAL GOVT TITLES AS ‘GOVERNMENT’ STILL



KNOWHAX

Rule Based Model

KNOWHAX

Created 12 rules in order to create a target label

- Created several rules based on the business problem and trial and error
- Rules are ordered with importance so once one rule is true it will not be overwritten
- All text was lowercased for standardization and tokenized (tokens around 2 or 3 proved to be important in classifying a ‘government’ from ‘person’ entity)

Examples of some of the rules

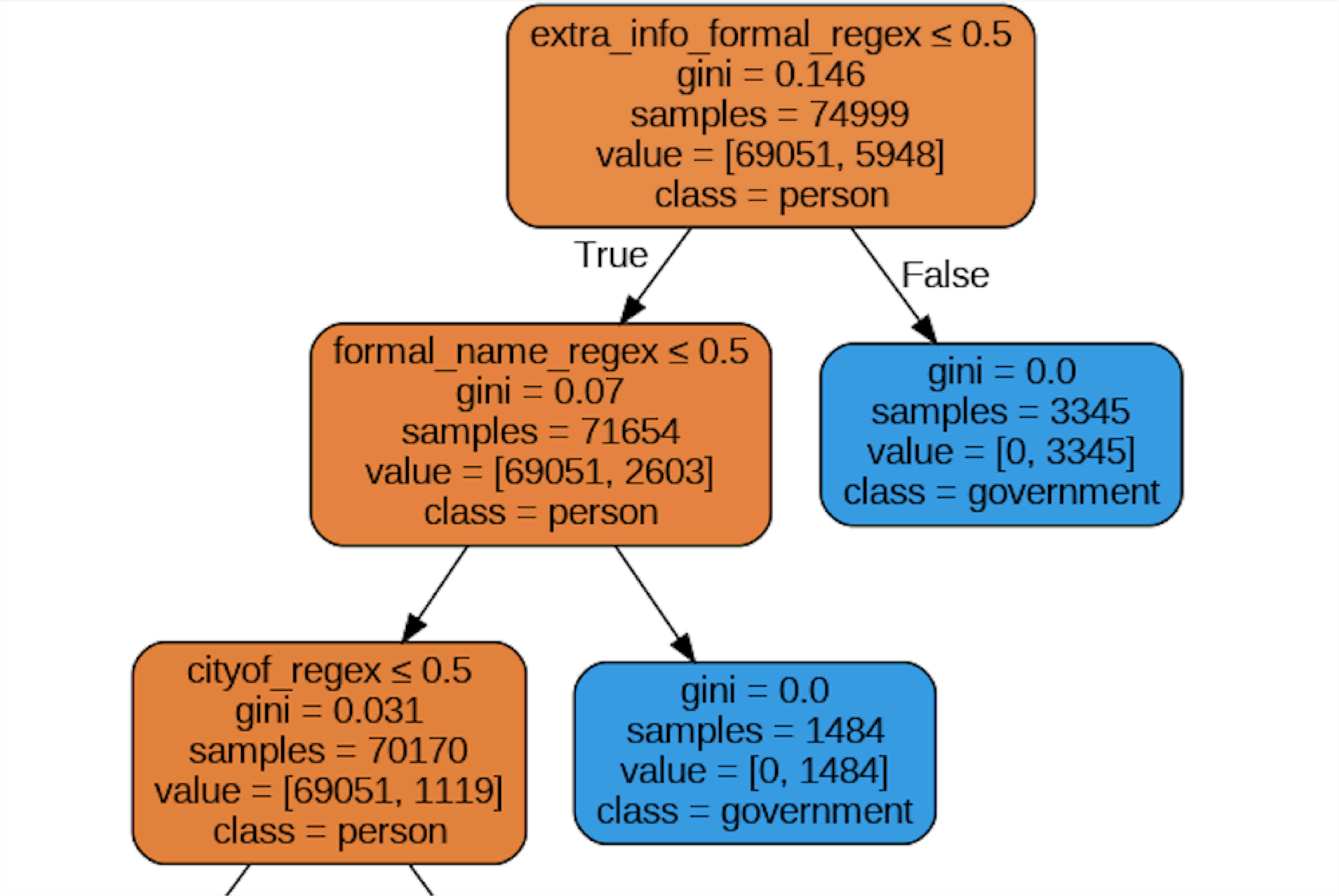
Contains u.s.
Contains 'department of'
Contains 'city of'
DOES NOT Contains 'LLC'
Formal Titles in NAME or Extra_Info fields (secretary, officer, judge, d.a.)

Formal Titles proved to be most important in Decision Tree Model

- Top Three Features of the Decision Tree (ran on 100k obs)
  - Formal Titles in “Extra Info” column (Police Officer, D.A., etc.)
  - Formal Titles in “Name” column
  - “City of” in “Name” column

- Accuracy: 99%
- Recall: 91%

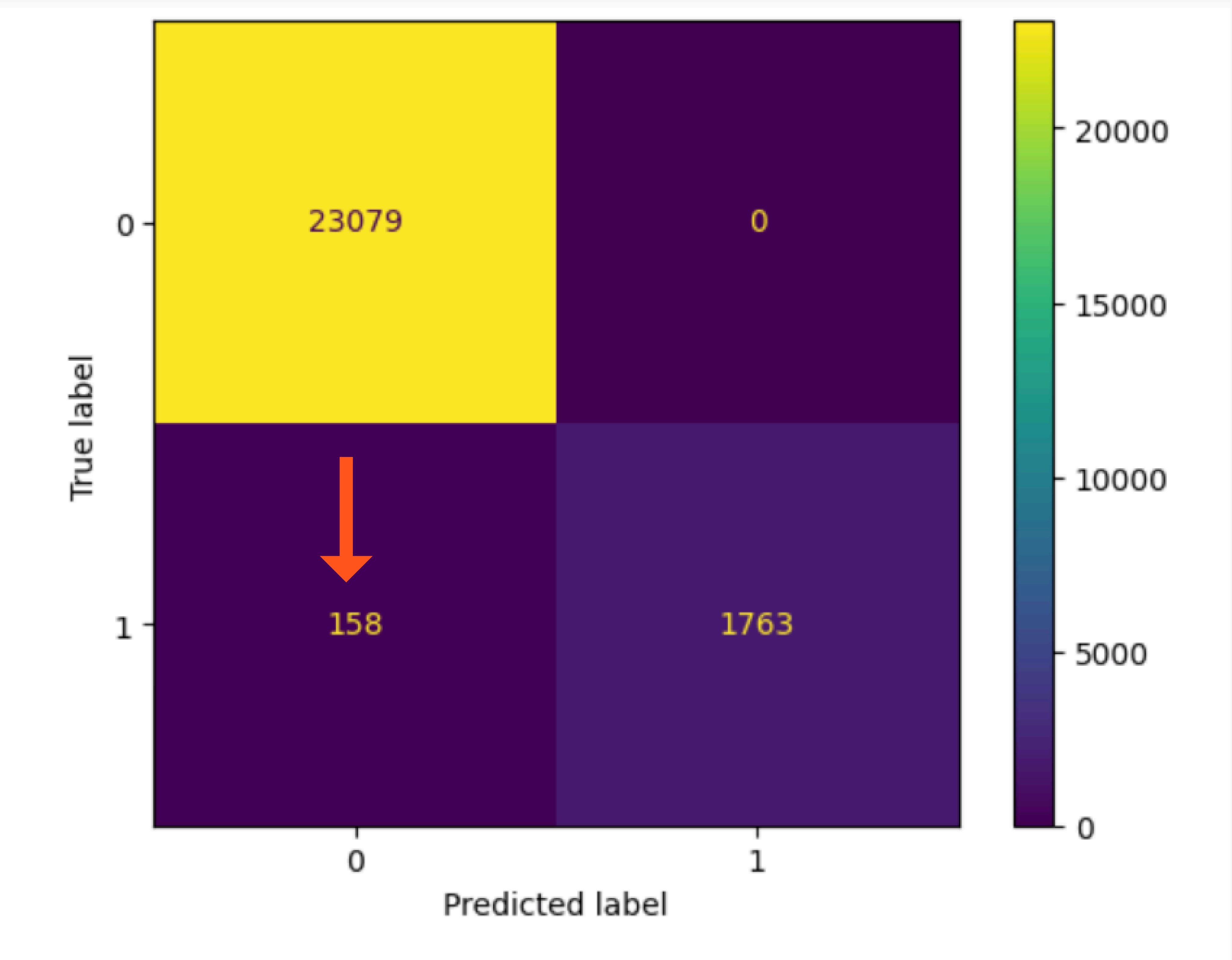
- False Negative: Identify an entity as ‘person’ (0) when in fact it is ‘government’ (1)



KNOWHAX

False Negatives happened ~0.6% of the time

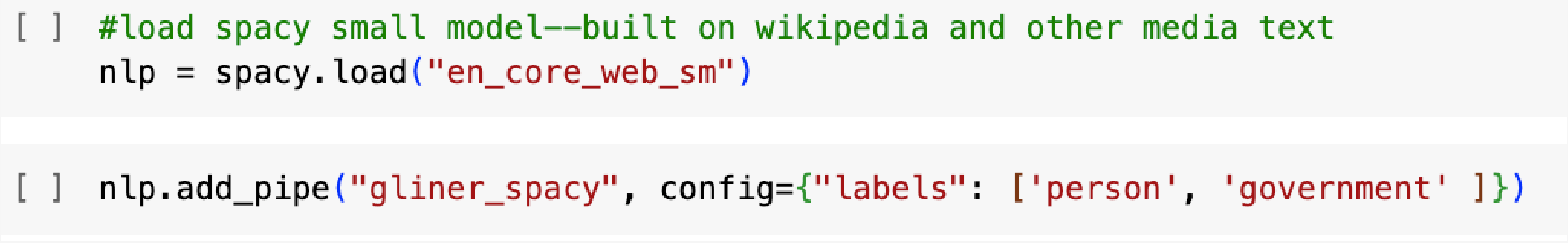
- Model resulted in ~0.6% of False Negatives
  - Identifying entity as a person when it was actually ‘govt’



Zero shot GLiNER LLM

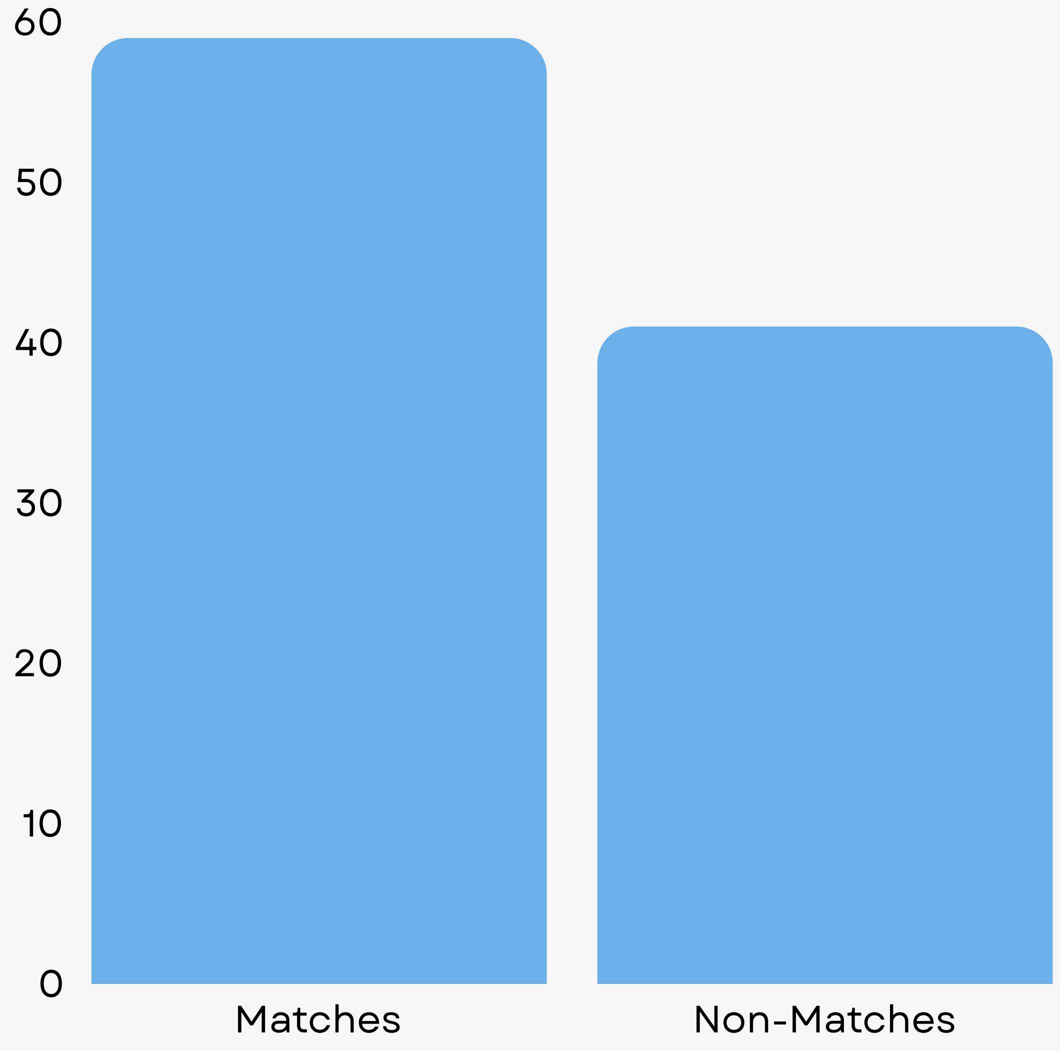
Zero-Shot GLiNER Model Matched 59% of Target

- Zero-Shot model was created to see how well it can classify the created rules-based target of “government” and “person”
- The HuggingFace GLiNER medium-v2.1 was used on a sample of 2k taken from the 100k data set (for computational and time constraints)
- Roughly the model matched 59% of our created target correctly



```
[ ]  #load spacy small model--built on wikipedia and other media text
     nlp = spacy.load("en_core_web_sm")

[ ]  nlp.add_pipe("gliner_spacy", config={"labels": ['person', 'government' ]})
```



KNOWHAX

Tuned spaCy model (in works)

KNOWHAX

Impact of custom training data added to spaCy model

Created Training Data with formal titles in front of names and labeled ‘government’

Created Training Data with person suffixes (i.e. llc, .inc, l.l.c) and labeled ‘person’

Used the spaCy small english language pipeline as a foundation and added this custom training data

Ran 100 plus iterations and got different results each time with the worse being mostly classifying everything as ‘government’

This will be investigated further in another sprint

KNOWHAX

Learnings & Recommendations

Learnings & Reccs

- Creating finite rules to determine a target then running a Machine Learning model against that target performs pretty well
- The limitations is the difficulty in being able to define almost all the rules needed to create a robust target
- John Does in court cases can still be classified by using the ‘extra info’ column
- Even though the zero-shot LLM model was ~59% accurate, it can be used to identify names that were matched to both labels for further investigation or to help create more rules
- Custom training data could potentially be used to train the model but computational (cloud) resources would be needed since it has difficulty performing on just a local machine

THANK YOU!

PROJECT REPO

Angelica “Jelly” Spratley, MSc

Daniel Burdeno, MSc

Uohna Thiessen,PhD