2024 /11 /12

20:00 - 21:00

11310CS460200 Group 23 Meeting Minutes

|  |  |
| --- | --- |
| Topic | Progress Report |
| Place | Discord Voice Chat |
| Agenda | Discuss about our dataset collection and input format |
| In attendance | All present |
| Task Assigned | 游松澤: collecting starting pitcher datasets  曾柏勲: collecting closer pitcher datasets  蕭以勝, 楊立慈, 賴允中: model design |
| Next meeting | Date: 11/19  Time: 8:00 ~ 9:00 pm  Objective: Keep working on phase3 (finish training our first implementation of RNN and try to optimize it)  Location: Discord Voice Chat |
| Meeting Summary:   1. I eventually find a python package called pybaseball, which is a platform that combines statcast, Baseball Reference, Baseball Savant, and FanGraphs data. Just easily import it and we can get the required dates data and features easily.   Below is a partial screenshot of my data collection process. I selected **Shota Imanaga** as our starting pitcher and **Mason Miller** as our closer, gathering their complete data for the 2024 season. This resulted in **2,590 rows** for Imanaga and **995 rows** for Miller. In the future, we can expand by adding more features or extending the dataset to include additional years if needed.     1. We discussed and tried to build the prototype of our RNN model, maybe we can use the scraped data into training next week if everything works properly.   Because some length of pitch sequences are not the same as sequence\_length , so we want to use padding length to the sequence\_length. To ignore the padding values, we use masking layer to train based on the real value. Each time step we need to get the probability of pitch type,so return\_sequences ==True. We think we need more layers to build up some model and join more features. | |
| A group photo of the discussion session: | |