2024 /11 /05

20:00 - 21:30

11310CS460200 Group 23 Meeting Minutes

| Topic | Task Assigned and Model structure confirmation |
| --- | --- |
| 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: 11/12  Time: 9 p.m  Objective: Keep working on phase3 (naive implementation of our designed model)  Location: Discord Voice Chat |
| Meeting Summary:   1. We discuss what actually the model input and output is, and also think about the first (easier) version of the dataset csv columns.   This is our concept thoughts of csv:   | Date | Pitcher Name | Batter Name | 此打席第幾顆球 | Pitch-type | isStrike | | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |   We may add feature like whether this ball isHit (since it may be a foul ball but it will count in isStrike) or weather and so on, but this is for advanced part if we have time to do it. (Since we find out that actually the collection of pitch by pitch dataset is not as easy as we think since we can’t find the raw data as detailed as this in statcast. The game log only provides the ball that has result (out or been a hit) as below)    But we can find the detail data in MLB each game like this:    However, we need to first find what date this pitcher are pitching. Then, we need to click this game and click every batter in every inning to get this page. This is kinda not systemic and kinda hard to do web crawler. Hence, we aim to find api or other website that have better dataset to collect.   1. We also discussed how we should train our model. Our original thought was that each person should train a model on their own for a pitcher of their choice (possibly of different pitcher types). However, we realized that this is pretty ineffective since everyone will have to do similar work; Also, it fails to apply to a more general setting. We decide that we can instead simply add the pitcher’s ID as a feature and let the model figure it out in the training process. This enables us to train our model on a lot more data.   Additionally, we decided to write the same RNN code and train it on a starting pitcher and a closer. We hope to be able to compare their results and gain insights to what factors might affect their accuracy difference.   1. We decided to divide our work into 3 groups: 2 people on processing the collected data to fit our model’s usage, 1 person on sketching the model’s rough backbone, and 2 person on implementing the details of the sketch code. This is because we’ve been busy preparing for midterms of other subjects, so we plan to execute phase 2 & 3 of our timeline in parallel. | |
| A group photo of the discussion session: | |