**Hungry Hornets: Checkpoint 5**

1. Can we build a natural language processing model that allows users to ask questions directly to it, in order to retrieve information on officers who have drug and alcohol abuse and medical allegations against them? The aim is to create a tool that would allow novice users to find out more information about these individuals without coding experience. We will do this by adapting the TAPAS language model created by Google and available publicly. Due to constraints in computational power with the model itself, we will limit our analysis to officers with more than 2 allegations of drug and alcohol abuse, of whom have had at least 1 sustained allegation and have a numerical value listed for their current salary. We will measure the success of our approach by asking the following questions to the model across every single officer and testing the overall accuracy per question. Here, [X] represents each individual officer in our table.
   1. What is the race of [X]?
   2. What is the gender of [X]?
   3. What is the birth year of [X]?
   4. what is the allegation count of [X]?
   5. What is the sustained count of [X]?
   6. What is current salary of [X]?

We wanted to create a natural language model that would allow users to ask questions about officers with drug and alcohol and medical abuse allegations directly to it and return answers. We felt this was an important task as it could allow this type of information to be readily accessible to non-technical people in the public.

When we ran each of the above questions for all the officers in the table, here is the accuracy overall accuracy per question:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gender** | **Race** | **Birth Year** | **Allegation Count** | **Sustained Count** | **Current Salary** |
| 86% | 73.33% | 68.89% | 71.11% | 68.89% | 66.67% |

As can be seen with any NLP approach, how the language is phrased, contextualized and formatted within the database can influence the provided outputs. We expect that with incremental innovation, involving iterative training of this algorithm, our accuracies will continue to improve.

We also expect that as these models improve and computational power increases, we can expand the size of the table and add additional columns, to allows users to ask more types of questions to it directly and truly allow the public to identify information that is important to them.