NBA PERFORMANCE ANALYSIS

David Rafanan & Cory Dickson

1) What did you propose to do? What is the motivation/background?

Many people follow the NBA; always on the lookout for great basketball performances from players as well as highlight reels, amongst others. Aside from watching actual games, they use other mediums (i.e. ESPN, NBA website, etc.) to keep up to date with the latest games. Through these other means, they can see the stats of the players as well as live updates whether it be a foul committed from one player or a made shot from the other; these stats can illuminate the many different outcomes in the game of basketball. However despite all of these, these alternatives are often shown to be inaccurate in comparison to actually watching the game. Maybe there is a way to deduce more information about the game based on the people who watch it. Through live tweets during the game, we are hoping to find ways of describing the game better, or if at all. There is a way how to analyze player behavior based on keywords.. In twitter, you can keep updated with the game from the latest tweets which will describe the different kinds of aspects of the game not seen in the aforementioned. Furthermore we have a working theory that players who are mentioned more may be more efficient or was involved in some playmaking aspect of the game. We are going to use sentiment analysis on tweets to determine if their is a correlation between mentions, and a players’ efficiency rating. Furthermore, we are interested in finding a correlation between player performance and the overall negative or positive sentiment in given statements. For example if people make mention of a players shooting and it’s positive we want to see if they are performing better or worse than average.As for background we had to learn several R libraries such as quite obviously the twitter API, ngrams, NLP, to deal with the text processing and classification.

2) Explain the data you used and model in detail.

Overall the data consists of tweets that contain unigrams, bigrams and trigrams of interest. For example unigrams are considered to be player names, bigrams for actions and trigrams that contain name and action. This assumption was used to find tweets in where a user made specific mention of a player action or attribute. We had to denoise the data of hashtags and retweet headings as well as emoji’s. The text was then converted into lower case to make the parsing easier. Furthermore the tweets came with temporal data, the date and time published hence allowing us to filter the tweets received over a time interval which was during the game. We used the assumption that tweets sent out during the live game were talking about the game, this was the best heuristic we could come up with. The model was to use ngrams to depict the most freqeunt basketball terms and players mentioned, used for sentiment analysis

3) What did you end up doing?

We ended up following the proposed process rather closely. We retrieved the tweets from five Lakers games and two Oklahoma City Thunder games so we would have an adequate sized dataset. Each of the games with exception of the Oklahoma games (which had 1000-1500) had about 200-300 tweets. This was due to the limit that twitter imposes on tweets that it can retrieve after some period of time. Therefore we tried to do the most recent games to retrieve the most amount of tweets. We sanitized the dataset as described above then converted it into a corpus with a frequency table of words used. We discounted the hashtags we used to retrieve the tweets as it would obviously be the most mentioned thing. For example for the Lakers we removed “Los Angeles Lakers” and “OKCThunder” where not used in the analysis of the tweets. We then retrieved the stats, including the PER of the players of each game. We also put together a list of basketball jargon to search the tweets for in order to find relevant tweets or phrases to analyze. The model was to use sentiment analysis to compare the words in the tweets to two basketball corpuses (Good and bad) and determine a score between 1-10 by subtracting the positive and negative.

4) What if anything did you change about your approach and why?

As for the data collection we did as what was described in the project proposal, the only main difference is instead of focusing solely on the Lakers we chose the Oklahoma City Thunder as they have been having a better season than the former team. We were worried that the string of losses that the Lakers have been experiencing would cause the tweets to be primarily negative as the loss would skew any positive plays made by any given player.

5) What visualization(s) have you included? Explain what is conveyed in the visualization and why.

We chose to use wordclouds to be the best way to display the data as it provides a topological view of the frequencies. This also allowed us to see which jargon is primarily being used during the game. For example when the Lakers played the Jazz, rebounds was the most common talked about action, so this helped us approach the analysis for each game by giving us a sense of what kind of observations were being made during the game.

‘’’

#display settings

par(mfrow=c(1,1))

col=brewer.pal(6,"Dark2")

#Wordcloud: LakersVJazz

set.seed(2245)

wordcloud(corpusLJ, min.freq=25,rot.per = 0.35, random.color=F, max.word=45, random.order=F,colors=col)

#Wordcloud: LakersBulls

set.seed(1550)

corpusLB=tm\_map(corpusLB,PlainTextDocument)

wordcloud(corpusLB, min.freq=25,rot.per = 0.35, random.color=F, max.word=45, random.order=F,colors=col)

#Oklahoma Port

set.seed(4211)

corpusTP=tm\_map(corpusTP,PlainTextDocument)

wordcloud(corpusTP, min.freq=25,rot.per = 0.35, random.color=F, max.word=45, random.order=F,colors=col)

#Oklahoma Celtics

set.seed(3011)

corpusTC=tm\_map(corpusTC,PlainTextDocument)

wordcloud(corpusTC, min.freq=25,rot.per = 0.35, random.color=F, max.word=45, random.order=F,colors=col)

’’

6) What evaluation method did you propose?

We chose to use the Naive Bayes Algorithm in order to classify text in positive or negative sentiment. For this specifically, we classified basketball terms and plays into positive or negative connotation into corpuses. This was the basis on evaluating the tweets from live games, whether a player had a positive or negative impact based on the words that were said while tweeted.

7) How did your model perform according to this evaluation?

We have recorded the games the Lakers have played from 12/3/2016 - 12/11/2016, which were a total of five games. We anticipated the naive bayes performance analysis to work, but what resulted was a lot of noise. This was due to the injuries that occurred during the early games of the timespan; leading to many twitter users to tweets about that instead of the game. In addition, all of the games in this span were losses, so what we obtained in addition to the noise from other events were negative sentiment for the most part.

8) Based on your results what conclusions do you draw?

Unless there were unexpected circumstances (i.e. injury, technical foul, fights, etc), the model would work as expected. As stated above, a backup team (Oklahoma City Thunder) was recorded for the sake of testing the model. The reason why the that team was chosen for testing is due to a greater win-loss ratio and no injuries. As a result, the model performed better with this team than the Lakers, so it will work better on those teams with better records.

9) Based on your results what further studies would you do or are warranted?

Due to the noise we’ve received from the injuries, we are planning to incorporate what is considered noise (injuries, technical/flagrant fouls, brawls, etc) into additional factors to determine performance analysis.