

## Project Overview

We want to test a hypothesis from the perspective of behavioural finance, that the social media have a noticeable impact on individuals' decisions, thus, building a direct correlation between "public sentiment" and "market sentiment". Specifically, we hope to discover certain pattern between the stock price of Regal Entertainment Group (REG) and related Twitter discussion data of new movie releases. REG operates the largest movie theatre chain in USA. We assume that the more people discuss recent new movies on Twitter, the more they will go to cinemas to watch, hence, the investors hope they will bring more revenues to REG.

## Approach to Extract Data

### 1. List of movies released in April 2015, US

<http://www.movieinsider.com/movies/march/2015>

Movie	Release Date (all on Friday)	BO Revenue USD (M)
Furious 7	April 3	1.516B
The longest ride	April 10	62.94
Unfriended	April 17	64.1
Child 44	April 17	13
Monkey Kindom	April 17	17.1
Paul Blart: Mall Cop 2	April 17	107.6
The age of adaline	April 24	57.7
Little boy	April 24	17.4
The water diviner	April 24	30.8

We only choose the national wide released movies as there were many movies or dramas released in limited places in US.

From the above list, we further limited the list of movies to:

### **Furious 7, The longest Ride, Paul Blart: Mall Cop 2, and The age of adaline**

they were across almost the whole month of April and had much more BO revenue then the rest.

### 2. Define Tweets keyword

For each movie, we choose to use the movie title, official movie account in Twitter, and some hash tag to scan all the relevant data.

e.g. for Furious 7, keywords are: Furious 7, @FastFurious, #furious7

### 3. Regal Entertainment stock price

Date	Open	High	Low	Close	Volume	Adj Close*
1 May 2015	22.03	22.60	21.88	22.13	1,512,500	21.15
30 Apr 2015	22.25	22.46	21.90	22.00	1,584,600	21.03
29 Apr 2015	22.23	22.49	22.09	22.33	888,000	21.34
28 Apr 2015	22.27	22.41	22.07	22.29	672,700	21.31
27 Apr 2015	22.69	22.72	22.31	22.33	450,000	21.34
24 Apr 2015	22.37	22.73	22.24	22.59	797,400	21.59
23 Apr 2015	22.02	22.37	21.89	22.27	572,600	21.29
22 Apr 2015	22.09	22.16	21.77	22.00	651,900	21.03
21 Apr 2015	22.37	22.43	22.03	22.12	1,136,600	21.14
20 Apr 2015	22.25	22.48	22.22	22.29	547,500	21.31
17 Apr 2015	22.36	22.37	22.04	22.19	914,900	21.21
16 Apr 2015	22.22	22.43	22.11	22.37	517,800	21.38
15 Apr 2015	22.40	22.52	22.24	22.25	494,500	21.27
14 Apr 2015	22.32	22.48	22.12	22.34	703,600	21.35
13 Apr 2015	22.32	22.59	22.18	22.32	931,600	21.33
10 Apr 2015	22.79	23.11	22.78	22.79	395,600	21.78
9 Apr 2015	22.85	22.95	22.47	22.79	783,800	21.78
8 Apr 2015	22.72	23.04	22.55	22.96	993,400	21.95
7 Apr 2015	23.55	23.56	23.17	23.23	791,000	22.20
6 Apr 2015	23.29	23.67	23.10	23.60	1,379,100	22.56
2 Apr 2015	22.89	23.36	22.80	23.26	713,000	22.23
1 Apr 2015	22.80	22.92	22.36	22.87	996,100	21.86

#### 4. Scan Tweets by keyword

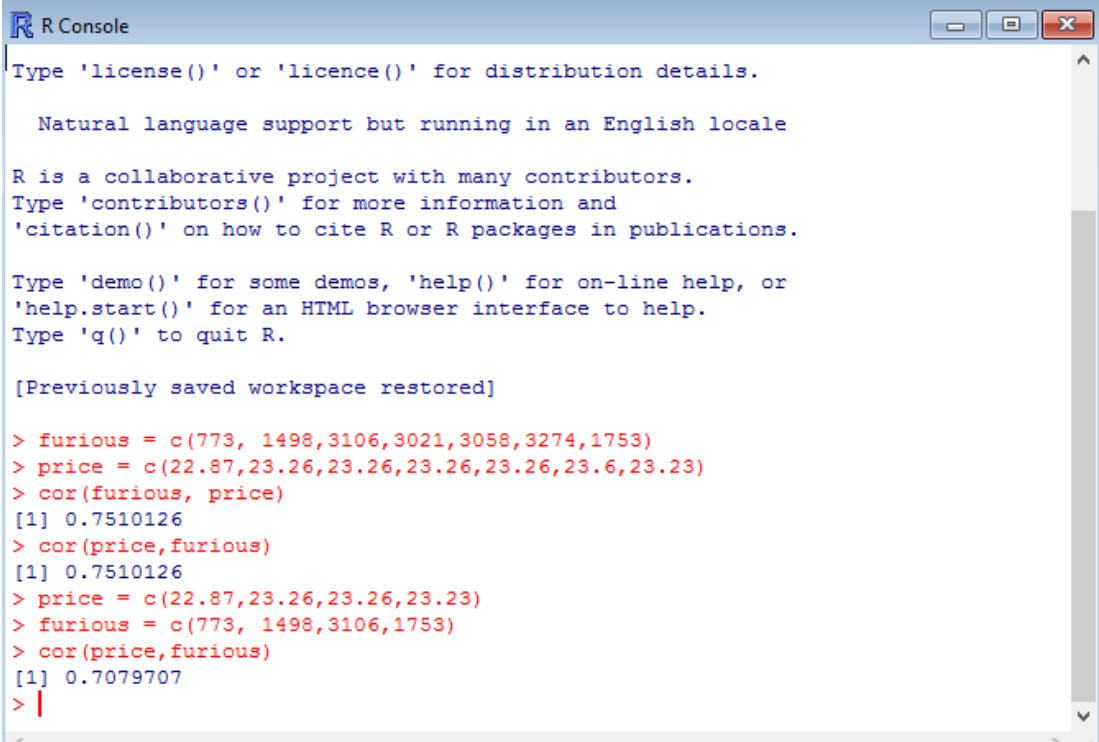
1. We first clean up the whole April tweets data from the online archive by removing those unnecessary attributes, and limited the data size to 20GB instead of 160GB.
2. We use Spark code to filter tweets data for each movie, and load them to MongoDB.
3. We use JS scripts to get each movie tweets data for each date as follows:

#### Target analysis date and Movie table:

Movie	Analysis period
Furious 7	April 1 - 7
The longest ride	April 8 - 14
Paul Blart: Mall Cop 2	April 15 -21
The age of adaline	April 22 - 28

## 5. Initial Result

<b>Furious 7</b>	Stock price	Tweets Count	Timestamp_ms UTC + 0
April 1	22.87	773	1427846400000
2	23.26	1498	1427932800000
3	Close	3106	1428019200000
4	Close	3021	1428105600000
5	close	3058	1428192000000
6	23.6	3274	1428278400000
7	23.23	1753	1428364800000
<b>The longest ride, thelongestride</b>			
8	22.96	200	1428451200000
9	22.79	170	1428537600000
10	22.79	268	1428624000000
11	Close	399	1428710400000
12	Close	285	1428796800000
13	22.32	161	1428883200000
14	22.34	92	1428969600000
"Paul Blart: Mall Cop 2", "BlartRidesAgain"			
15	22.25	466	
16	22.37	454	
17	22.19	1056	
18	Close	1273	
19	Close	747	
20	22.29	365	
21	22.12	171	



```
R Console
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> furious = c(773, 1498, 3106, 3021, 3058, 3274, 1753)
> price = c(22.87, 23.26, 23.26, 23.26, 23.26, 23.6, 23.23)
> cor(furious, price)
[1] 0.7510126
> cor(price, furious)
[1] 0.7510126
> price = c(22.87, 23.26, 23.26, 23.23)
> furious = c(773, 1498, 3106, 1753)
> cor(price, furious)
[1] 0.7079707
> |
```

Our initial simple correlation calculation is that:

During the target period, the stock price of REG is correlated to the movie tweets as high as 0.71 for Furious 7, which is a very hot movie.

When the tweets amount reached highest on April 6, the stock price was highest in the period.

For longest ride, the correlation is 0.68.

For Paul Blart: Mall Cop 2 the correlation is -0.1, which means not correlated.

There were not cooperate events and stock split or dividend during this period.

However, we don't have much background in finance and statistic analysis, we want to know how to conduct in-depth analysis of our topic.