

# Project Update 2

Ben Jakubowski

Haonan Zhou

# Review from Previous Update

- **Objective:** Predict NYC real estate sale prices and identify real estate comps.

# Review from Previous Update

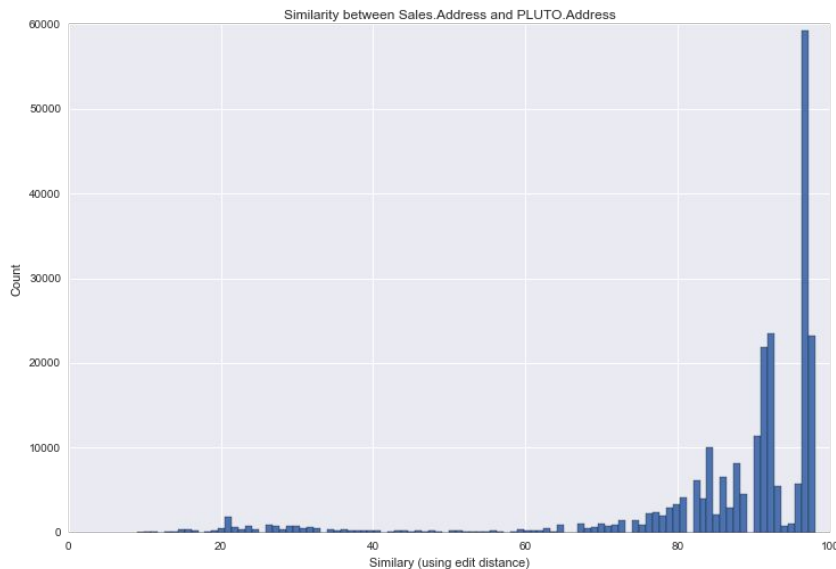
- **Objective:** Predict NYC real estate sale prices and identify real estate comps.
- **Previously work:**

# Review from Previous Update

- **Objective:** Predict NYC real estate sale prices and identify real estate comps.
- **Previously work:**
  - Iterated through current New York City sales pages on Streeteasy and scraped data for 12626 current listings.

# Review from Previous Update

- **Objective:** Predict NYC real estate sale prices and identify real estate comps.
- **Previously work:**
  - Iterated through current New York City sales pages on Streeteasy and scraped data for 12626 current listings.
  - Attempted to merge public datasets (NYC PLUTO and Annualized Sales data) yielding approximately 360k sale records from 2011-2015.
  - This merge failed for condos and still yielded a feature poor dataset.



# New Approach

- Instead of using public civic data (which is feature poor), exclusively focus on scraping Streeteasy sale pages (which are feature rich).

# New Approach

- Instead of using public civic data (which is feature poor), exclusively focus on scraping Streeteasy sale pages (which are feature rich).
- Challenge:

# New Approach

- Instead of using public civic data (which is feature poor), exclusively focus on scraping Streeteasy sale pages (which are feature rich).
- Challenge:



**403 Forbidden**

---

nginx



# New Approach

- Instead of using public civic data (which is feature poor), exclusively focus on scraping Streeteasy sale pages (which are feature rich).
- Challenge:

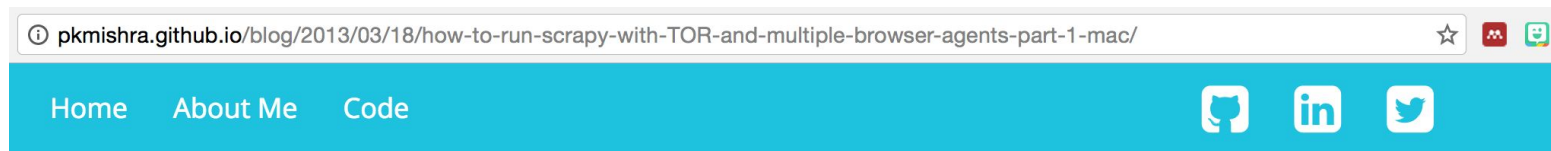


**403 Forbidden**

nginx

Needed to develop scraping strategy to avoid being blocked.

# Solution



## Scrapy: Run Using TOR and Multiple Agents

📅 Mar 18th, 2013

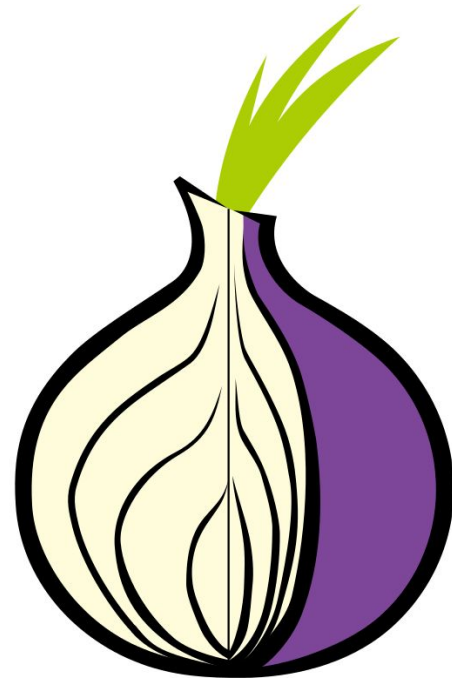
[Scrapy](#) is a brilliant and well documented crawler written in python. Though it is not as scalable as Apache Nutch but it can easily handle thousands of sites easily. You can get up and running very quickly using the official documentation. [Tor](#) gives you power to keep your privacy and security. Tor can hide you so that website can not track your identity. You may read more about TOR in [official site](#). However Tor only works for TCP streams and can be used by any application with SOCKS support.

# Solution



# Scrapy

An open source and collaborative framework  
for extracting the data you need from websites.  
In a fast, simple, yet extensible way.



# StreetEasy Data

- Saved HTML source of all scraped StreetEasy sales pages as text files
  - We chose to save the entire source since we anticipate scraping being a bottleneck, and didn't want to have to re-scrape if we later decided to try additional features from the sale pages.
- Wrote python script to parse the HTML files using BeautifulSoup
- Filtered out all records for properties outside of New York City
- Yielded 372699 records from the current set of 491581 scraped pages
- Current implementation of feature extraction script parses 50000 html pages per hour using a single CPU core
- Trivial to run script in a parallel computing framework to reduce runtime

# StreetEasy Data - Features

45 West 67th Street #26B

**\$2,850,000** FOR SALE

1,086 ft² | \$2,624 per ft² | 4.5 rooms | 2 beds | 2 baths

Condo in [Lincoln Square](#)

- ★ SAVE
- ✉ SHARE
- 🖨 PRINT
- ⚠ PROBLEM?

DAYS ON MARKET

0 days on StreetEasy

MONTHLY CHARGES

Common Charges: \$1,037

Monthly Taxes: \$915

ESTIMATED PAYMENT

Est. Payment: **\$12,190**

## DESCRIPTION

Location and PARK Views!!! Stunning views from this Renovated 2 Bedroom 2 Bath Condo that has Eastern and Southern Exposures. The Windowed Chef's Kitchen, has Stainless Steel Appliances. There are beautiful hardwood floors throughout. The 2nd bedroom has been turned into a den with lots of built ins, 45 West 67th Street is a boutique smoke free condo in the heart of Lincoln Center and steps to Central Park. Great shopping and restaurants await you!

## HIGHLIGHTS

-  Cats Only - No Dogs 
-  Elevator 
-  Full-time Doorman 

# StreetEasy Data - Features

## AMENITIES

### BUILDING AMENITIES

Bike Room Valet  
Concierge  
Laundry in Building  
Live-in Super  
Storage Available

## BUILDING

45 West 67th Street New York, NY 10023

Condo in [Lincoln Square](#)

173 units | Built in 1983

SALES LISTINGS: [2 active and 74 previous](#)

RENTALS LISTINGS: [5 active](#)

DOCUMENTS AND PERMITS: [266 documents](#)

[MORE ABOUT THE BUILDING](#)

## NEARBY

### TRANSPORTATION

## Subways

 at 66th St **under 500 feet**

  at 72nd St **0.19 miles**

   at 72nd St **0.3 miles**

     at 59th St-Columbus Circle **0.56 miles**

  at 81st St **0.57 miles**

[View subway lines on Google Maps](#) ▶

# StreetEasy Data - Features

List of features extracted from HTML pages:

- saleid
- address
- price
- neighborhood
- borough
- status
- date
- num\_beds
- num\_baths
- num\_sqft
- type
- url
- monthly\_cost
- amenities\_list
- transit\_list
- gps\_coordinates
- school\_district
- building\_name
- built\_date
- building\_num\_units
- building\_url
- description

# StreetEasy Data - Sales Data

Terminal state of HTML pages:

- sold 187975
- no\_longer\_available 151486
- delisted 23249
- temporarily\_off\_market 5461
- current 1724
- expired\_owner\_listing 1472
- in\_contract 1332



# StreetEasy Data - Sales Data

## Number of listings by borough:

- manhattan 205510
- brooklyn 100874
- queens 36325
- staten\_island 17528
- bronx 12462

## Most popular neighborhoods:

- upper\_west\_side 19253
- upper\_east\_side 14007
- lincoln\_square 13441
- lenox\_hill 12648
- yorkville 11577

# StreetEasy Data - Sales Data

Most popular types of listings:

- co\_op 121996
- condo 96294
- resale 40517
- sponsor\_unit 33640
- house 25117
- multi\_family 23104
- townhouse 17007
- apartment 4863
- condop 4014

# StreetEasy Data - Sales Data

Listings bucketed by square footage:

- NaN 126549
- 1000 102990
- 2000 92580
- 3000 27836
- 4000 12493
- 5000 4882
- 6000 1917
- 7000 1179
- 8000 604
- 9000 453
- 10000 257
- 11000 204
- 12000 124

# StreetEasy Data - Sales Data

Listings bucketed by price per square foot:

- 250 20398
- 500 45083
- 750 45768
- 1000 43576
- 1250 34316
- 1500 21391
- 1750 12489
- 2000 7566
- 2250 4687
- 2500 2975
- 2750 2116
- 3000 1365
- 3250 993
- 3500 712

# StreetEasy Data - Sales Data

Listings in Manhattan bucketed by price per square foot:

- 250 741
- 500 5580
- 750 20308
- 1000 32229
- 1250 29847
- 1500 19958
- 1750 12045
- 2000 7416
- 2250 4630
- 2500 2953
- 2750 2088
- 3000 1355
- 3250 989
- 3500 707

# Next Steps

- **Featurization:** Examples include:

## DESCRIPTION

Location and PARK Views!!! Stunning views from this Renovated 2 Bedroom 2 Bath Condo that has Eastern and Southern Exposures. The Windowed Chef's Kitchen, has Stainless Steel Appliances. There are beautiful hardwood floors throughout. The 2nd bedroom has been turned into a den with lots of built ins, 45 West 67th Street is a boutique smoke free condo in the heart of Lincoln Center and steps to Central Park. Great shopping and restaurants await you!

---

## HIGHLIGHTS



Cats Only - No Dogs 



Elevator 



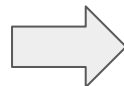
Full-time Doorman 

# Next Steps

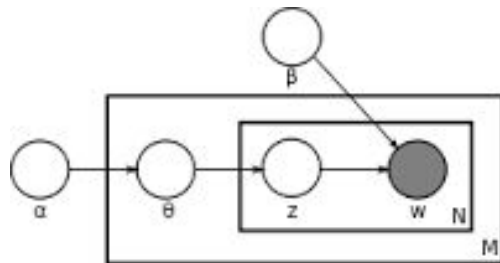
- **Featurization:** Examples include:

## DESCRIPTION

Location and PARK Views!!! Stunning views from this Renovated 2 Bedroom 2 Bath Condo that has Eastern and Southern Exposures. The Windowed Chef's Kitchen, has Stainless Steel Appliances. There are beautiful hardwood floors throughout. The 2nd bedroom has been turned into a den with lots of built ins, 45 West 67th Street is a boutique smoke free condo in the heart of Lincoln Center and steps to Central Park. Great shopping and restaurants await you!



Use LDA to preprocess description, optimizing K (number of topics) as hyperparameter).



## HIGHLIGHTS



Cats Only - No Dogs ✓



Elevator ✓



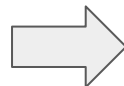
Full-time Doorman ✓

# Next Steps

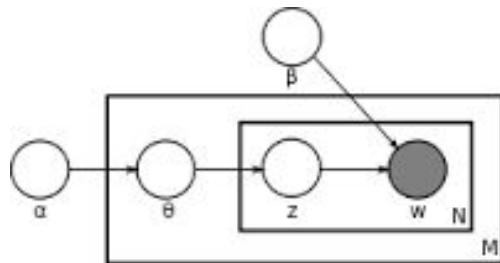
- **Featurization:** Examples include:

## DESCRIPTION

Location and PARK Views!!! Stunning views from this Renovated 2 Bedroom 2 Bath Condo that has Eastern and Southern Exposures. The Windowed Chef's Kitchen, has Stainless Steel Appliances. There are beautiful hardwood floors throughout. The 2nd bedroom has been turned into a den with lots of built ins, 45 West 67th Street is a boutique smoke free condo in the heart of Lincoln Center and steps to Central Park. Great shopping and restaurants await you!



Use LDA to preprocess description, optimizing  $K$  (number of topics) as hyperparameter).



---

## HIGHLIGHTS



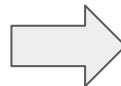
Cats Only - No Dogs ✓



Elevator ✓



Full-time Doorman ✓



One-hot encodings of highlights/other categorical features.



# Next Steps

- **Modeling:**
  - **Model 1:**
    - Model sale prices citywide, using geographic features as predictors to allow for geographic effects in model.
    - Augment obvious features (i.e. number of bedrooms, square feet) with topics from LDA
    - Use nonlinear models (random forests, boosted regression trees) to allow for complex interactions.

# Next Steps

- **Modeling:**
  - **Model 1:**
    - Model sale prices citywide, using geographic features as predictors to allow for geographic effects in model.
    - Augment obvious features (i.e. number of bedrooms, square feet) with topics from LDA
    - Use nonlinear models (random forests, boosted regression trees) to allow for complex interactions.
  - **Model 2:**
    - Model each neighborhood separately using simple linear regression model.
    - If competitive, this approach will allow for comparisons between neighborhoods' sale price sensitivity to different predictors.