**Key Phrase Identification and Statement Clustering for FBI Witness Statements**

**Description of the Problem Statement**:   
 The Federal Bureau of Investigation (FBI) collects witness statements as part of its investigation. These statements contain valuable information crucial to solving crimes, but they are often lengthy, unstructured, and diverse in content. Analyzing these statements manually is time-consuming and may lead to overlooked key information. To streamline this process, the project aims to develop an automated system for key phrase identification and statement clustering from witness statements.

**Who is Impacted by this Project**:   
 This project will have a significant impact on various stakeholders within the FBI and the broader law enforcement community.

* Investigative Teams: Field agents and investigators will benefit from a more efficient and effective means of processing witness statements.
* Analysts: Data analysts will be able to access structured data, facilitating pattern recognition.
* Witnesses: If witnesses are aware that this system is being used, it will allow them to over explain and include potentially important details they may have left out had they been trying to keep their responses brief.
* Victims: Speeding up the investigation process will result in the victims being returned more quickly.

**System in Place**:  
 As of now, to the best of our knowledge, the processing of witness statements within the FBI is primarily manual. Field agents and analysts read through statements, identify key information, and manually categorize or extract relevant details.

**Plan to Address the Problem**:  
 We plan to train a key phrase detection model and use it to identify the key phrases in the witness statements so that the analysis can skim through the report faster and still get the main points. We also want to cluster descriptions. For instance a few reports could describe the appearance of the criminal, whereas a few could describe their car. We could try to cluster these statements together and present the information in a cohesive manner through unsupervised clustering techniques.

## 

**Named Entity Recognition**

### **Task: Identify and categorize the named entities like names, places or events from text**.

### **Dataset: Conll -** [**conll2003 · Datasets at Hugging Face**](https://huggingface.co/datasets/conll2003)

**Creating Custom Datasets:**

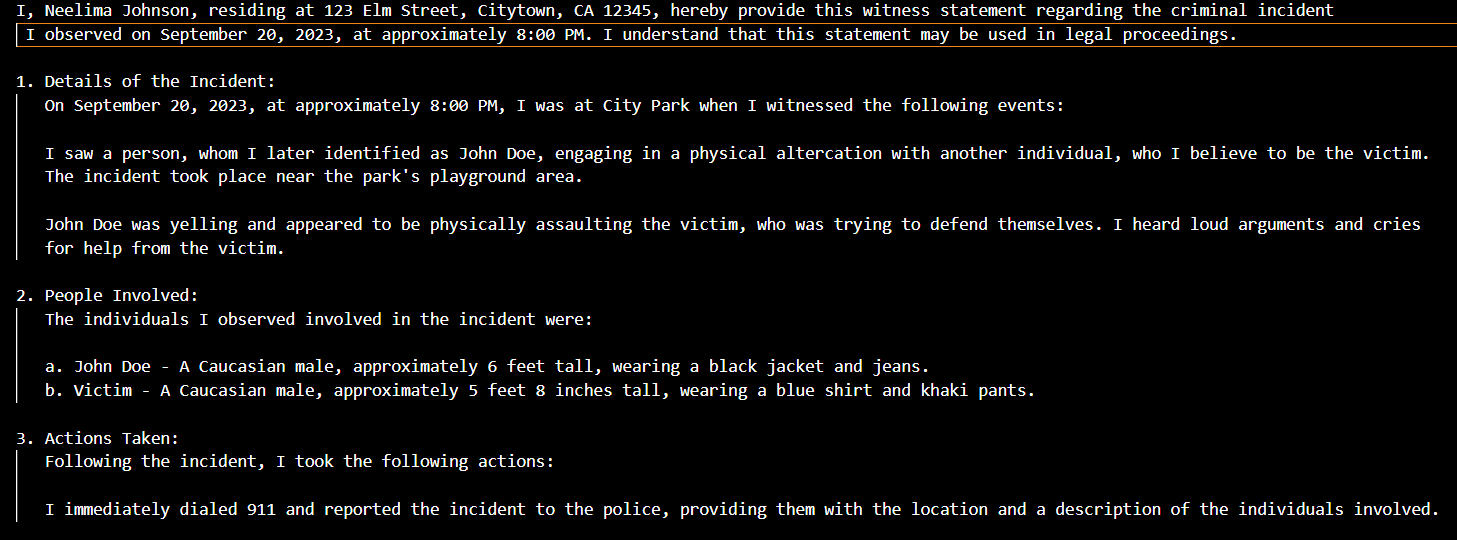
#### **Deep Active Learning as proposed in:** [**https://arxiv.org/abs/1707.05928**](https://arxiv.org/abs/1707.05928)

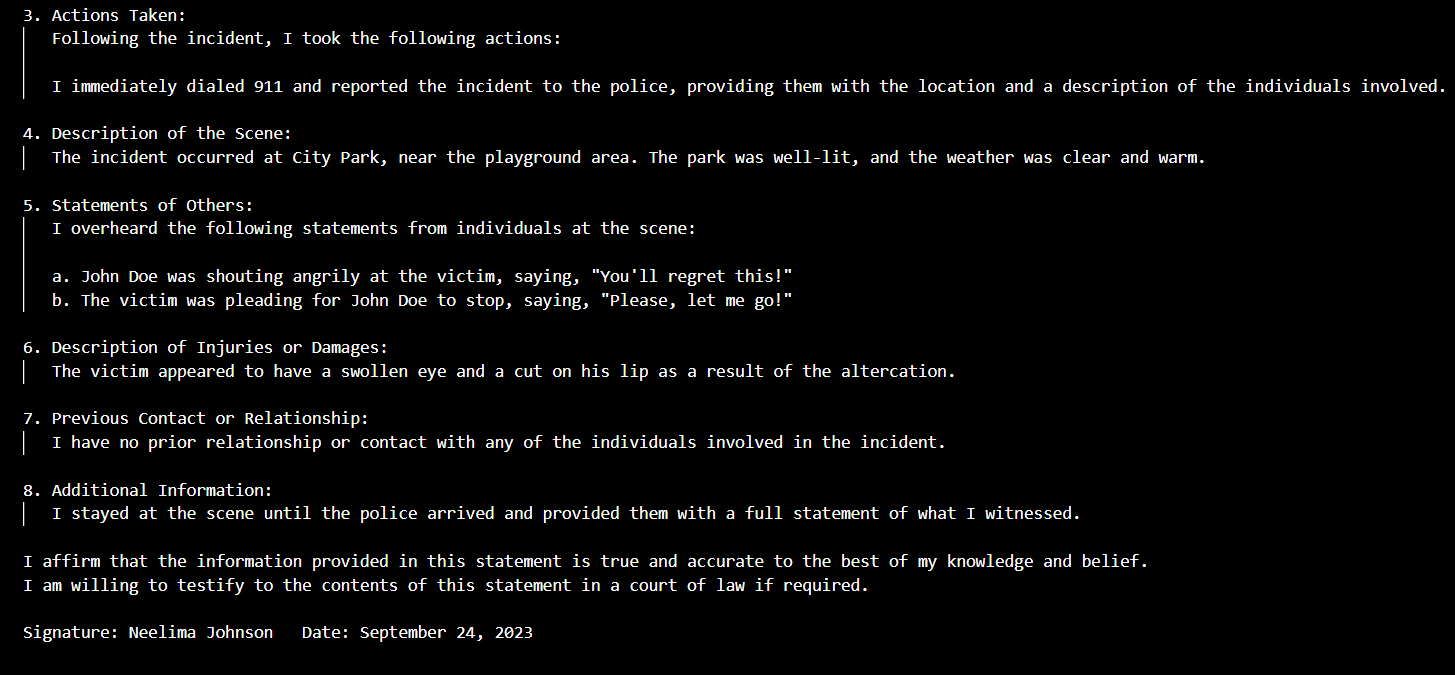
* 1. In this paper, they authors use Active learning to construct a dataset
  2. Active Learning is a technique to label more informative samples of data
  3. We annotate a small sample of data and train the model, and re annotate those samples from the inference set that the model was most unsure about.
  4. Upon adding those samples to the training set, the model learns something it did not know initially
  5. Thus we can get better quality data samples

#### **Self Training**

* 1. In self learning, we annotate a small set of samples and train a model on those samples
  2. The model runs an inference on some data, and we take those samples and retrain on the ones it was most confident on.

#### **Zero Shot Prompting**





In the provided witness statement, the named entities are as follows:

1. Neelima Johnson
2. John Doe
3. Citytown, CA
4. September 20, 2023
5. 123 Elm Street
6. City Park
7. September 24, 2023
8. 12345 (ZIP Code)
9. Phone: (555) 555-5555
10. Email: neelima@email.com

Prompt tuning techniques are required to improve the results.

The output needs to be in a format that is suitable for coding

Questions: Will we have access to GPT API? Perhaps GPT 4?

Models To be used:

* BERT
* GPT

Will we have access to Jupyter Labs for prompt tuning or quantization

* CRF(Conditional Random Field) - <https://www.mdpi.com/1099-4300/19/6/283> - Used a combination of LSTM and CRF model for Bio NER.

Annotation Guidelines to be used: BIOs tagging - (Beginning, intermediary, other)

**Key Phrase Identification:**

We can use models like KeyBERT for Key phrase analysis. If we need a new dataset to train a Key phrase model, we can use <https://huggingface.co/datasets/midas/oagkx> or create our own dataset which fits our requirements with the methods explained above.

**Clustering:**

After Named entities and key phrases are recognized, we can provide embeddings to phrases and we can implement unsupervised clustering models. If unsupervised clustering models are not giving out optimal results, we can try using topical modeling where the key phrases are associated to a certain topic and we can cluster by topics and subtopics.

**Prompting to obtain Test data:**

**Task: To obtain witness data through GPT**

Prompt Engineering Attempt:

1. <https://chat.openai.com/share/80a93559-0674-4e2d-81d2-0814cdd43de7> - Zero shot success
2. <https://chat.openai.com/share/3c71f199-abd5-4922-b47a-9db42e1407f2> - Zero shot failure
3. <https://chat.openai.com/share/3c71f199-abd5-4922-b47a-9db42e1407f2> - prompt engineering to get a report

Question: How do we make this a full-stack project?

Key Phrase Extraction: