CS 579 PRESENTATION

NBA player salary predication and rationality analysis

Group Member:
Hang Li
Lu Wang

THE IDEA

Goal:

we try to tell whether the current salary for a certain NBA player is to high or too low.

Approach:

We use Twitter to collect tweets that mentioned a certain player
Compare these tweets to the tweets that for other player
Select the most similar players by comparing their tweets
Return the average salary of the most similar players
Compare the average salary to this player's salary

DATA COLLECTION

Before we started:

1. Build python environment
2. Install mongoDB in your machine
3. Build your own config file
4. Make everything works!

DATA COLLECTION

Build python environment:

1. Clone the repository

```
$git clone https://bitbucket.org/gatesice/iit-cs579-project.git
```

2. Install the library. Add `sudo` at the beginning if necessary.

```
$cd iit-cs579-project
$pip install -r requirements.txt
```

3. (optional) You can use `virtualenv` for isolated python environment.

DATA COLLECTION

Install mongoDB in your machine:

1. Download mongoDB at mongodb.org
2. Extract the package to where you like
3. You can add mongoDB/bin to your PATH

DATA COLLECTION

Build your own config file:

1. Make a copy of `iit-cs579-project/collector/config.sample.py` and rename it to `config.py` in the same folder:

```
$cd iit-cs579-project/collector
$cp config.sample.py config.py
```

2. Modify `config.py`, add your Twitter API. Feel free to create multiple subclasses that implements `Config` class.

```
class YourOwnConfig(Config):
    CONSUMER_KEY = ‘YOUR_CONSUMER_KEY'
    CONSUMER_SECRET = ‘YOUR_CONSUMER_SECRET'
    ACCESS_TOKEN = ‘YOUR_ACCESS_TOKEN'
    ACCESS_TOKEN_SECRET = ‘YOUR_TOKEN_SECRET'

    MONGODB_IP = ‘192.168.1.10'
    MONGODB_PORT = 2400
```

DATA COLLECTION

Make everything works:

1. Start mongod on your machine which installed mongoDB.

```
$mkdir –p /path/to/your/db
$mongod --dbpath /path/to/your/db --bind_ip 192.168.1.10 --port 2400
```

2. On the machine that contains the collector, start collector to collect data

```
$cd iit-cs579-project
$python –m collector “Kobe Bryant” --config YourOwnConfig
```

3. When you meet any issue (e.g: network issue). Just restart the program and it will continue the search from current progress.

DATA COLLECTION

Module structure:

```
collector/
    __init__.py
    __main__.py
    config.sample.py
    config.py (you create your own config file!)
    models.py
    timing.py
    twitter_wrapper.py
```

DATA ANALYSIS

Build up features:

How many people have POSITIVE opinion toward the player
How many people have NEGATIVE opinion toward the player
How many people have NEUTRAL opinion toward the player

How many media/group have POSITIVE opinion toward the player
How many media/group have NEGATIVE opinion toward the player
How many media/group have NEUTRAL opinion toward the player

DATA ANALYSIS

Build up features:

Use word analysis, go through the database, determine the type of the feature for each tweets

After collecting feature data, return a feature matrix containing all feature information of all players

DATA ANALYSIS

Search for similar player:

Compare each feature for each player, we have 6 features, so that we have 6 rating attributes for each player, each player will get a rate for each feature based on how similar they are when comparing to the selected player. The rating are assigned from 0 to the number of the players in the database.
Adding up the ratings and rank then by order, we will able to get the most similar player comparing to the selected one.

```
['Kobe Bryant', 23.5, 29382, 10730, 15867, 79012, 41911, 20184, 0]
['Lebron James', 20.6, 21772, 10935, 12514, 37523, 19021, 17127, 11]
['Derrick Rose', 18.9, 21698, 6322, 16849, 28781, 18191, 9679, 20]
['Kevin Durant', 20.0, 6746, 7851, 3890, 68849, 53836, 12034, 23]
['Chris Paul', 20.1, 7025, 3130, 3791, 25013, 14937, 6137, 37]
['Dwight Howard', 21.4, 3930, 7016, 2406, 12225, 7219, 12377, 40]
['Paul George', 15.8, 4279, 1712, 2673, 10412, 6499, 3003, 51]
['Tim Duncan', 10.4, 10819, 1718, 8895, 4992, 3143, 2466, 55]
['Blake Griffin', 17.6, 4812, 1178, 3346, 9931, 4789, 2408, 56]
['Eric Bledsoe', 13.0, 5111, 1024, 5942, 6032, 4393, 1435, 65]
['DeMarcus Cousins', 13.7, 3906, 1185, 2594, 6327, 3335, 1991, 69]
['Nene', 13.0, 14737, 8438, 4414, 161746, 31019, 31017, 76]
['Marc Gasol', 15.8, 3182, 748, 1942, 7785, 4684, 1336, 78]
['Ty Lawson', 11.6, 1430, 5643, 633, 5164, 2758, 10054, 83]
['Jeremy Lin', 14.9, 1860, 1043, 1093, 5094, 3059, 2210, 89]
['Kevin Love', 15.7, 21341, 984, 13828, 2156, 799, 1443, 94]
['Carmelo Anthony', 22.5, 1354, 1568, 573, 5285, 2343, 3392, 94]
['Chris Bosh', 20.6, 2891, 546, 1909, 3422, 1858, 901, 106]
['DeAndre Jordan', 11.4, 2023, 445, 1160, 2566, 1352, 719, 130]
['Rajon Rondo', 12.9, 1441, 401, 523, 2853, 1340, 1114, 139]
['Dwyane Wade', 15.0, 1200, 545, 521, 2616, 1045, 1158, 141]
['Josh Smith', 14.0, 860, 506, 901, 2232, 1410, 584, 143]
['Rudy Gay', 19.3, 1067, 513, 650, 2116, 1035, 832, 151]
['JaVale McGee', 11.3, 2214, 180, 1921, 1867, 1181, 496, 155]
['Chandler Parsons', 14.7, 972, 412, 702, 1989, 1576, 494, 156]
['Kyle Lowry', 12.0, 1056, 364, 622, 2371, 1360, 476, 161]
['Eric Gordon', 14.9, 354, 1145, 305, 1419, 1043, 1825, 164]
['Tony Parker', 12.5, 1187, 222, 535, 3023, 1222, 357, 167]
['Joakim Noah', 12.7, 743, 703, 423, 1738, 834, 922, 169]
['Deron Williams', 19.8, 473, 371, 231, 2618, 1371, 759, 171]
```

DATA ANALYSIS

Reason for outliners

Players that have the same name with other famous person

Players which are famous, and would like to win champions and don't care about their salary that much

Players which are not famous, but will help their team to win a lot, and the club would like to pay them a lot

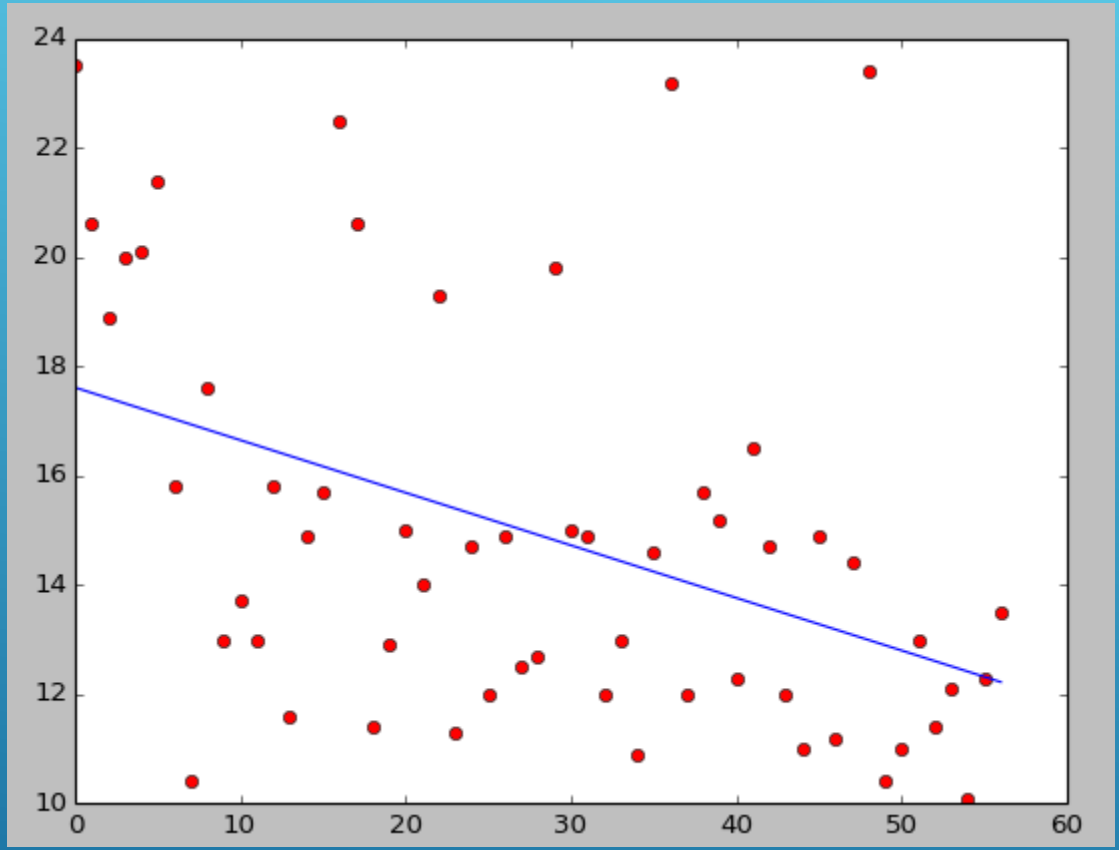
DATA ANALYSIS

Remove outliners

choose the highest/lowest paid players

draw the trend line for similar players, x-axis for player count, y-axis for salary

pick the points that are far away from the line, count the appearance time for the player(point). Remove top several player than in the outliners list.


24
22
20
18
16
14
12
10
0
10
20
30
40
50
60

DATA ANALYSIS

Predict

enter a player name, list similar list, choose the top 5 similar player, return the average salary of those player

Predict average

try to predict the salary for all the players that are not outliners.

RESULTS

player	predicted	actual	accuracy
Rajon Rondo	14.34	12.90	0.89
Gerald Wallace	11.46	10.10	0.87
Brook Lopez	13.34	15.70	0.85
Kevin Garnett	13.98	12.00	0.83
AI Jefferson	11.88	13.50	0.88
Derrick Rose	18.60	18.90	0.98
Joakim Noah	14.70	12.70	0.84
Lebron James	18.60	20.60	0.90
Kevin Love	14.82	15.70	0.94
Chandler Parsons	13.54	14.70	0.92
Tyson Chandler	13.46	14.60	0.92
JaVale McGee	13.30	11.30	0.82
Danilo Gallinari	14.56	10.90	0.66
Josh Smith	15.00	14.00	0.93
David Lee	14.60	15.00	0.97
Andrew Bogut	13.70	13.00	0.95
Andre Iguodala	11.66	12.30	0.95
Paul George	15.56	15.80	0.98
Roy Hibbert	14.48	14.90	0.97
David West	13.98	12.00	0.83
Chris Paul	16.04	20.10	0.80
Blake Griffin	15.18	17.60	0.86
DeAndre Jordan	13.92	11.40	0.78
Jeremy Lin	13.98	14.90	0.94
Zach Randolph	12.76	16.50	0.77
Marc Gasol	14.30	15.80	0.91
Dwyane Wade	14.48	15.00	0.97
Larry Sanders	11.60	11.00	0.95
Nikola Pekovic	12.06	12.10	1.00
Omer Asik	12.54	14.90	0.84
Eric Gordon	13.60	14.90	0.91
Tyreke Evans	11.60	11.20	0.96
Jrue Holiday	11.90	11.00	0.92
Andrea Bargnani	12.70	12.00	0.94
Kevin Durant	17.44	20.00	0.87
Russell Westbrook	13.80	15.70	0.88
Serge Ibaka	13.14	12.30	0.93
Eric Bledsoe	15.00	13.00	0.85
LaMarcus Aldridge	13.98	15.20	0.92
Brandon Roy	11.60	14.40	0.81
Nicolas Batum	12.06	11.40	0.94
Rudy Gay	15.64	19.30	0.81
DeMarcus Cousins	15.54	13.70	0.87
Tony Parker	12.70	12.50	0.98
Kyle Lowry	13.64	12.00	0.86
Gordon Hayward	12.72	14.70	0.87
Derrick Favors	11.40	13.00	0.88
Marcin Gortat	11.42	10.40	0.90

('average accuraccy is', 0.8919929416436814)

CONCLUSION

1. The salary of each player have strongly relationship with the tweets that collected for those players.

2. The predicted salary will trend to be closer to major group, which means the prediction will not be that accurate when dealing with most popular players and most unpopular players. However, increase the data will help to improve the prediction a lot.

3. The most important feature for a player's salary is how popular they are, or how often their name appeared in people's conversation.

4. Based on the predicted salary, we can say that the current salary for a certain player is too high or too low, or maybe just OK.

FUTURE WORK

1. We only collected the players that have a salary greater than 10 million. When considering lower salary players, they will have a greater change to have a same name with someone that is more famous than him. Thus we may not able to get the expected tweets. So one of the future work is to try to get the tweets that we are looking for.

2. More data means more accuracy, if we can collect more expected data, it will increase the prediction accuracy a lot.

3. By follow this predicting pattern, we can also predict the players that transfer from one team to the other. And predict their salary if they change their teams.

THANKS FOR COMING