**Natural Language Processing Final Project Milestone #1**

**Group:**

Neython Lec Streitz, Jordan Jung

**Main Question:**

What professional women’s soccer player is most similar to a given men’s player?

**Motivation:**

While this might seem like a strange or unimportant question, our motivation for exploring this question revolves around wanting to boost enthusiasm for women’s soccer. Historically, women’s soccer was banned around the world by FIFA, the international soccer association, until 1752. The ban was in place specifically because men’s soccer organizations were threatened by competition from an emerging popularity in women’s soccer. As a result, women’s soccer has always had to operate in the shadow of men’s soccer.

Our hope is to create a model that takes open source data from STATSBOMB, international leaders in data analytics for soccer, and embeds data into words that describe actions. From there, we want to represent players as a “bag of actions” so that they are composed of action vectors. The end goal then, is to be able to use the player vectors to find similarities and analogies between players. In this way, we can take a popular men’s soccer player and find a women’s soccer player that is similar to them in the kind of actions they perform during a game. For instance, inputting Uruguayan player Luis Suarez might return Dutch player Vivianne Miedema. Both players are prolific goal scorers for their respective clubs and national teams. Yet, Suarez has 16.8 million Twitter followers compared to Miedema’s 88.9 thousand. Ultimately, our model could be a gateway for people who enjoy men’s soccer to get into women’s soccer.

**Existing Research:**

Natural language processing is quickly gaining traction in the soccer analytics community. A lot of the work currently published revolves around talent identification and enhancing the fan experience. For example, in a report by Buitelaar et al., they generate a knowledge base extracted from the FIFA and UEFA websites. With this knowledge base, they created automatically inserted hyperlinks for teams and players so that you can learn more about that specific entity. Similarly, work by van der Lee et al., takes data from Dutch soccer games and generates corpus-based texts that are geared towards a specific fan base through tone and other small linguistic features. In a paper by Tanaka et al., they present an automatic commentary system for soccer matches. Comparably, Taniguchi et al., also describe a commentary system that takes event data and turns it into live commentary. While these four papers explore natural language processing and soccer, they do not use their models to explore similarities between players. In this way, our project is unique.

In one of the closer examples to our own research question, work by Zhong et al., used Word2Vec processing on event data in order to analyze the strategies of RoboCup matches and using the similarity of teams and players, generate strategies to combat other teams. In their work, they focused only on passes and dribbles.We hope to use all of the data we can from the StatsBomb open source database, including different kinds of passes (chipped, inswinging cross, outside the foot cross, etc.), shots (finesse, driven, etc.), dribbles, and saves just to name a few. Another important differentiator is that the RoboCup data they used was only in 2D. Our data set is represented in 3D, which will add complexity to our models. Nevertheless, their use of Word2Vec and clustering of players are techniques we are planning on implementing as well.

**Data:**

For our data, we plan on using the StatsBomb open source repository. This repository includes around 900 different matches between 2003-2020 with over 4,000 men’s and women’s players. Each event document includes team and competition metadata which is somewhat useful for labeling purposes, but also labeled event data of player actions. In total, there are about 3 million actions in the dataset, described by their action type, location on the field, player, time, and optional information like success or body part used. Because StatsBomb is a leader in the soccer data analytics world, and the StatsBomb data set is used widely in data analytics research, we feel confident that it is an accurate data set to pull event data from. The complete data set can be found here: <https://github.com/statsbomb/open-data>. We’ve included 50 matches with event data as our sample, but we have access to all of them already.

**Citations:**

Buitelaar *et al*., "Generating and visualizing a soccer knowledge base." In *Demonstrations*. 2006.

Tanaka *et al*., "MIKE: An automatic commentary system for soccer." In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*, pp. 285-292, 1998.

Y. Taniguchi, Y. Feng, H. Takamura, and M. Okumura, “Generating Live Soccer-Match Commentary from Play Data”, *AAAI*, vol. 33, no. 01, pp. 7096-7103, Jul. 2019.

van der Lee *et al*., "PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences." In *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 95-104, 2017.

Zhong *et al*.,"A study on the analysis of soccer games using distributed representation of actions and players." *ICIC Express Letters* vol. 13, no. 4, pp. 303-310, 2019.