Using Social Media To Predict Stock Prices

Florida International University

Thomaz L. Santana

April 26, 2013

About

- Mining the Web
- Machine Learning
- Using Social Media to Predict Stock Prices

Web Scraping

- Computer program that extracts information from websites
- Also called a bot, web crawler, or spiders

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml" class="no-js" lang="en">
<!--<![endif]-->
<head>
    <meta charset="utf-8"/><!-- Set the viewport width to device width for mobile -->
    <meta content="width=device-width, initial-scale=1.0 maximum-scale=1, user-scalable=no"
name="viewport"/><meta content="text/html; charset=utf-8" http-equiv="content-type"/>
<meta content="Florida International University - Web Communications" name="author"/>
<meta content="florida, international, business, miami, university, universities, college, colleges, higher
education, academics, school, worlds ahead" name="keywords"/>
<meta content="Florida International University is a vibrant, student centered public research university,
ideally located in Miami that is WORLDS AHEAD in its commitment to learning, research, entrepreneurship,
innovation and creativity so that our graduates are prepared to succeed in the global marketplace."
name="description"/><title>Home - Florida International University - FIU</title><!-- Included CSS Files -->
    <link href="/_assets/min/?b=_assets&amp;f=css/foundation.css,css/vr.css,css/fiu.css,webfonts/icons/ss-
social.css,webfonts/ss-standard.css,css/print.css" rel="stylesheet"/>

<!--[if IE 7]>
<link rel="stylesheet" href="http://www.fiu.edu/_assets/css/ie7.css">
<![endif]--><!--[if IE 8]>
<link rel="stylesheet" href="http://www.fiu.edu/_assets/css/ie8.css">
<![endif]-->

   <link href="http://www.fiu.edu/favicon.ico" rel="icon" type="image/png"/>
```

Example of HTML code for FIUs homepage

Tools

- Python
  - Learned Python programming on codecademy.com
- BeautifulSoup (A Python module)
  - Makes parsing HTML easy

Data Source


Store
Mac
iPod
iPhone
iPad
iTunes
Apple Support Communities
Welcome, Guest
Sign in
Apple Support Communities > iPhone > Using iPhone > Discussions
Is there a way to turn off 3g on iOS 6.1.3?
9 Views 1 Reply Latest reply: Apr 14, 2013 3:43 PM by Allan Sampson
emirgrbich
Apr 14, 2013 3:04 PM
If not, how do I restore to previois version without LTE support so that my battery can again
last longer?
Level 1 (0 points)
Categories: Wi-Fi, 3G and Bluetooth Tags: 3g
I have this question too (0)
Allan Sampson Central Texas
Re: Is there a way to turn off 3g on iOS 6.1.3?
Apr 14, 2013 3:43 PM (in response to emirgrbich)
This depends on your carrier.
If your carrier is AT&T, the option is no longer available.
Level 10 (114,680 points)
Like (0)

Storing Data

	A	B
1	2/16/2013 5:24	"I think ive sorted it worked for me anyway.went to sys prefs - Users & Groups - unlocked the padlock and clicked on my ac
2	2/16/2013 11:30	"I have similar issues. There is likely not much we can do this is planned obsolescence. The iOS updates have been rampin
3	2/16/2013 18:19	"Really after update 10.8.3 nomore hdmi problem . it fixes perfect Mac Mini Late 2012 HDMI TO HDMIThank you Fouad Abb
4	2/16/2013 19:59	"Thanks for the news. Will wait for the final release to update and see the magic."
5	2/16/2013 11:02	" Hello I am currently an owner of an Ipod Touch 4th generation. I see you are having problems with your Ipod Touch/Ipad/
6	2/16/2013 3:14	"Ok So I found out what was happening. The Magsafe charger was BAD and it showed me the lights but was not charging th
7	2/16/2013 21:10	"Omg this worked it solved my problem thanks sooo much"
8	2/16/2013 20:58	"hi i have the same thing to i some one stolen my ipod to and i thrid find my iphone on icloud but it nevers show up to my r
9	2/16/2013 22:10	"My iPhoto also crashes every time I try to edit a photoIll edit a couple of them and then the wheel starts spinning and it cr
10	2/16/2013 23:12	"Process: iPhoto [6378]Path: /Applications/iPhoto.app/Contents/MacOS/iPhotoIdentifier: com.apple.iPhotoVersion: 9.4.2
11	2/16/2013 21:51	"Do you sync your iPhone and iPad on the same computer? If so that could be the issue. It could be remedied by plugging i
12	2/16/2013 14:20	"Like @lepaapl I upgraded to 7.6.3 and for the first 24 hours it seemed as if Apple had actually fixed this. Alas after a day th
13	2/16/2013 6:19	"I just had the same problem too. I have an iPad 2 and when I first downloaded three books that I had on my iPad before th
14	2/16/2013 6:11	"no my husband never accesses my account but I do occasionaly acsess his. I have had a few Ids so there might be a couple
15	2/16/2013 1:18	"I have had exactly the same problem. I was so happy with the battery life on my 5th gen (after replacing my 3rd gen). I mo
16	2/16/2013 10:00	"There is no Battery Usage button per se on your ipod 5G:Goto:Settings --> General-->Usage then scroll all the way down w
17	2/16/2013 14:03	"imessage still works!!"
18	2/16/2013 23:58	"Im trying to move files from my old computer to a new one. Using the old computer itunes I can see the new windows 8 it
19	2/16/2013 23:53	"I just set up my new Apple TV today and I noticed when I log in with my Apple account there are a few movies I can see in
20	2/16/2013 23:45	"I forgot answers to my security questions"
21	2/16/2013 23:48	"If you have a rescue email address (which is not the same thing as an alternate email address) set up on your account then
22	2/16/2013 23:20	"Has anyone figured out how a newbie can use microsoft exchange on mountain lion?"
23	2/16/2013 23:14	"Hello again.Tonight my client ran into an interesting problem with his HP DeskJet F4280 USB printer. It refused to print be
24	2/16/2013 23:06	"I am unable to pair with my apple white wireless keyboard. Im running 10.5 cant type passkey and no passkey "
25	2/16/2013 22:55	"I took our iMac in at the direction of Apple to have the hard drive replaced. When we got it back I restored from Time Ma
26	2/16/2013 23:18	"Apart from native apps only web links are showing. The above occurred after a failed iOS 6.1.1 update. This gave Error 160
27	2/16/2013 23:32	"You can restore from your icloud backup if youve had backup with that."
28	2/16/2013 23:18	"Wie funktioniert die Datenbertragung vom macBook (2008) zum macBook pro (retina 13 zoll)? Welche Kabel brauche ich. V
29	2/16/2013 23:03	"My main library of 12000 photos appears to be at overload stage with a regular spinning ball. I want to try to split the librar
30	2/16/2013 22:57	"Hello everyone! I finally decided to ditch all my PowerPC Macs and get a Mac Pro! Ive gotten tired of having a gaming PC n
31	2/16/2013 23:46	"I should be specific. My budget is $1200 for the *base model* not for the Mac + upgrades. Ill get the GPU RAM and CPU up

Data Details

- Over 500 GB of HTML
- 6.5 GB of text from posts
- 2,400 days of data
- 15.6 million posts
- 1.08 billion words

Errors

1776	11/1/2011 18:23	"Could some one supply me with the email address of Apple cus	3456249
1777	11/1/2011 16:42	"When I get a call a +1 will appear in front of it and make it so the	3456248
1778	11/1/2011 16:38	"I was able to use Excel 2011 for several languages including Kor	3456247
1779	11/1/2011 20:02	"cynthiafrommountain view wrote: What do I need to do to get t	3456247
1780	11/1/2011 16:36	"I had to reboot my computer and lost all my music. Now when	3456246
1781	11/1/2011 18:34	"What version of windows are you running?"	3456246
1782	11/1/2011 16:35	"I have tried a few solutions I have read others had success with:Uninstalled iPhoto and reinstalled from iLife disk performed Software Up	
1783	[0x96517b5a] Kernel stack: 27 semaphore_wait_continue + 0 [0x22a88f] Binary Images: 0x1000 - 0x25fff com.apple.AirPortBaseStationAgent 1.5.5 (155.2) <0001		
1784	cf2] 27 _dispatch_call_block_and_release + 16 (in libSystem.B.dylib) [0x9654ba24] 27 ____StartKernelListener_block_invoke_8 + 36 (in DirectoryService) [0x332		
1785	box 1.6.5 (???) <21164164-41CE-61DE-C567-32E89755CB34> /System/Library/Frameworks/Carbon.framework/Versions/A/Frameworks/HIToolbox.fra mework/		
1786	n ??? (???) <EE2AFFE6-5157-1B18-F6B2-6859AED0DEA8> /usr/sbin/racoon 0x96517000 - 0x966beff7 libSystem.B.dylib ??? (???) <2DCD13E3-1BD1-6F25-119A-386		
1787	:threadMain() + 0 [	3456245	
1788	11/1/2011 17:00	"Did a full reboot and powered off and a restart. Empty iPhoto li	3456245
1789	11/1/2011 17:13	"Thank you. I did exactly that. Held down the ALT key and relaunched. New Library with images and one even without. After the reboot/p	
1790	es: 0x1000 - 0x25fff com.apple.AirPortBaseStationAgent 1.5.5 (155.2) <00010203-0405-0607-0809-0A0B0C0D0E0F> /System/Library/CoreServices/AirPort Base S		
1791	ceive_continue + 0 [0x210d84] Thread 3d4 DispatchQueue 2 User stack: 19 start_wqthread + 30 (in libSystem.B.dylib) [0x9653d5c6] 19 _pthread_wqthread + 390		
1792	ue_receive_continue + 0 [0x210d84] 2 lo_allintrs + 302 [0x2a1c2e] 2 interrupt + 192 [0x2ab423] 2 lapic_interrupt + 108 [0x2b32f2] 2 mp_kdp_exit + 868 [0x2b456		
1793	) [0x9653e382] Ker	3456245	
1794	11/1/2011 16:28	"My new phone is saying iPhone disabled connect to iTunes WH	3456244
1795	11/1/2011 16:24	"After the most recent upgrade to iTunes it now opens up at ran	3456243
1796	11/1/2011 16:23	"My MobileMe alias email is not being received after I transfere	3456242
1797	11/1/2011 16:18	"I have an AT&T modem/router in use as a bridge to enable my A	3456241
1798	11/1/2011 16:41	"If you really have the 2-Wire gateway configured as a bridge mo	3456241
1799	11/1/2011 16:45	"Bob I followed your instructions and under Connect Using it sho	3456241
1800	11/1/2011 16:46	"Bob I am located in the United States."	3456241
1801	11/1/2011 16:56	"under Connect Using it shows Ethernet.Then despite what it mi	3456241
1802	11/1/2011 16:57	"Im pretty new at this just bought the iMac three days ago. Assu	3456241
1803	11/1/2011 16:58	"Suggest that you leave things as is. A bridge only modem will n	3456241
1804	11/1/2011 17:01	"Thank you! I appreciate the help."	3456241

Errors

- When you press ENTER to go to a new line in a text document you add “\n”
- The code for counting posts was counting extra lines and inflating the actual volume of posts for the day
- Fixed it and re-scraped the data to make sure the results are consistent

Analyzing Data

- Using Pythons built-in CSV (comma separated value) module to do a word frequency count took 40 minutes for one day
- 1,600 processing hours or 200 hours on an 8-core CPU (8 days)
- Too long. After much research I rewrote the code using Pythons mmap (Memory-mapped file support) module
- From 8 days to just under 10 hours!

Analyzing Data

- How to iterate through dates in order?
- Use “num2date” from “matplotlib.dates”

Example:
734984 = 2013-04-26
734985 = 2013-04-27

734988 = 2013-04-30
734989 = 2013-05-01

Machine Learning

- The study of systems that can learn from data
- Types:
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Focus on simple supervised learning
  - Generate a function that maps inputs to desired outputs

Simple Example

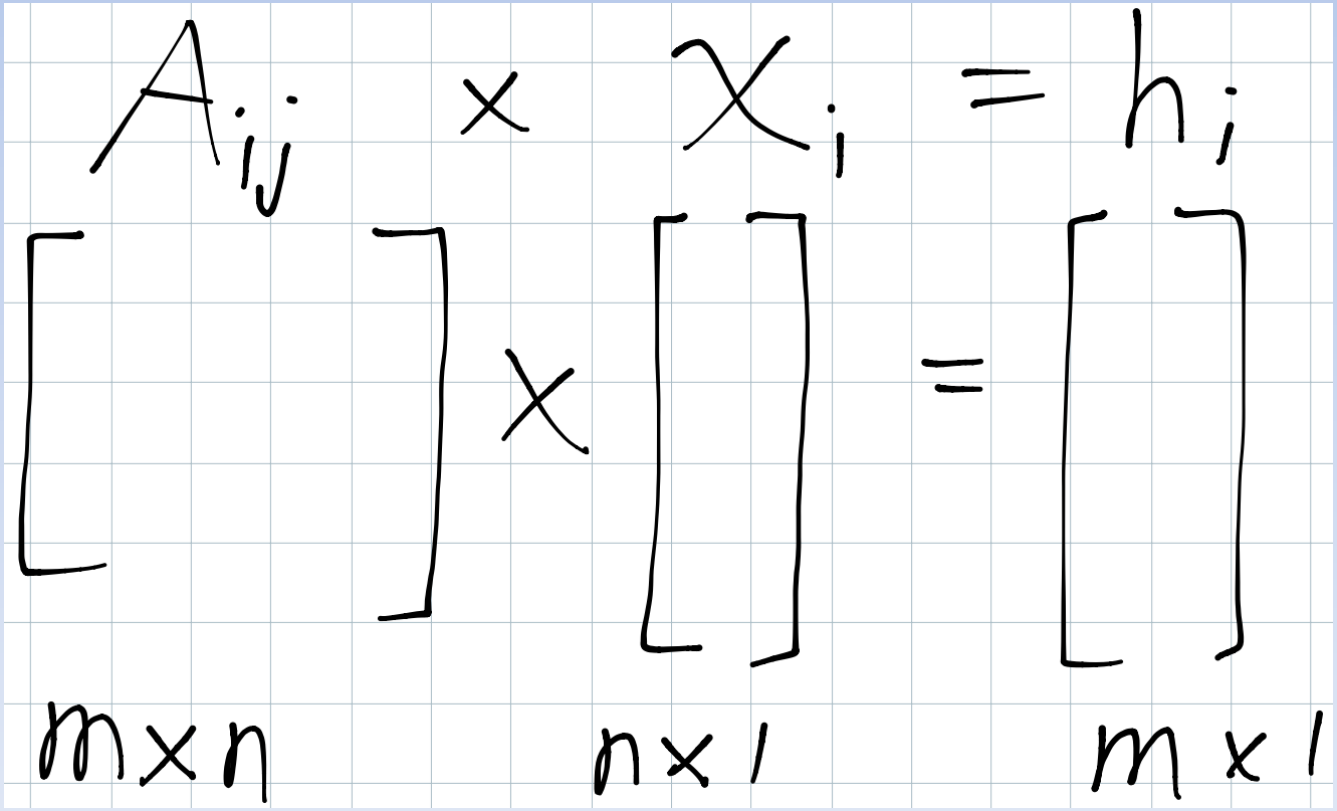
Training Set

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Picture from Andrew Ng, Coursera.org machine learning course

- Matrix “A” contains the features
- Vector “X” contains the coefficient's that the algorithm will solve for
- Vector “h” is the hypothesis



Cost Function

$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

- J is the cost function
- m is the number of training examples
- y is the desired output
- h is the hypothesis

Gradient Descent

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

- θ_j Coefficients to be solved for to minimize J
- α Learning rate

$$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$$

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

Updating Thetas

Correct: Simultaneous update

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\theta_1 := \text{temp1}$$

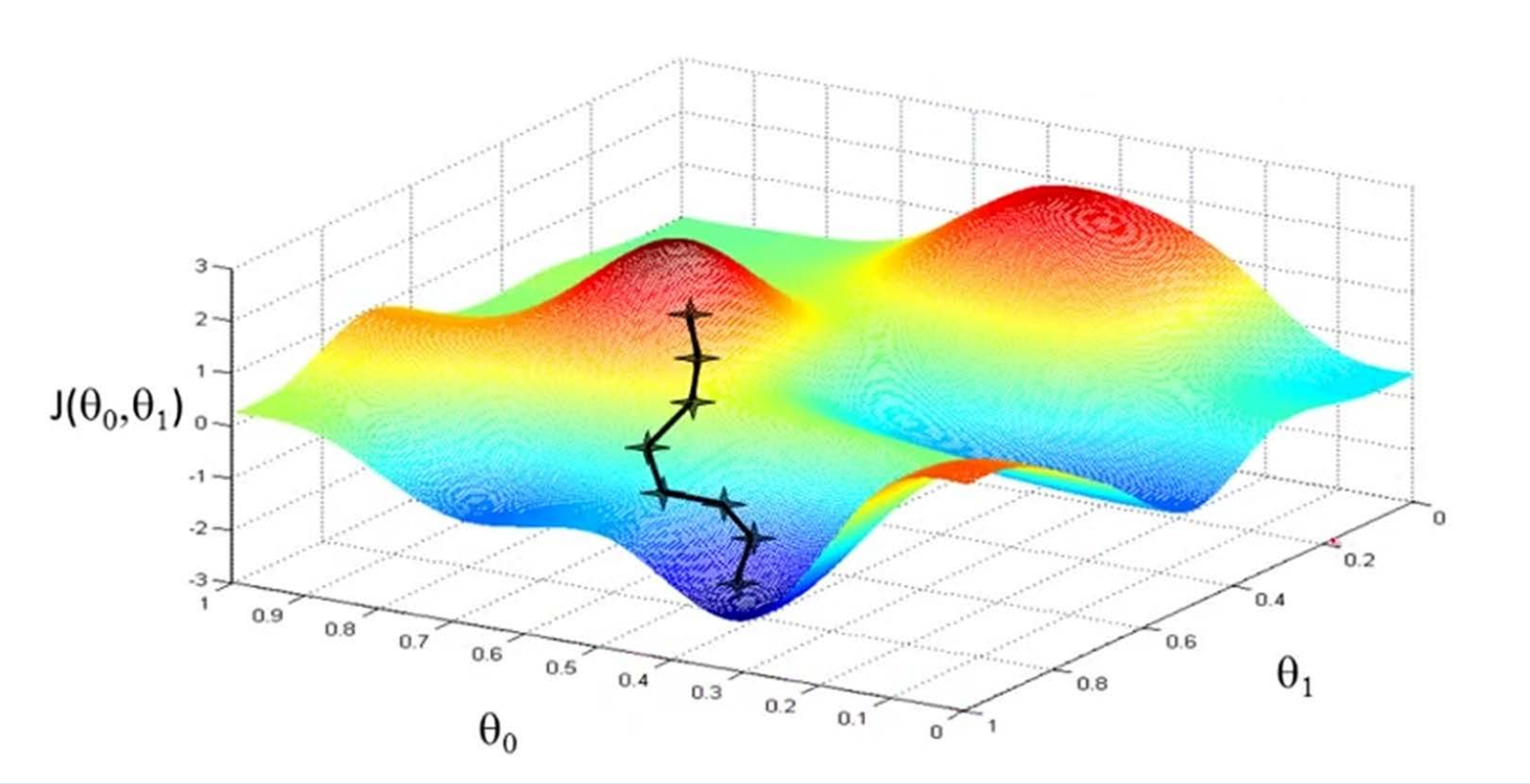
Incorrect:

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_1 := \text{temp1}$$

Picture from Andrew Ng, Coursera.org machine learning course

Local Optimum

- Optimal values depend on how the thetas were initialized



Picture from Andrew Ng, Coursera.org machine learning course

Thetas

- Features should have a positive or negative influence
- Initialize all thetas to zero
  - (My hypothesis)

Learning Rate & Features

- Features are data points for a specific training example e.g. house size: 950 sq-ft, 2 car garage
- Learning rate is tricky… Too large makes the function diverges to infinity. Too small takes allot of iterations to converge

Example

$$\begin{bmatrix} 1 & 950 & 1 \\ 1 & 1500 & 1 \\ 1 & 2000 & 2 \end{bmatrix} \times \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} h_0 \\ h_1 \\ h_2 \end{bmatrix}$$

3 training examples

"b" in $y = mx + b$

house size

garage

Initialized zeros

$Y_1 = 180{,}000$, $Y_2 = 240{,}000$, $Y_3 = 310{,}000$

house Prices for the 3 examples

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

$$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$$

$$\theta_0 = 0 - \alpha \left[(0 - 180{,}000)\tfrac{1}{3} + (-240{,}000)\tfrac{1}{3} + (310{,}000)\tfrac{1}{3} \right]$$

$$\theta_1 = 0 - \alpha \left[(-180k)\tfrac{956}{3} + (-240k)\tfrac{1500}{3} + (-310k)\tfrac{2000}{3} \right]$$

$$\theta_2 = 0 - \alpha \left[(-180k)\tfrac{1}{3} + (-240k)\tfrac{1}{3} + (-310k)\tfrac{2}{3} \right]$$

$$\theta_0 = 0 - \alpha(-243{,}333)$$

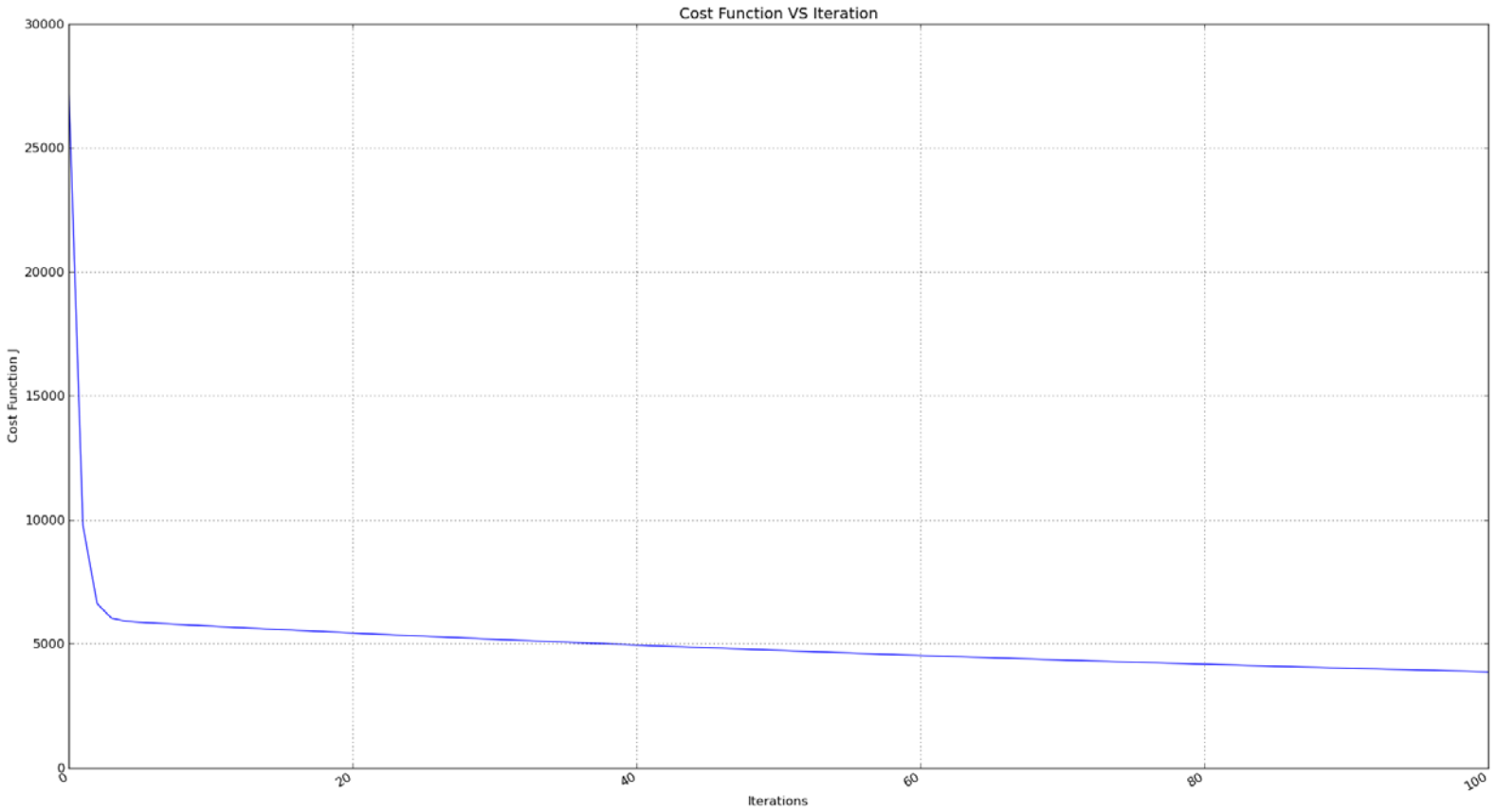
$$\theta_1 = 0 - \alpha(-383{,}666{,}666)$$

$$\theta_2 = 0 - \alpha(-346{,}666)$$

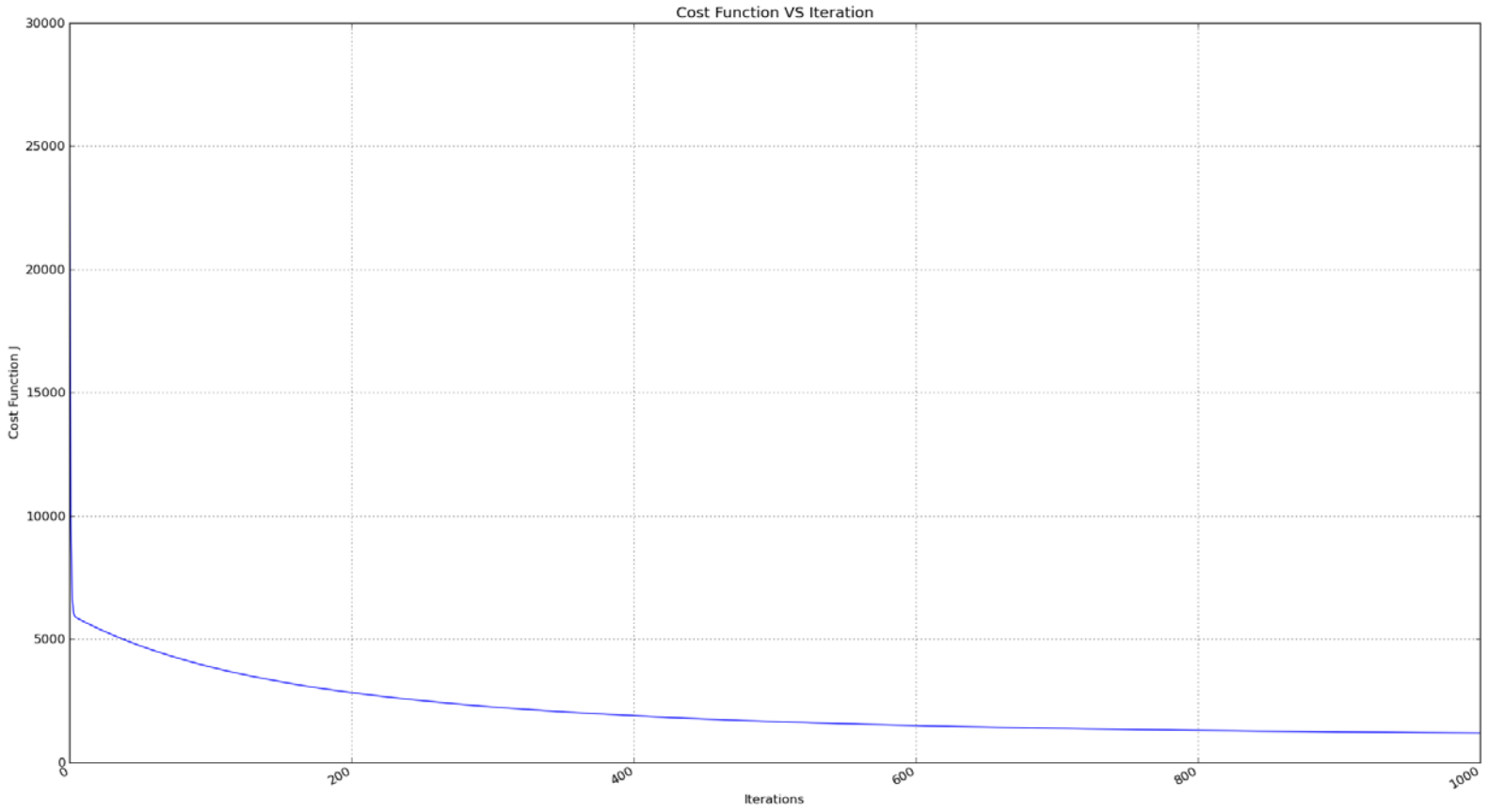
Learning Rate & Features

- Learning rate must be small enough to accommodate the largest feature
- Scale features to get them within the same magnitude of each other
- A feature with one order of magnitude larger than the rest will require one order of magnitude more iterations

Cost Function VS Iteration



Cost Function VS Iteration



Using Social Media To Predict Stock Prices

- The hypothesis: Prices have an instant response to the words used in social media
- Use word lists from a research from Tim Loughran & Bill McDonald
- Use Harvard IV negative words list

Details

- Training data from Jan, 2007 to Feb 1, 2012
- Testing data from Feb 2, 2012 to Feb 27, 2013
- Features used:
  - Inflation, Forum post volume (daily activity), volume of Google searches for Apples products, volume of Google searches for Microsoft, and 56 randomly selected words with sufficient volume from the word lists

Features

Features	Theta
b	0.277
Apple Trends	4.708
Microsoft Trends	-0.156
Post Volume	-0.664
Inflation	8.475
afraid	-0.272
annoying	-0.301
avoid	-0.043
bad	-0.222
break	-0.034
broken	0.122
bug	0.645
corrupted	-0.043
dead	-0.194
error	0.253
expensive	-0.152
failed	-0.302

Positive Theta: Positive correlation, the larger the feature is the higher the predicted stock price will be.

Negative Theta: Negative correlation, the larger the feature is the lower the predicted stock price will be.

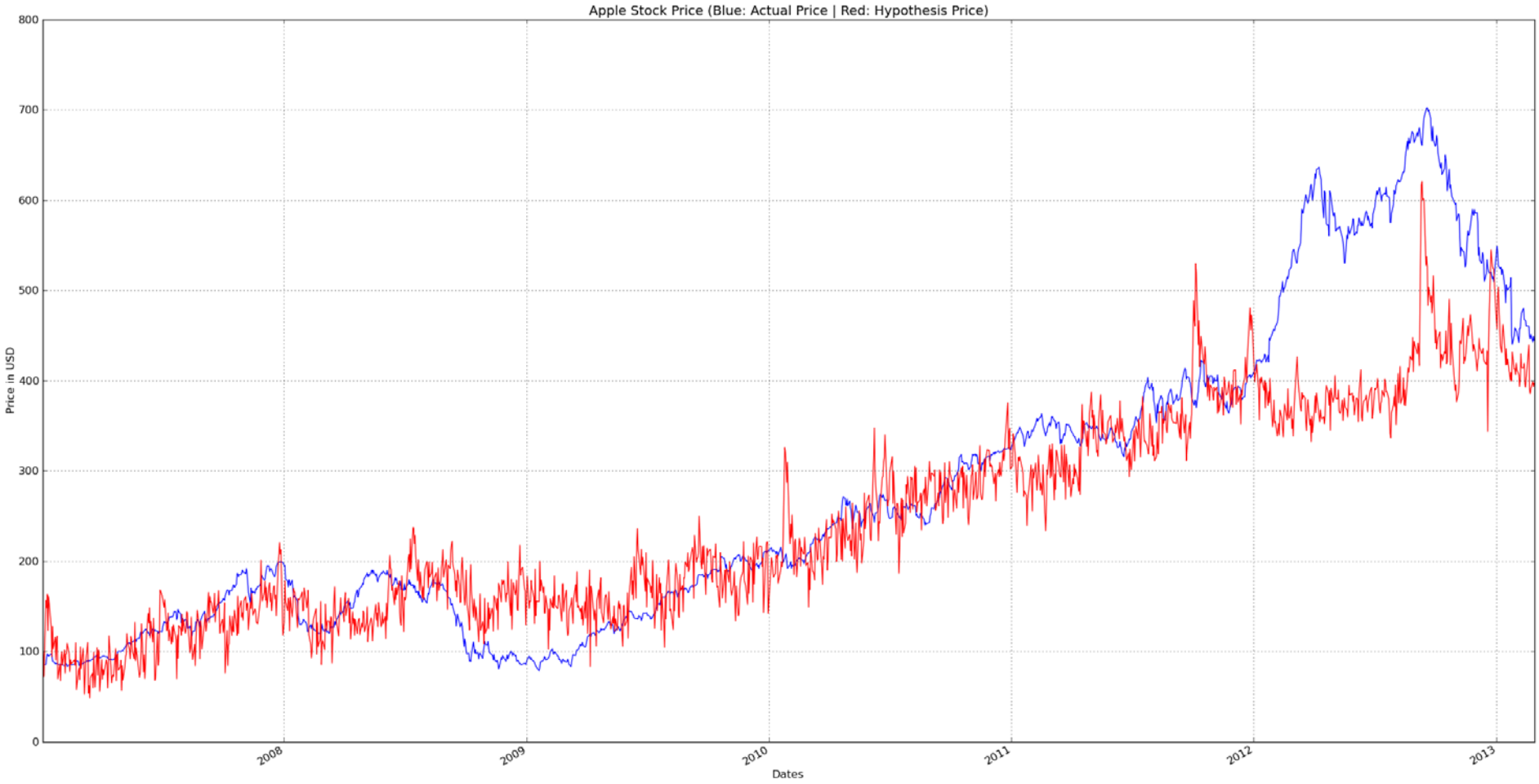
Magnitude of Theta: A large absolute value means there is a strong correlation.

All features are scaled so their average “feature” contribution is 30. (e.g. h(x)=(30)*(-0.302)+…)

Apple and Microsoft trends range from 10-100 (not scaled). Inflation ranges from 1-1.5 (The Theta value for inflation should be interpreted as 0.8)

Features	Theta
fix	1.118
missing	0.024
order	-0.235
problem	-1.431
problems	-0.464
trouble	0.168
unfourtunately	0.242
wrong	-0.612
annoying	-0.301
fustrating	0.064
quit	-0.343
better	-1.211
easy	-0.365
exellent	0.356
good	0.298
great	-0.391
happy	-0.443
perfect	0.123
perfectly	0.236
best	0.091
crash	-0.095
dont	-1.734

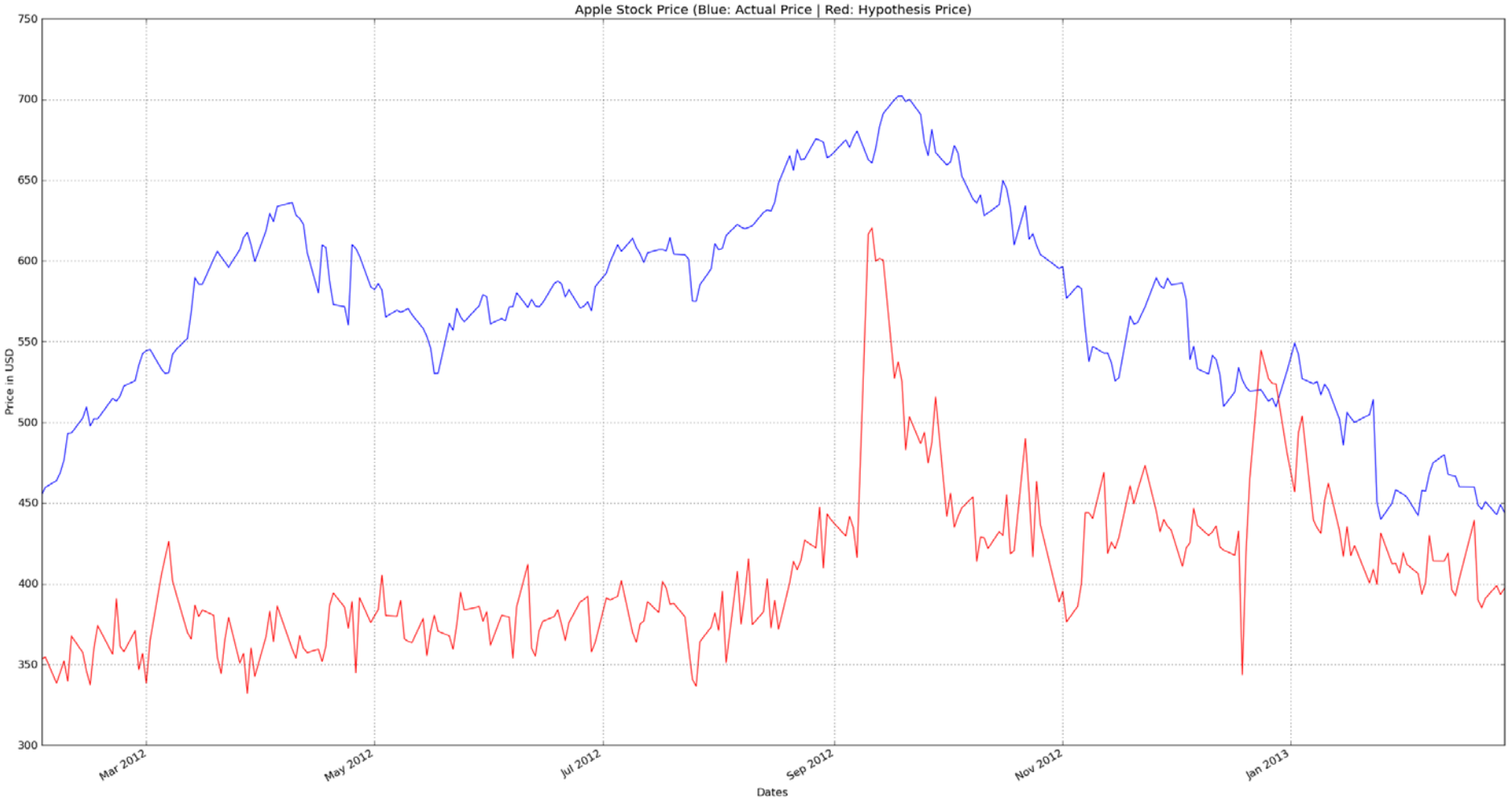
Features	Theta
hard	-1.782
lost	0.425
love	-0.276
luck	0.208
missing	0.024
no	1.869
not	1.909
smart	0.141
sorry	0.442
strange	-0.032
weird	-0.050
yes	-0.451
doesnt	0.240
update	-0.911
free	-0.593
upgrade	-0.231
buy	0.141
only	-2.007
security	0.431
isnt	0.194
downgrade	0.036
hate	-0.117



Blue line: Actual Stock Price Red line: Hypothesis Stock Price

X-axis: Dates (Jan 1, 2007 – Feb 27, 2013) Y-Axis: Stock Price in USD

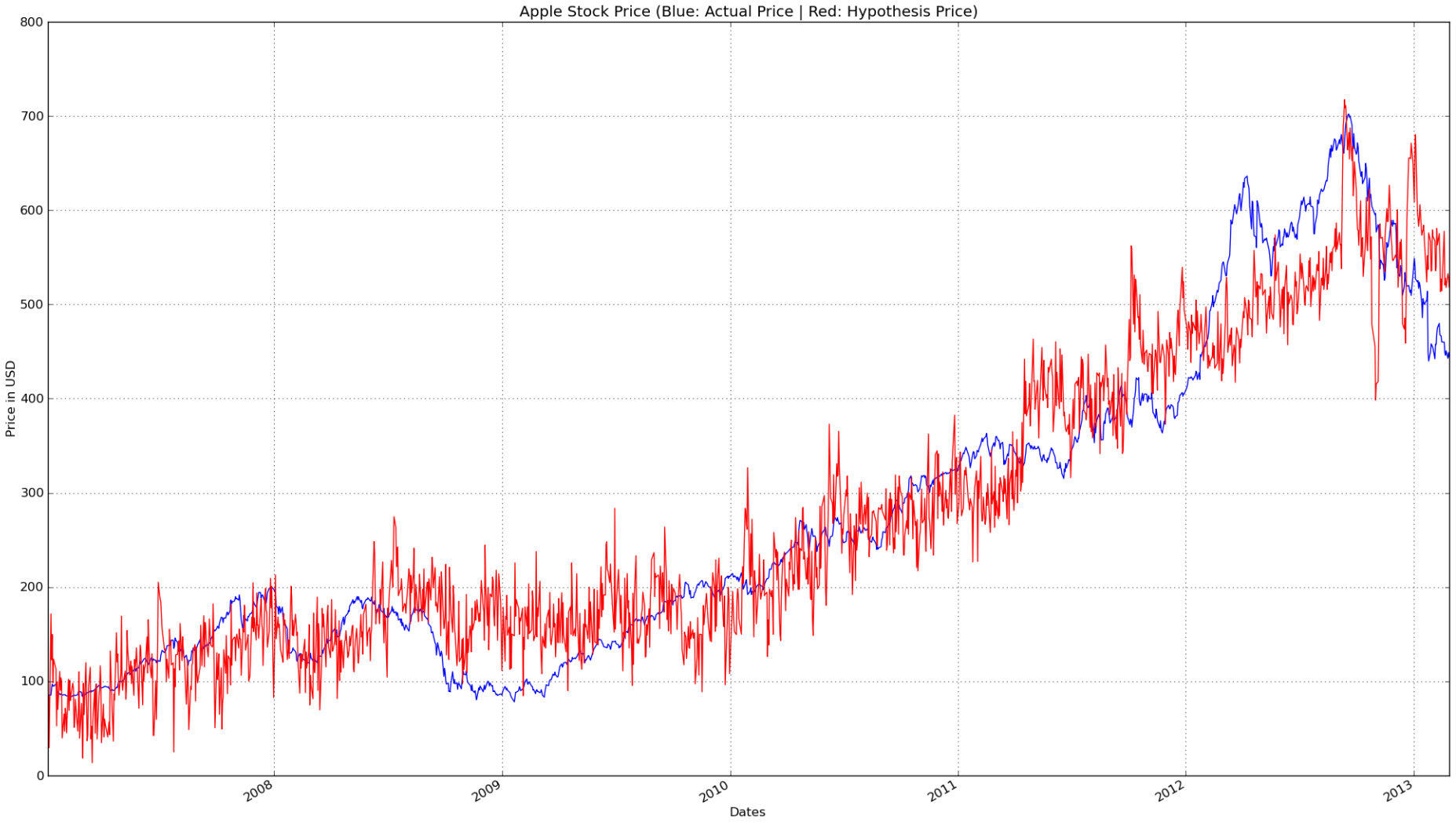
Learning Data: Jan 1, 2007 – Feb 1, 2012 Testing Data: Feb 2, 2012 – Feb 27, 2013



Blue line: Actual Stock Price Red line: Hypothesis Stock Price

X-axis: Dates (Feb 2, 2012 – Feb 27, 2013) Y-Axis: Stock Price in USD

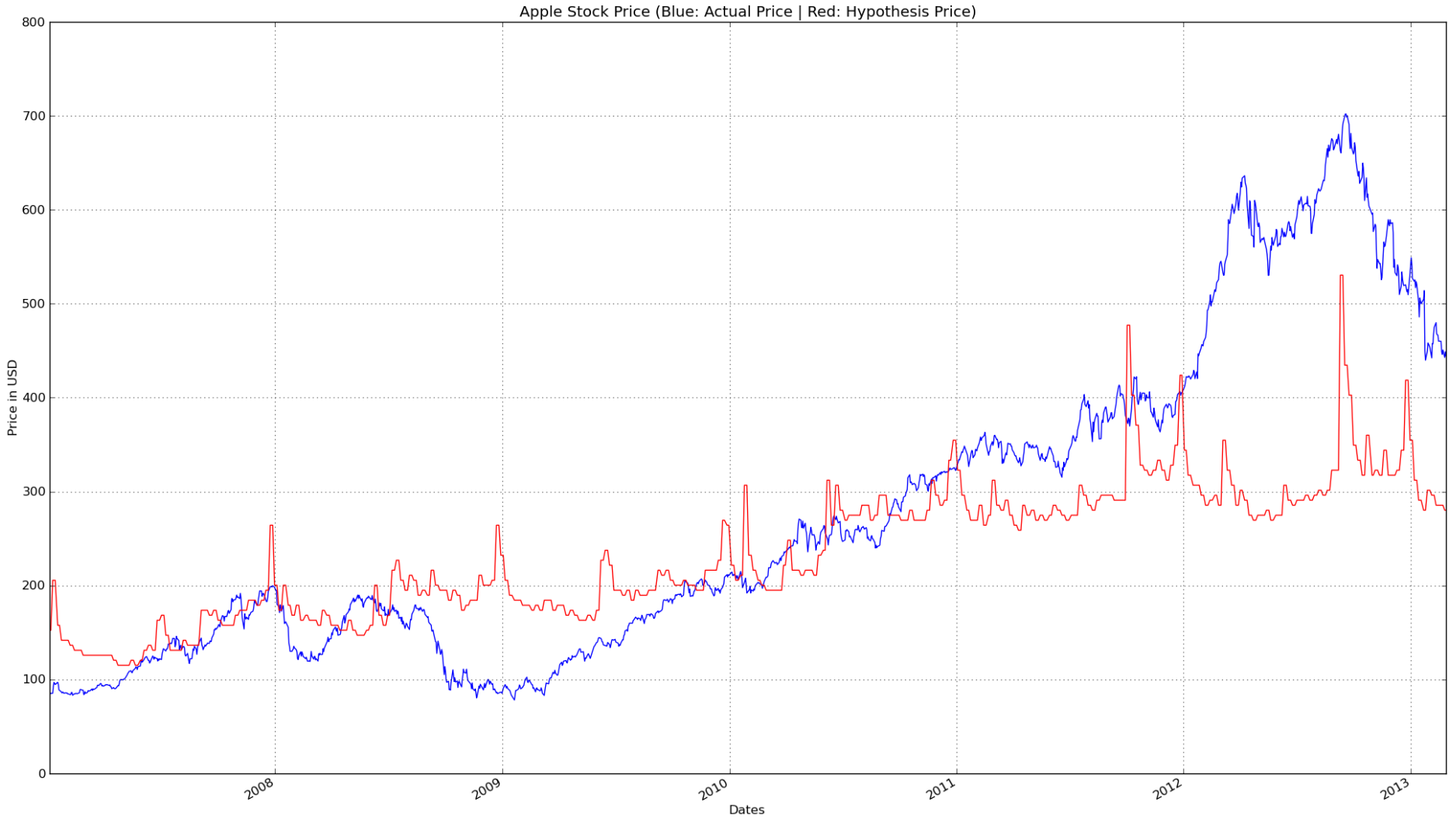
For Fun... Train the data on all dates



Conclusion

- Strong positive correlation to search volume of Apples products
- Google Trends alone do a remarkable job predicting stock prices

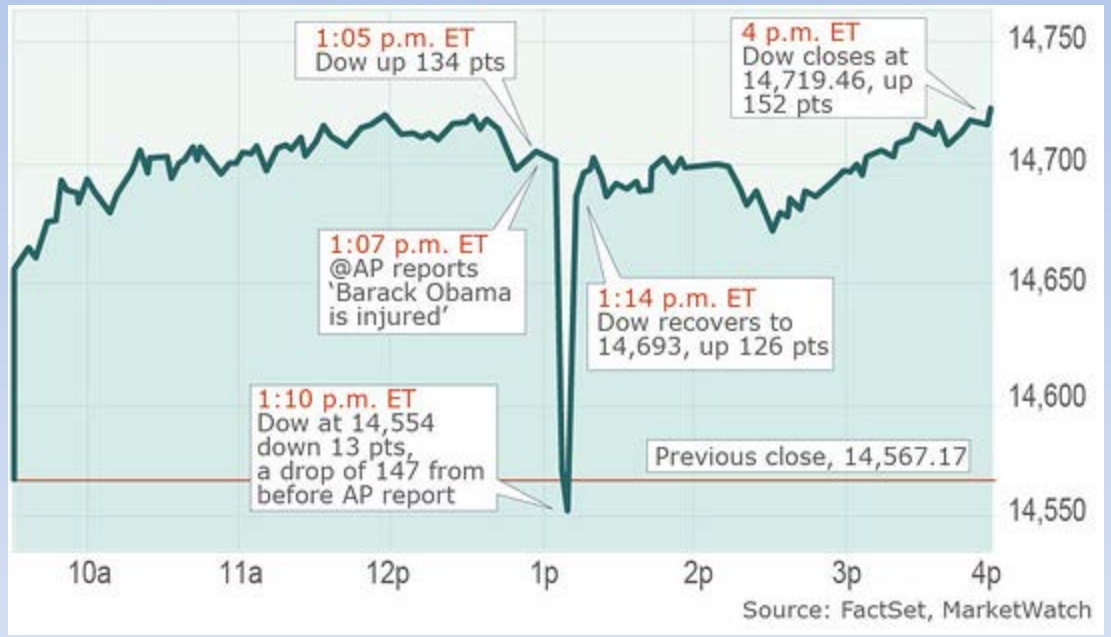
Google Trends Only



Discussion

- Apples forums are for resolving issues
- Should repeat analysis for Apple rumor forums
- Lacking features that relate to company fundamentals
- Lacking features that relate to Apples overseas activity
- Lacking features of Apples other competitors
- Weekends were excluded

News twitter account was hacked and posted a fake tweet about a bomb at the white house and Barack Obama was injured… Within seconds the Dow fell 147 points. Traders have programs analyzing social media and automatically execute trades based on words used.



Run Example

- Learn Python on Codecademy.com
- Learn Machine Learning on Coursera.org
- Word lists: http://www3.nd.edu/~mcdonald/Word_Lists.html