**INSY 5378: Project 2: Pokemon Go! Analytics**

**Group Members:**

Sathya Narayanan Manivannan

Sandeep Ramesh

Vaithiraj Sokkalingam

**Introduction:**

The primary goal of the project is to scrape data from web pages provided by the professor to build a prediction model for the augmented reality game Pokemon GO! We have also performed a secondary optional task which was to analyze the images from the web pages using tensorflow for Deep Learning.

We have used BeautifulSoup package of python for scraping data from html files and we have used pandas dataframes to clean and process data. We have used matplotlib and scatter matrix for visualizing the data at various steps and finally we have used sklearn to build a machine learning model to predict the total number of ratings for iOS and android on 2016/11/01 11:50 PM.

**Step1: Web Scraping**

First we used OS.Walk to traverse through all files in the folder and used Beautiful Soup to do web scraping of all Html files in the folder. The result set consisting of relevant numerical values was then stored in 2 separate Json files (counts\_size\_android.json, counts\_size\_ios.json).

1. **iOS files:**

We created a soup object for every iOS html file and extracted the required values like Current ratings, All ratings and file size. We navigated to the “span” and “li” tags in the html file and extracted the data. For the files for which data was missing we initially populated it as “NA”.

1. **Android Files:**

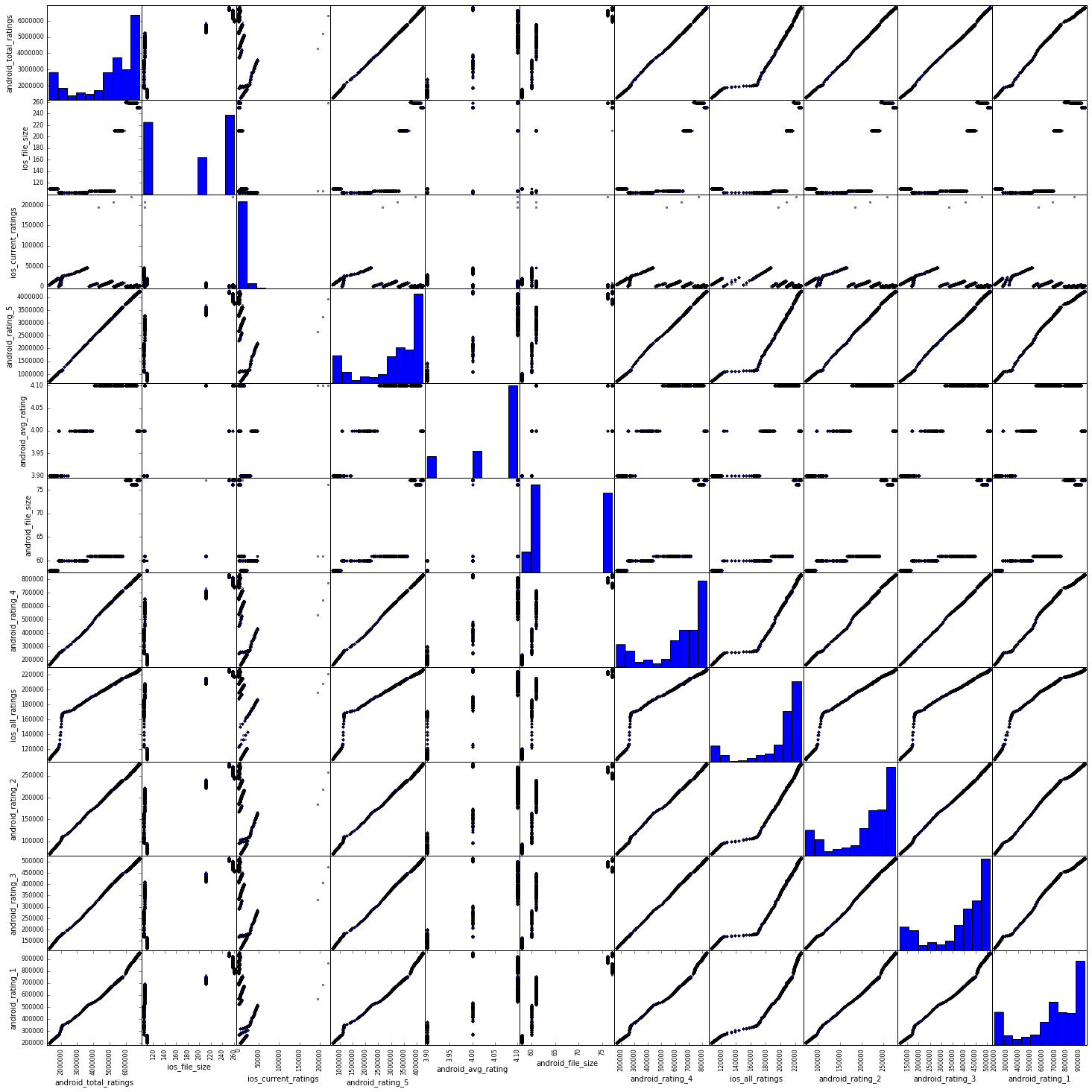
We again created a soup object for every android html file and extracted the required values like Average rating, total ratings, number of ratings with 5 star, number of ratings with 4 star, number of ratings with 3 star, number of ratings with 2 star, number of ratings with 1 star and file size. We used the “span” and “div” tags to extract the required values for android files. For the files for which data was missing we initially populated it as “NA”.

**Step2: Data Organization**

1. The Extracted values from the 2 json files were then placed into a dictionary date\_time\_count\_dict() and it was converted into a pandas dataframe where the index is datetime and values are a combination of 11 iOS and Android values.
2. The resulting data frame was also converted to 3 file types namely data.json, data.csv and data.xlsx.

**Step3: Data Exploration**

1. We now used the describe method to explore the data.
2. A scatter matrix was then plotted to find pairs of variables with high correlations.



1. We also found the Pearson correlation coefficient for the identified pairs.

android\_rating\_1, android\_rating\_5: 0.993504986113

android\_rating\_1, android\_total\_ratings: 0.995139670361

android\_rating\_3,android\_rating\_1: 0.992714549072

android\_rating\_3, android\_rating\_5: 0.999576019713

android\_rating\_2, android\_rating\_1: 0.994361440471

android\_rating\_2, android\_rating\_3: 0.999477140687

android\_rating\_2, android\_rating\_5 : 0.999493317211

android\_rating\_4, android\_rating\_1 : 0.993417513569

android\_rating\_4, android\_rating\_2 : 0.999379835274

android\_rating\_4,android\_rating\_3 : 0.999884538467

android\_rating\_5, android\_rating\_4 : 0.9996718162

android\_rating\_3, android\_total\_ratings : 0.99954234989

android\_rating\_2, android\_total\_ratings : 0.999629379239

android\_rating\_5,android\_total\_ratings : 0.999864336081

android\_rating\_4, android\_total\_ratings: 0.999699177204

android\_rating\_1, ios\_all\_ratings: 0.949641890542

1. We plotted a time series graph for each of the 11 variables using MatplotLib. Some of them have been plotted together as a group.

Datetime vs android\_total\_ratings,

Datetime vs ios\_all\_ratings,

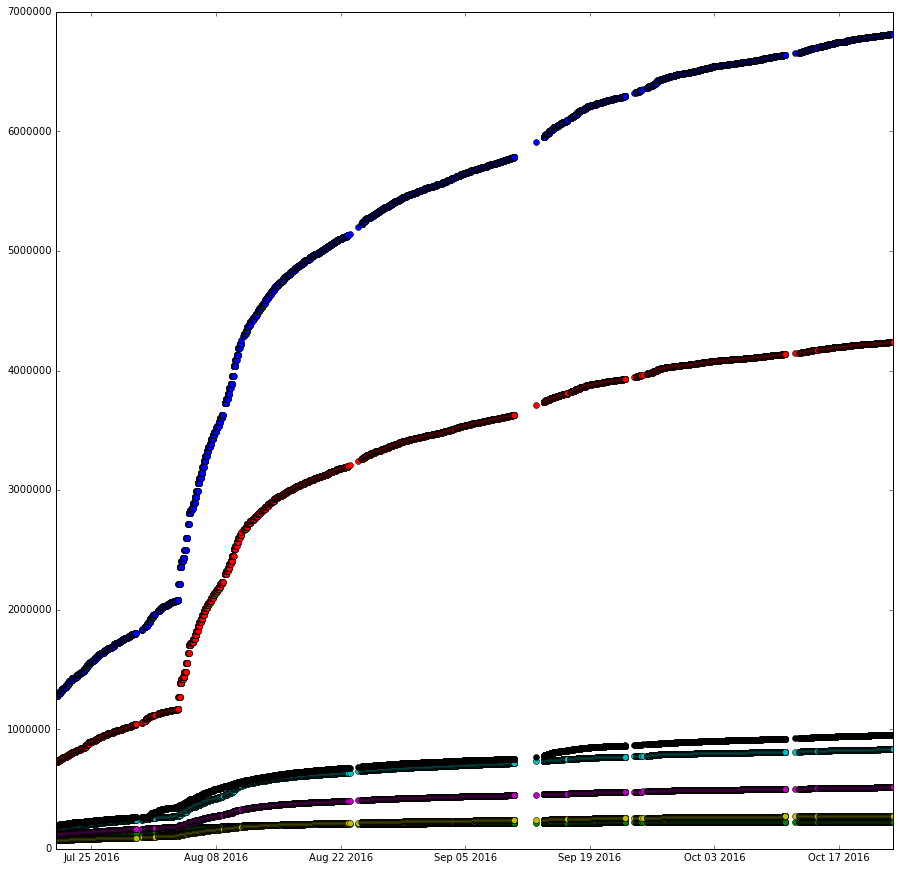
Datetime vs android\_rating\_5,

Datetime vs android\_rating\_4,

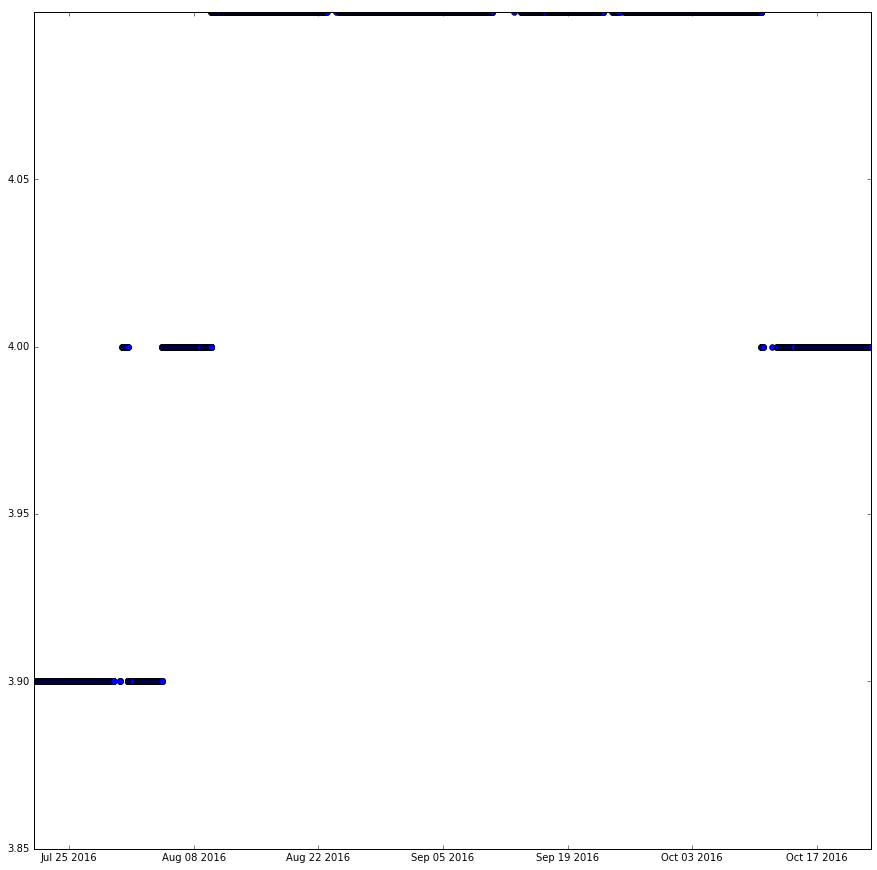
Datetime vs android\_rating\_3,

Datetime vs android\_rating\_2,

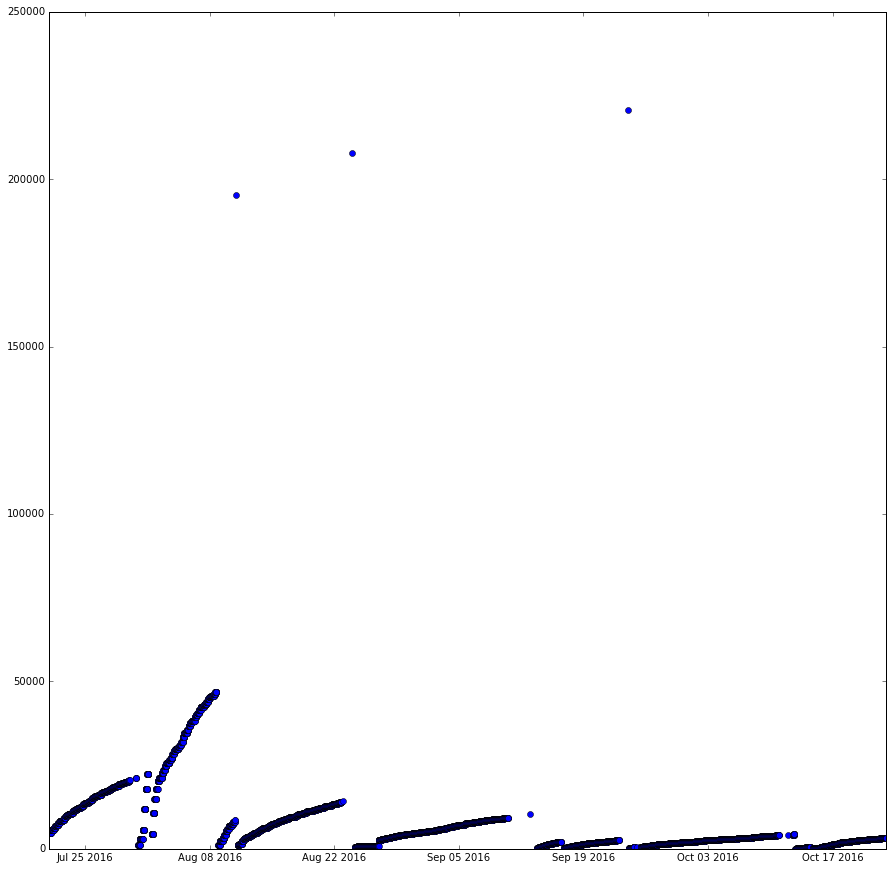
Datetime vs android\_rating\_1



Datetime vs android\_avg\_rating

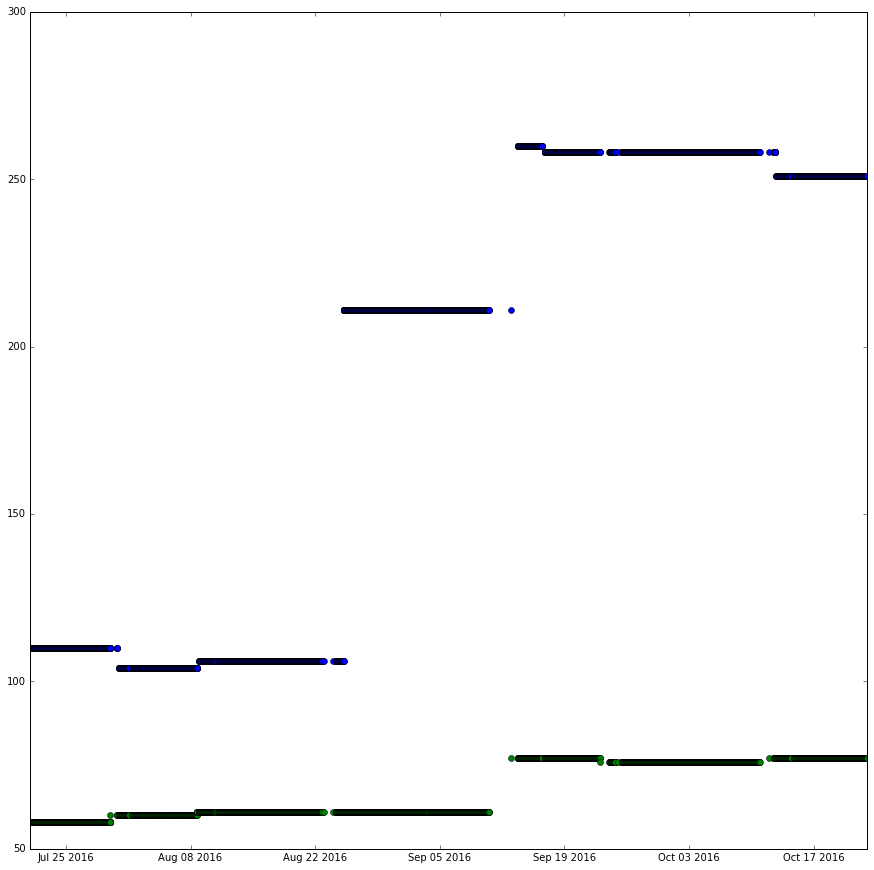


Datetime vs ios\_current\_ratings



Datetime vs ios\_file\_size

Datetime vs android\_file\_size



**Step 4: Prediction Model:**

We then built machine learning models using Linear Regression in Sklearn . The model was used to predict the values of iOS\_all\_ratings and Android\_total\_ratings using Linear regression. The predicted values are given below. We have used the features and values of 30th October 23:50 to predict the value of 1st October.

[ 6805482.91760286 5618731.81351341 6702793.36238438 ...,

1825697.93812002 6221902.17059595 5114467.4947342 ]

**Predicted Value of Android Total Rating for 11/01/2016 23:50:00 is [ 7035760.58457758]**

[ 227494.79106695 210589.86036687 224410.21363409 ..., 132359.98980016

219108.61442503 212834.17589185]

**Predicted Value of IOS Total Rating for 11/01/2016 23:50:00 is [ 231473.80602788]**

**Step5: Deep Learning**

**Objective:**

To identify and download all unique screenshots from iOS and Android pages. To analyse the content of these images using Tensorflow.

**Code Description:**

We used OS.Walk to traverse through all files in the folder and used Beautiful Soup to do web scraping of all Html files in the folder. All the URL for iOS and Android images were scraped from all the webpages and were stored in a set to make the entries unique. We found that few URLs were pointing to JPEG images and the rest were pointing to images which were not JPEG. So, we used StringIO and PIL to convert these images to JPEG and saved all the images. We had 20 unique images in total including iOS and Android.

Our next step was to analyze these images using tensorflow python script provided by google.

Tensorflow works in Python 3.5 and above versions. So we used image\_classify\_by\_tensorflow.py in Python 3.5.2 and passed the image names as arguments to the script.

**Sample Execution:**

F:\MS 2nd Sem\Data Science\Project 2\Command_line_input.JPG

F:\MS 2nd Sem\Data Science\Project 2\Command_line_output.JPG

The following are the images and their respective tensorflow analysis results.

imgfile0.jpg



web site, website, internet site, site 0.8907701373100281

menu 0.0036376425996422768

monitor 0.0018526039784774184

screen, CRT screen 0.001841831486672163

analog clock 0.0017735683359205723

imgfile1.jpg



web site, website, internet site, site 0.42240995168685913

comic book 0.03247756138443947

carousel, carrousel, merry-go-round, roundabout, whirligig 0.02088976837694645

fountain 0.01781134307384491

safety pin 0.014400497078895569

imgfile2.jpg



web site, website, internet site, site 0.6088576316833496

television, television system 0.056650057435035706

monitor 0.019958283752202988

notebook, notebook computer 0.016072208061814308

iPod 0.01179817970842123

imgfile3.jpg



aircraft carrier, carrier, flattop, attack aircraft carrier 0.09968294948339462

pole 0.03657454252243042

wing 0.02655319683253765

lakeside, lakeshore 0.024369418621063232

magnetic compass 0.023960646241903305

imgfile4.jpg



comic book 0.19361448287963867

maze, labyrinth 0.1932980716228485

web site, website, internet site, site 0.05235723778605461

monitor 0.029567185789346695

book jacket, dust cover, dust jacket, dust wrapper 0.027674004435539246

imgfile5.jpg



laptop, laptop computer 0.49859192967414856

web site, website, internet site, site 0.10645917803049088

monitor 0.06384018808603287

screen, CRT screen 0.029848331585526466

notebook, notebook computer 0.02801426686346531

imgfile6.jpg



web site, website, internet site, site 0.1163666844367981

laptop, laptop computer 0.0807962566614151

notebook, notebook computer 0.05348580703139305

joystick 0.04790791869163513

monitor 0.04169079661369324

imgfile7.jpg



fountain 0.20302647352218628

carousel, carrousel, merry-go-round, roundabout, whirligig 0.08313611894845963

comic book 0.05170505866408348

toyshop 0.0334254615008831

monitor 0.03227037563920021

imgfile8.jpg



space shuttle 0.2304239720106125

joystick 0.059921521693468094

racer, race car, racing car 0.05625808611512184

scoreboard 0.04957202821969986

airliner 0.04575682431459427

imgfile9.jpg



web site, website, internet site, site 0.6115028262138367

sunglasses, dark glasses, shades 0.04096995294094086

electric fan, blower 0.030481770634651184

comic book 0.019868792966008186

sunglass 0.018985025584697723

imgfile10.jpg



ashcan, trash can, garbage can, wastebin, ash bin, ash-bin, ashbin, dustbin, trash barrel, trash bin 0.15497566759586334

joystick 0.06404933333396912

cannon 0.03585103154182434

maraca 0.02726832963526249

pedestal, plinth, footstall 0.027154820039868355

imgfile11.jpg



web site, website, internet site, site 0.2275332808494568

envelope 0.09162567555904388

Band Aid 0.03712098300457001

pinwheel 0.029456494376063347

airship, dirigible 0.024857865646481514

imgfile12.jpg



web site, website, internet site, site 0.3661857843399048

safety pin 0.020037969574332237

sunglasses, dark glasses, shades 0.0167746189981699

toilet seat 0.015619936399161816

washer, automatic washer, washing machine 0.014380029402673244

imgfile13.jpg



ant, emmet, pismire 0.1064392551779747

monitor 0.09437637776136398

aircraft carrier, carrier, flattop, attack aircraft carrier 0.09096026420593262

wing 0.048978470265865326

web site, website, internet site, site 0.04596283659338951

imgfile14.jpg



web site, website, internet site, site 0.5248120427131653

monitor 0.06628231704235077

notebook, notebook computer 0.04221828281879425

television, television system 0.04088204726576805

laptop, laptop computer 0.011699662543833256

imgfile15.jpg



lawn mower, mower 0.3191564679145813

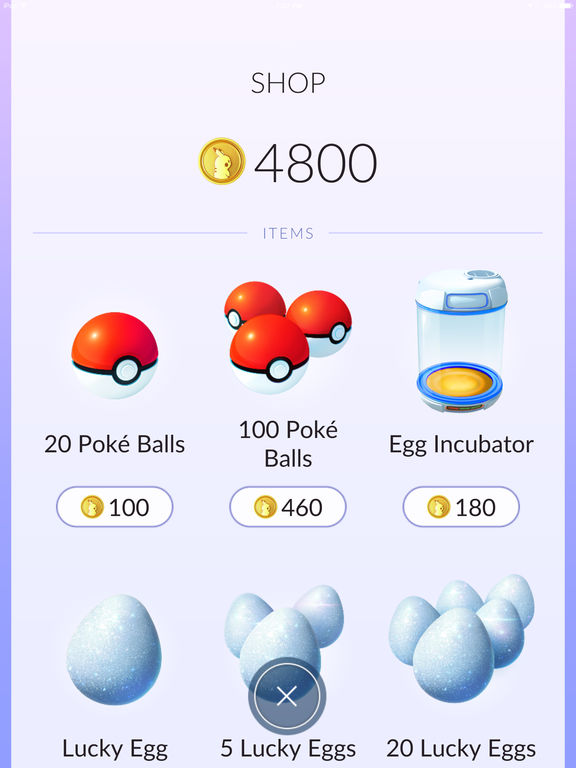
golf ball 0.05855679139494896

bow 0.03301812708377838

barrow, garden cart, lawn cart, wheelbarrow 0.028211992233991623

croquet ball 0.023382484912872314

imgfile16.jpg



web site, website, internet site, site 0.883571207523346

menu 0.008027322590351105

slot, one-armed bandit 0.004043694585561752

washer, automatic washer, washing machine 0.003706002375110984

hand-held computer, hand-held microcomputer 0.002964386250823736

imgfile17.jpg



web site, website, internet site, site 0.12342077493667603

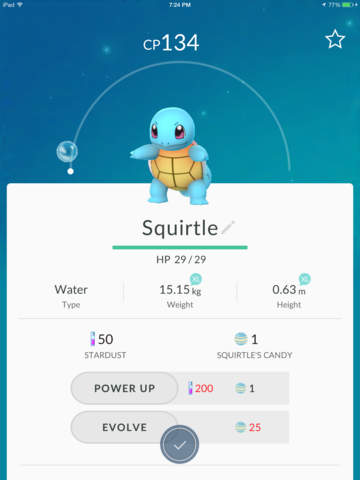
maze, labyrinth 0.07148678600788116

comic book 0.04789261892437935

joystick 0.04420957341790199

television, television system 0.037576720118522644

imgfile18.jpg



web site, website, internet site, site 0.9409151673316956

analog clock 0.0036712565924972296

envelope 0.002909436821937561

monitor 0.0022511144634336233

screen, CRT screen 0.0021692963782697916

imgfile19.jpg



web site, website, internet site, site 0.706946074962616

television, television system 0.02659899927675724

monitor 0.019617343321442604

screen, CRT screen 0.016214534640312195

pool table, billiard table, snooker table 0.0154481902718544

imgfile20.jpg



web site, website, internet site, site 0.36779189109802246

envelope 0.16913513839244843

binder, ring-binder 0.05812036246061325

tray 0.017636902630329132

monitor 0.017210371792316437

imgfile21.jpg



web site, website, internet site, site 0.5862426161766052

monitor 0.07197427749633789

television, television system 0.05955268442630768

comic book 0.04756322130560875

teapot 0.014249833300709724