


Humans of New York

Topic Modeling of Instagram Posts

Matt Barrett, Tim Lai, Christine Mulcahy, Elena Reynolds, and Brett Scroggins

Social Media Analytics
Spring 2018




Motivation

Our project explored the ‘Humans of New York’ Instagram account and its relationship between topics presented and interaction from fans.

Data Description

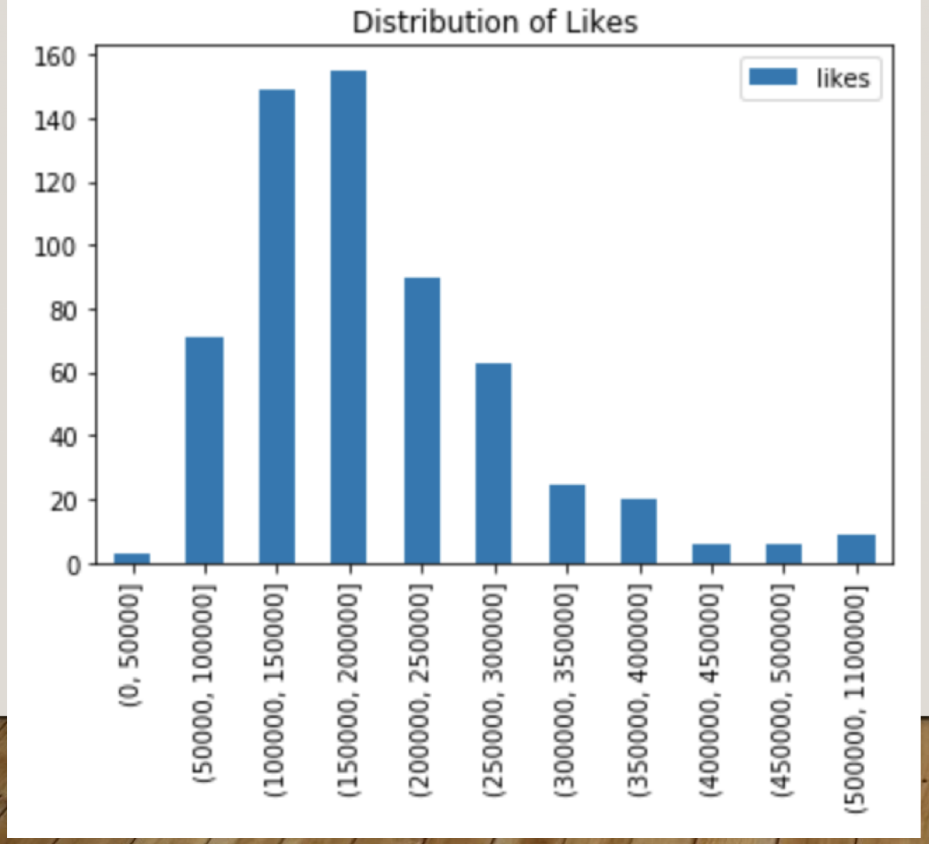
We used the Instagram_Vision.py to scrape 598 unique images posted by Humans of New York within the last two years.

From the output, we extracted:

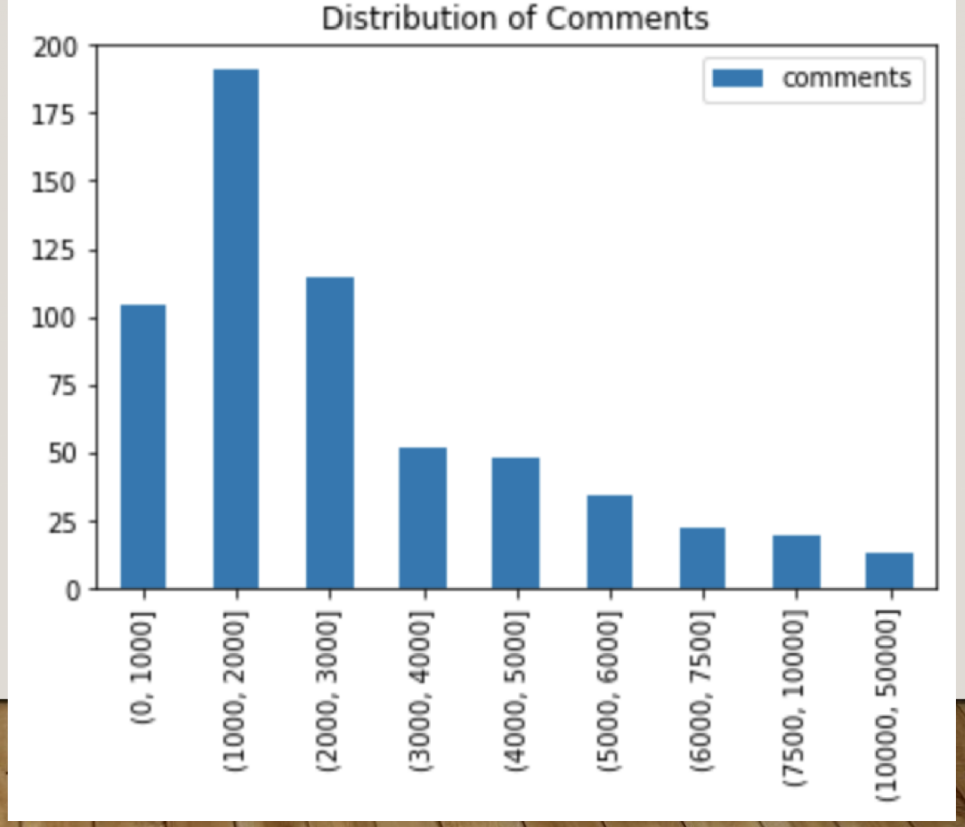
- Number of likes
- Number of comments
- Post location
- “Story” words to use in topic modeling
- Google image labels to use in topic modeling

Variance of Likes and Comments

Mean of Likes = 195492
Median of Likes = 177000
Standard Deviation of Likes = 102398



Mean of Comments = 2923
Median of Comments = 2016
Standard Deviation of Comments = 3256



Geographic Locations

Top Locations:

1. NYC
2. Mumbai
3. Jaipur
4. São Paolo
5. Rio de Janeiro

Topic Modeling from “Stories”

Topic 0: Pet care

Topic 1: Daily routine

Topic 2: Health

Topic 3: Separation anxiety

Topic 4: Business travel

Topic 5: Milestones

Topic 6: Mourning

Topic 7: Fathers/kids

Topic 8: Jobs

Topic 9: Mom/kids

Topic 10: Mothers/school

Topic 11: Struggle

Topic 12: Love/family/parents

Topic 13: Sons/bravery

Topic 14: Marriage

Analysis of Topics from “Stories”

Likes Regression

Likes = 186408.8 + 5854.9*T0 + (-13525.9)*T1 + (-22933.6)*T2 + 11551.4*T3 + 11670.6*T4 + (-19903.5)*T5 + 7751.8*T6 + 12748.5*T7 + (-10224.8)*T8 + 26662.4*T9 + 32074.5*T10 + (-12265.3)*T11 + 27366.3*T12 + 2406.2*T13 + 10248.4*T14

- RMSE = 99,353.4

Comments Regression

Comments = 3062.2 + 596.3 *T0 + (-1223.7)*T1 + (-210.3)*T2 + 220.2*T3 + (-562.9)*T4 + (-756.6)*T5 + (-647.5)*T6 + 448.8*T7 + (-482.9)*T8 + 85.6*T9 + 201.6*T10 + (-469.7)*T11 + 150.6*T12 + 1650.5*T13 + 381.0*T14

- RMSE = 3,186.2

Topic Modeling from Google Cloud Vision Image Labels

Topic 0: Black Tie

Topic 1: Traffic

Topic 2: Food + Drink

Topic 3: Dogs

Topic 4: Eyewear

Topic 5: Outdoors

Topic 6: Beach

Topic 7: Transportation

Topic 8: Technology

Topic 9: Elderly

Topic 10: Marketplace

Topic 11: Fashion

Topic 12: Socializing

Topic 13: Business

Topic 14: Happiness

Analysis of Topics from Google Cloud Vision image labels

Likes Regression

Likes = 206235.8 + 70010.4*T0 + (-10029.6) *T1 + 17458.9*T2 + 44863.7 *T3 + (-3534.6)*T4 + (-14569.9)*T5 + (-23271.3) *T6 + (-29999.1)*T7 + (-68204.5)*T8 + (-7530.9)*T9 + (-5435.6)*T10 + 14356.5*T11 + (-23949.6)*T12 + (-22760.4)*T13 + 30276.3*T14

- RMSE = 97,787.0

Comments Regression

Comments = 3277.8 + (-284.5)*T0 + (-502.7)*T1 + (-1044.1)*T2 + 1919.2*T3 + (-520.6)*T4 + (-424.8)*T5 + 767.8*T6 + 181.9*T7 + (-1217.6)*T8 + (-54.1)*T9 + (-836.3)*T10 + 891.8*T11 + 579.0*T12 + (-131.0)*T13 + (-99.9)*T14

- RMSE = 3,186.7

Next Steps

- **Scrape all remaining posts**
- **Unexplored factors that might influence the model:**
  - Change in followers over time
  - Day of week/time of day posted
  - Length of the caption
  - Location of post
  - Current events that influence the public's emotional reaction
- **Possible changes to the target variable:**
  - Regress on a combination of likes and comments together
  - Filter out meaningless comments (e.g. "@barua" - tagging a friend)