**NLP FINAL PROJECT - MILESTONE 2**

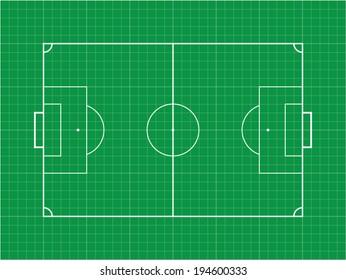
**Neython Lec Streitz and Jordan Jung**

**Link to code is shared on Google Colab and included here:**

[**https://colab.research.google.com/drive/1gCNhT5SylcSI8AyWR3t7vtDJeBBg-kGm?usp=sharing**](https://colab.research.google.com/drive/1gCNhT5SylcSI8AyWR3t7vtDJeBBg-kGm?usp=sharing)

**Work Completed So Far:**

* Data is uploaded from the open source github. We get data for all events, player metadata, event metadata, and match metadata. Unsure if we will use the event or match metadata but the player metadata will be useful when creating the Player vectors.
* Event data is turned into a corpus and vocabulary ready for Word2Vec. The only ‘real’ preprocessing we do on the events is lower casing them and then tokenize them. We tokenize each event into a string built by the action name and location on the pitch. We would like to include extra information for the actions (like type of shot, left foot/right foot, where it ended up), but still need to figure out how to represent it well across different kinds of actions.
* Location is represented by a 6x4 grid like below. Upper left hand corner of the field would be 1/6 , 1/4 . That is, box 1 of 6 horizontally, and box 1 of 4 vertically.



**Needs to be completed:**

* Next, we need to implement our actual Word2Vec and Doc2Vec models. The Word2Vec will be applied to our corpus of actions. In this way, we will know what actions are similar to each other. To test the metrics of this model, I think it will be easy to print similar actions and just see if they make sense by hand. For example, if we get that a pass on the ground in the middle of the field is similar to a pass in the air in the middle of the field, then our model is likely working. Doc2Vec will be applied to collections of actions performed by the same player. In this way, the players can finally be compared to each other.
* Once the model is completed we need to generate a handful of similar players on the men’s and women’s team as the ultimate goal of our model.

**Metrics To Consider:**

* Because both of us play soccer, and also know a lot of people who are avid fans of soccer, we will be able to measure our model with our own domain knowledge. As you suggested, finding articles about specific players and then switching out the names of the players with the most similar ones found by our model will be a great way to test how well it finds similarity. To note is that we don’t know many people who are avid fans of women’s soccer, so testing to make sure it makes sense might have to be done not by replacing the man’s name with their women’s counterpart, but the other way around. If we can find articles about women soccer players and replace them with men who are similar, we can see if the description still matches how the man plays (which most of the people we ask to read them will know more about/have seen play). We might also be able to compare performance the more additional information we give to each event. How does the model change when actions are just the name and locations versus name, location, and additional information?