APACHE PIG ASSIGNMENT

Download the data files from :

http://www.utdallas.edu/~axn112530/cs6350/Pig/data.zip

It contains a set of data files that you will use in this classroom assignment.

The commands to load them are as follows:

```
divs = load 'NYSE_dividends' as (exchange, symbol, date, dividends);
prices = load 'NYSE_daily' as (exchange, symbol, date, open, high, low, close, volume, adj_close);
bball = load 'baseball' as (name:chararray, team:chararray, position:bag{t:(p:chararray)}, bat:map[]);
crawl = load 'webcrawl' as (url, pageid);
```

1. Using the file NYSE_daily, compute the average of each symbol's close price.

2. Using the file NYSE_dividends file to compute the max value of each stock's dividends.

3. Join the NYSE_daily and NYSE_dividends file on the symbol key, and generate the following columns for each symbol: average opening price and average dividends.

4. Flatten the baseball dataset as follows:

```
players = load 'baseball' as (name:chararray, team:chararray, position:bag{t:(p:chararray)}, bat:map[]);
pos = foreach players generate name, flatten(position) as position;
bypos = group pos by position;
```

After this, generate the count of players that have played at each position i.e. generate something like:

```
Catcher, 100
Designated_Hitter, 20
```

5. Using co-group operator, find those stocks that have never paid a dividend. This requires you to use the NYSE_daily and NYSE_dividends files.

6. Using the webcrawl file, flatten the dataset to find out the number of outlinks and inlinks for each of the pages.

The output should be something like:

Page	Number of outlinks	Number of inlinks
http://….	6	10

7. Using the output of question number 6 above (or otherwise), compute the pagerank of every page. If you use any external source as reference, please cite it clearly.

8. Using the baseball file, find the top 5 teams with the highest count of players.