Jonathan Greenberg

3/29/2020

Project Report #8

I am no longer waiting on crucial data to be sponsored by YU that will help in my predictions, because I have found a way to access the data that I need for my modeling aspect of the project. It is called basketball\_reference\_web\_scraper and the library is found here: https://pypi.org/project/basketball-reference-web-scraper/

I have figured out a way to get each player’s identifier which I can do a fast lookup of their stats with. Now all I need to do is iterate through my dataframe and fill out the columns with appropriate historical game data, or games prior to their injury, and then I can begin my modeling.

The other three new columns in my dataframe called: “body”, “injury”, and “severity” along with the game statistics for 10 days prior to the player’s injury, will help in my predictions and classifications.

I also want to get all the physical attributes (height, weight, age, etc.) of each player to help increase the number of features for my model, and I have found a dataset on kaggle that can supply this for me: <https://www.kaggle.com/drgilermo/nba-players-stats>

I would also like to predict a binary classification of injury or not after a 10 day span of data, so i would like to add just as many rows of players that did not get injured and have a binary column called “is\_injured” and try a binary classification prediction model.

All of my work thus far can be seen in this colab:

<https://colab.research.google.com/drive/1InZ1b1YDH97RYYncTBs5TKvWcNoX9PPr>

I am looking for other ways to do clustering with NLP and make word embeddings to see which players and teams have the most similar injury types. I can make use of the BERT link that you sent me and try to follow that.