Data Anonymization for Crisis Response: An NLP approach

a project by Aussie Frost for DSCI 410L

Background

- the CAHOOTS organization went on ## cases in 2023

Research Question

Can a script reliably remove identifying information such as names, phone numbers, and addresses from case narratives?

Data

1	time_of_call	call_type	call_transcription
2	2024-12-07 08:51:51	behavioral	Hi, this is Barrett Quill, Drake Vincent and Juno Jasp
3	2024-03-25 23:00:53	outreach	Hi, this is Xander Justice. I'm following up on my rec
4	2024-02-26 10:55:34	behavioral	My name is Fiona Quick, and I am contacting you reg
5	2024-04-25 19:32:59	medical	This is Yvette Ingram and Oscar Pace. I need to orde
6	2024-05-20 22:40:21	medical	We went on a case today. Finn Quest and Juno Frost

Figure 1. Case Narratives

Data

Surname	Approximate Number	% Frequency	Rank
SMITH	2,501,922	1.006	1
JOHNSON	2,014,470	0.81	2
WILLIAMS	1,738,413	0.699	3

→

```
SMITH,JOHNSON,WILLIAMS,JONES,BROWN,DAVIS,MILLER,WI
ALLEN,YOUNG,HERNANDEZ,KING,WRIGHT,LOPEZ,HILL,SCOTT
,MORRIS,ROGERS,REED,COOK,MORGAN,BELL,MURPHY,BAILEY
S,HENDERSON,COLEMAN,JENKINS,PERRY,POWELL,LONG,PATTI
ULLIVAN,WALLACE,WOODS,COLE,WEST,JORDAN,OWENS,REYNOI
KS,CRAWFORD,HENRY,BOYD,MASON,MORALES,KENNEDY,WARREI
```

Figure 2. Name Data

```
street_list = [
 "street",
 "st",
 "avenue",
 "ave",
 "boulevard",
 "blvd",
```

→

```
"street",
"st",
"avenue",
"ave",
"boulevard",
"blvd",
"road",
"rd",
```

→

```
street, st, avenue, ave,
ter, park, pkwy, alley,
expy, east, west, north,
acacia, acorn park, acor
ainsley, airport, alamed
aldron, alexander loop,
```

Figure 3. Location Data

Methods

- My methods are this and also this
- There were methods that were this

Results

- My methods are this and also this
- There were methods that were this