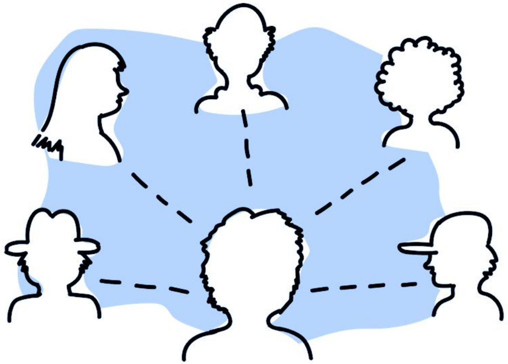
Instagram Hashtag Recommender System

Kaci Kus

Motivation

Instagram is the 3rd most popular social network

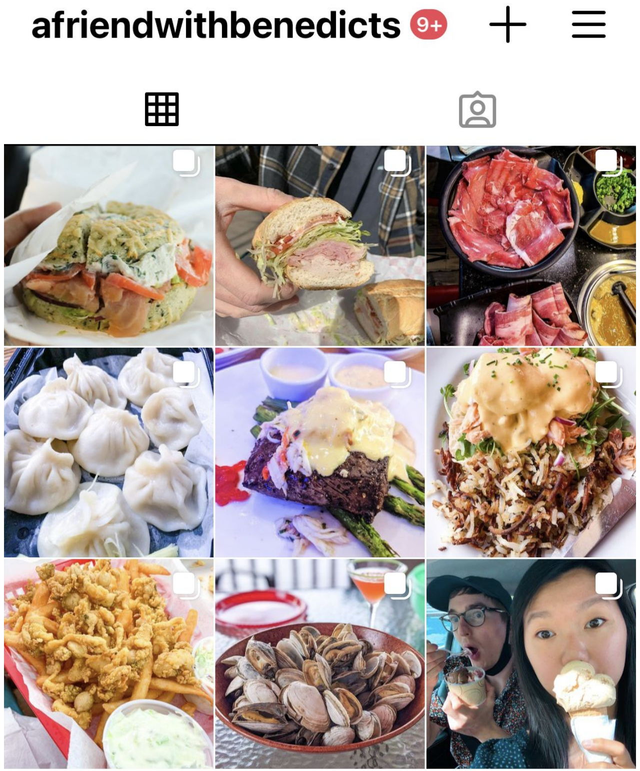
- >1 BILLION monthly active users
- Allows users to reach a wide audience and grow their following
    - Sponsorships
    - Customer acquisition



Motivation

Personal interest:

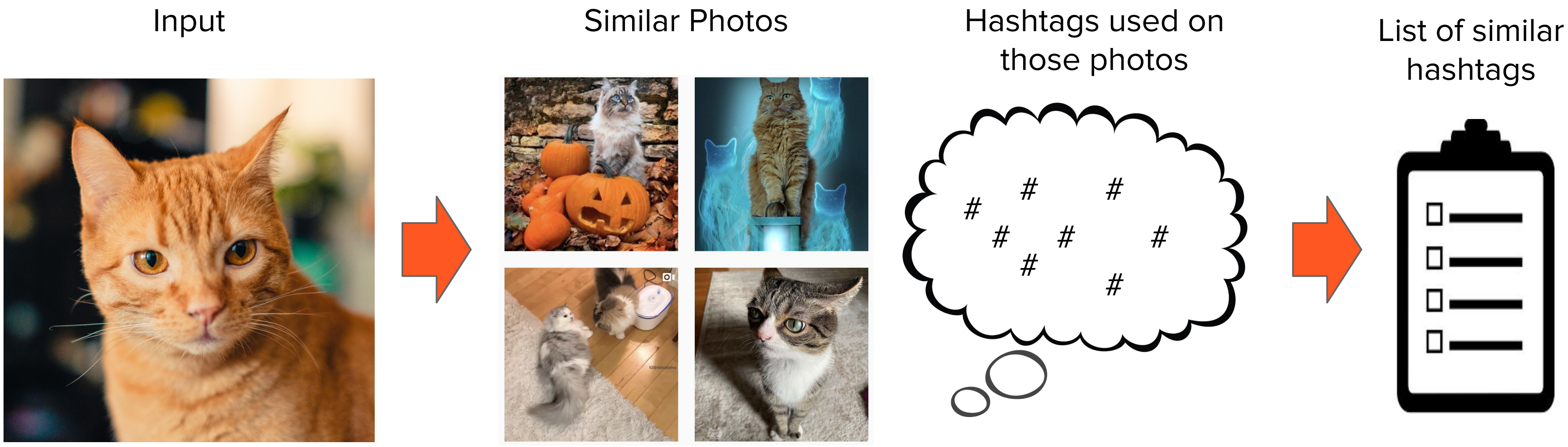
- I run an active food-review account where hashtags are critical for gaining new followers
- Currently keep a static list of general brunch related hashtags that I copy-and-paste
  - Inconvenient!
- Each post uses the same hashtags and is not post-specific



Goal

A recommender system would:

1. Take out the guesswork by automating the process of choosing hashtags
2. Give you confidence that you are choosing the optimal hashtags for your post



Input

Similar Photos

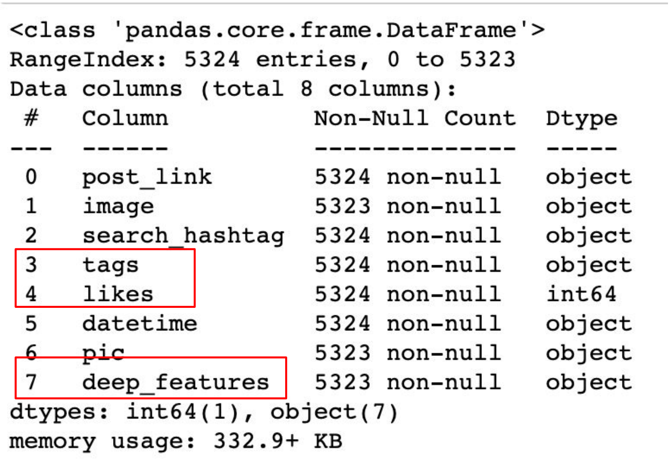
Hashtags used on those photos

List of similar hashtags

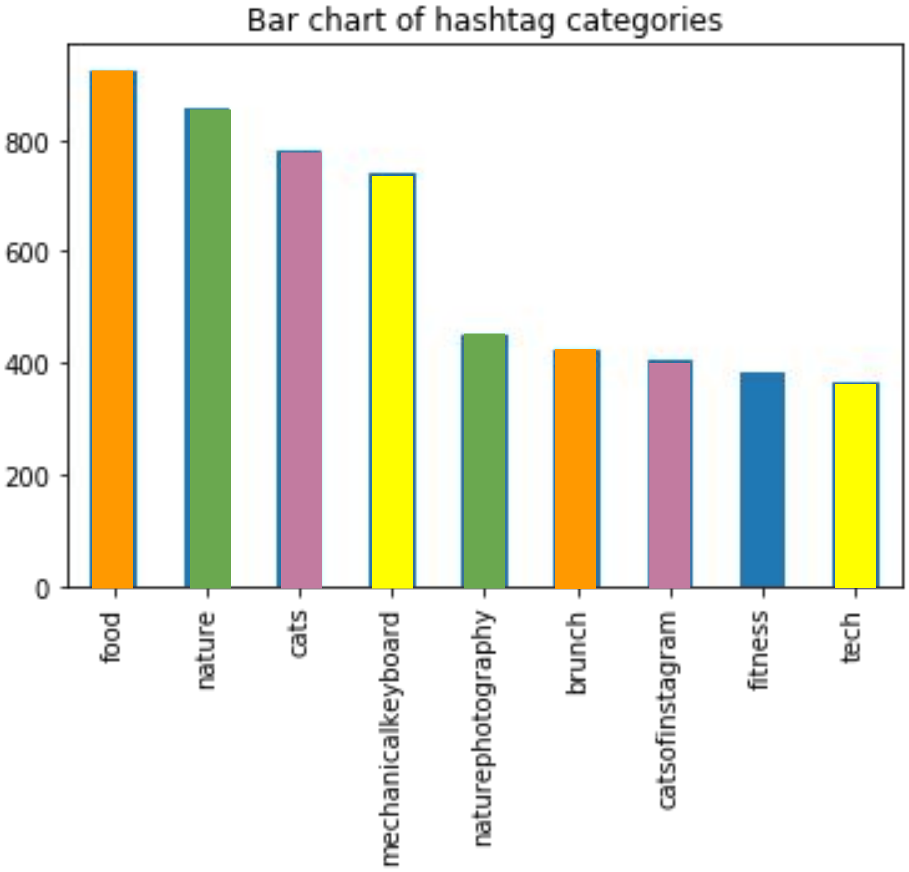
My cat, Parm!

Dataset

- 5,324 posts from instagram
- Function to scrape the post meta-data for most recent posts using a search hashtag
- Function to extract deep features using MobileNetV2 pre-trained neural network (by Google)



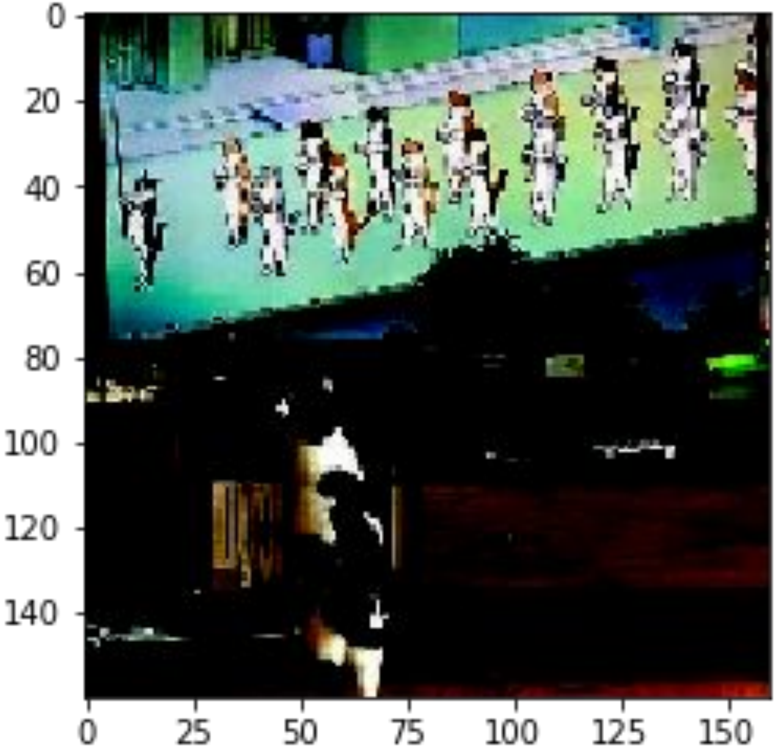
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5324 entries, 0 to 5323
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   post_link       5324 non-null   object
 1   image           5323 non-null   object
 2   search_hashtag  5324 non-null   object
 3   tags            5324 non-null   object
 4   likes           5324 non-null   int64
 5   datetime        5324 non-null   object
 6   pic             5323 non-null   object
 7   deep_features   5323 non-null   object
dtypes: int64(1), object(7)
memory usage: 332.9+ KB
```



Data Exploration

Example of one post

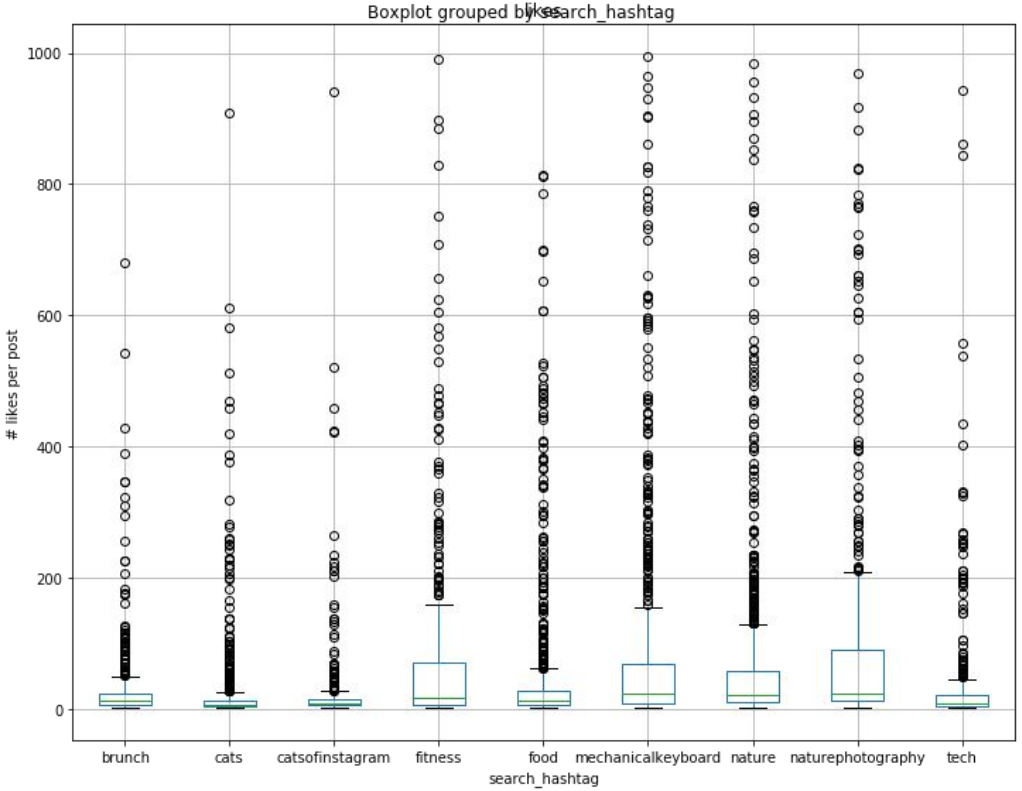
- Search hashtag: #Cats
- 459 Likes
- 27 hashtags used
- 'cats_of_instagram', 'catstagramcat', 'admiremycat', 'ilovemycat', 'catsogram', 'catsofig', 'catscatscats', 'pamperedcats', 'cats_features', 'persiancats', 'cats_of_instworld', 'cats_of_ig', 'bestmeow', 'cats', 'adorablecathttps', 'thedailykitten', 'purr', 'cats_of_day', 'excellent_cats', 'cat', 'catstagrams', 'fluffycat', 'catsuit', 'cats_of_the_globe', 'catstagram', 'cats_of_insta', 'catstgram'



Distribution of likes by search hashtag

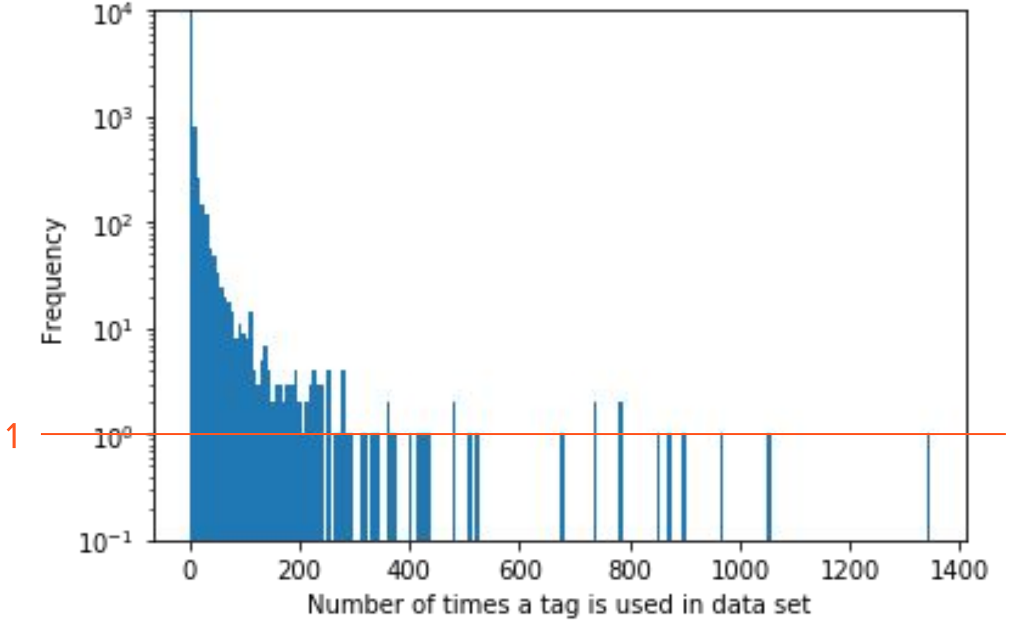
- #fitness, #mechanicalkeyboard, #nature, and #naturephotography seem to earn slightly higher median likes
- Most categories have similar ranges

Keep in mind that we don’t have any data on how many followers each account that posts has



Hashtags

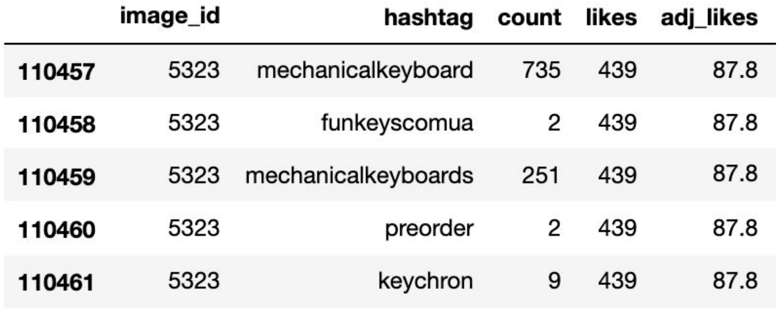
- 5,324 posts
- 110,462 hashtags total
- 31,562 unique hashtags used
- Each posts uses on avg. ~21 hashtags
- The majority of hashtags are actually used very few times!



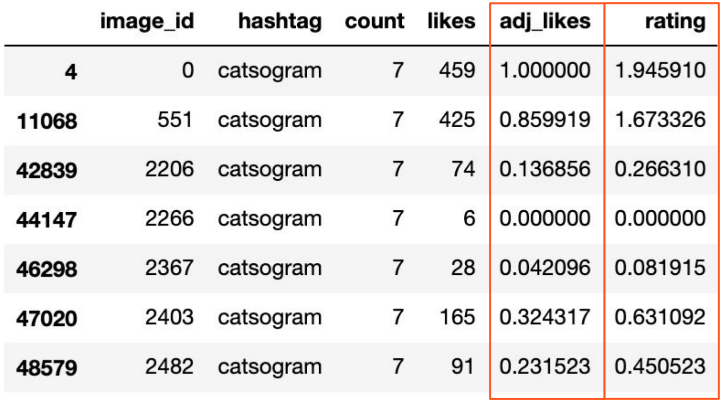
Creating a metric to “rate” hashtags

- Likes on a post can range anywhere from 1 to >800

1. Divide the number of **likes** on a post by the **number of hashtags used** to get **adj_likes**
2. For each unique hashtag, **normalize** all of the **adj_likes** from 0 to 1
3. Multiply the **log** of the **count** (number of times a hashtag is used) by the **adj_likes** to get final **rating**



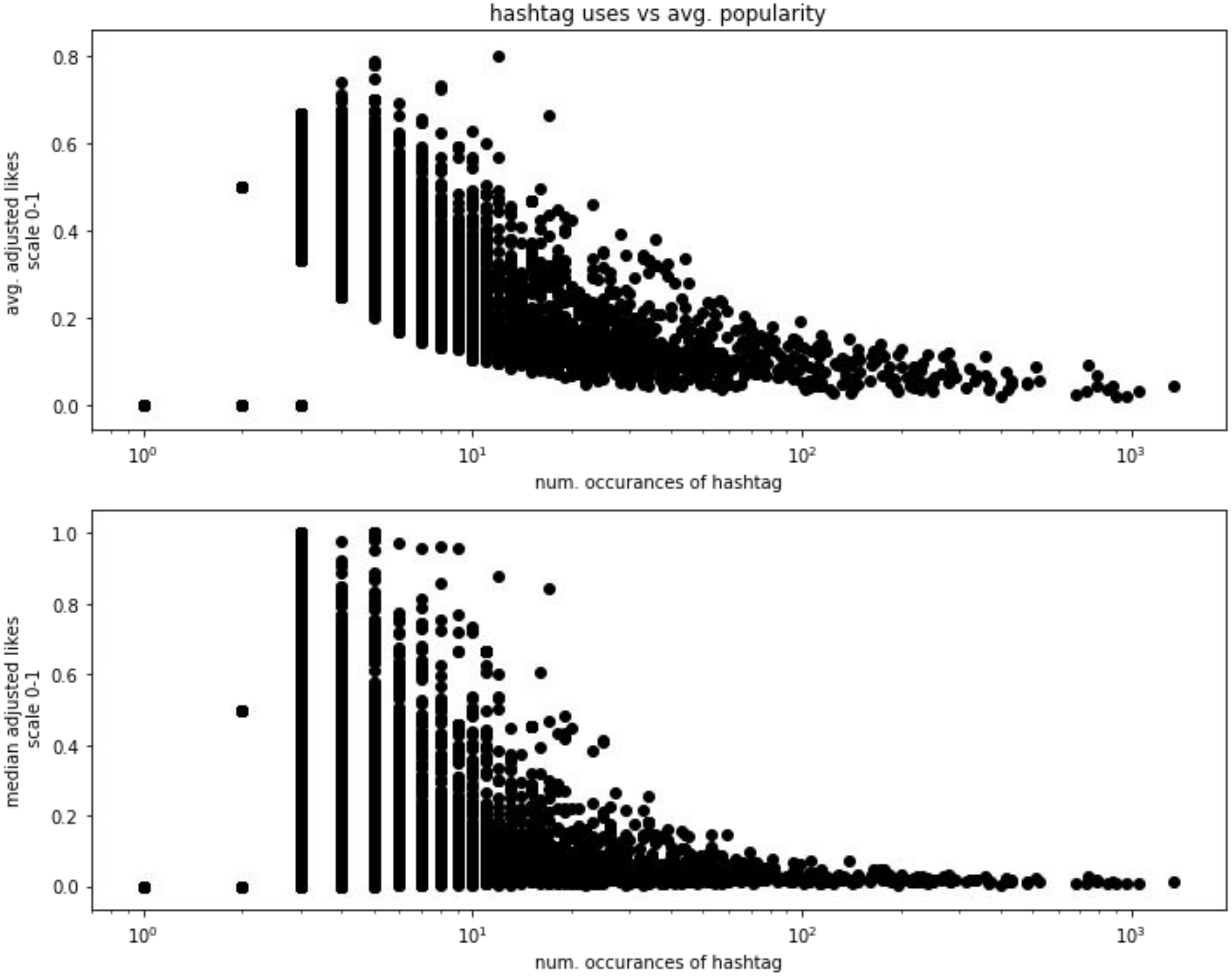
	image_id	hashtag	count	likes	adj_likes
110457	5323	mechanicalkeyboard	735	439	87.8
110458	5323	funkeyscomua	2	439	87.8
110459	5323	mechanicalkeyboards	251	439	87.8
110460	5323	preorder	2	439	87.8
110461	5323	keychron	9	439	87.8



	image_id	hashtag	count	likes	adj_likes	rating
4	0	catsogram	7	459	1.000000	1.945910
11068	551	catsogram	7	425	0.859919	1.673326
42839	2206	catsogram	7	74	0.136856	0.266310
44147	2266	catsogram	7	6	0.000000	0.000000
46298	2367	catsogram	7	28	0.042096	0.081915
47020	2403	catsogram	7	165	0.324317	0.631092
48579	2482	catsogram	7	91	0.231523	0.450523

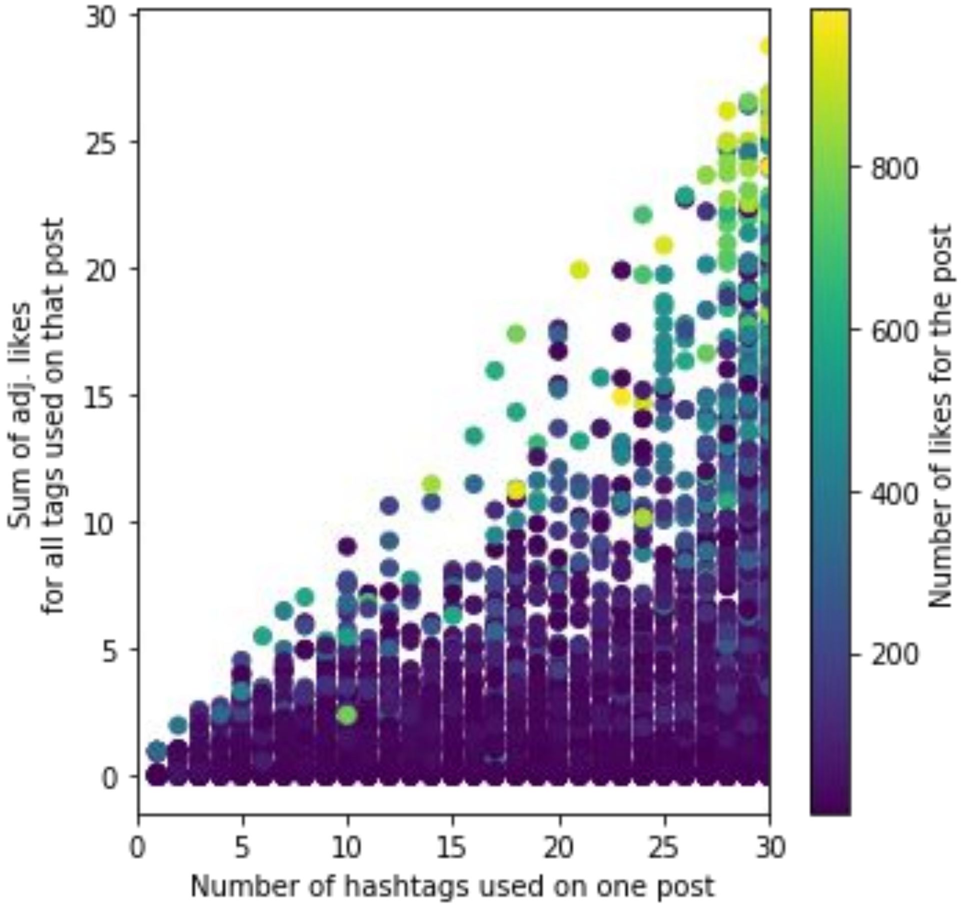
Hashtags

Hashtags that occur more frequently actually tend to have a lower number of likes associated with them!



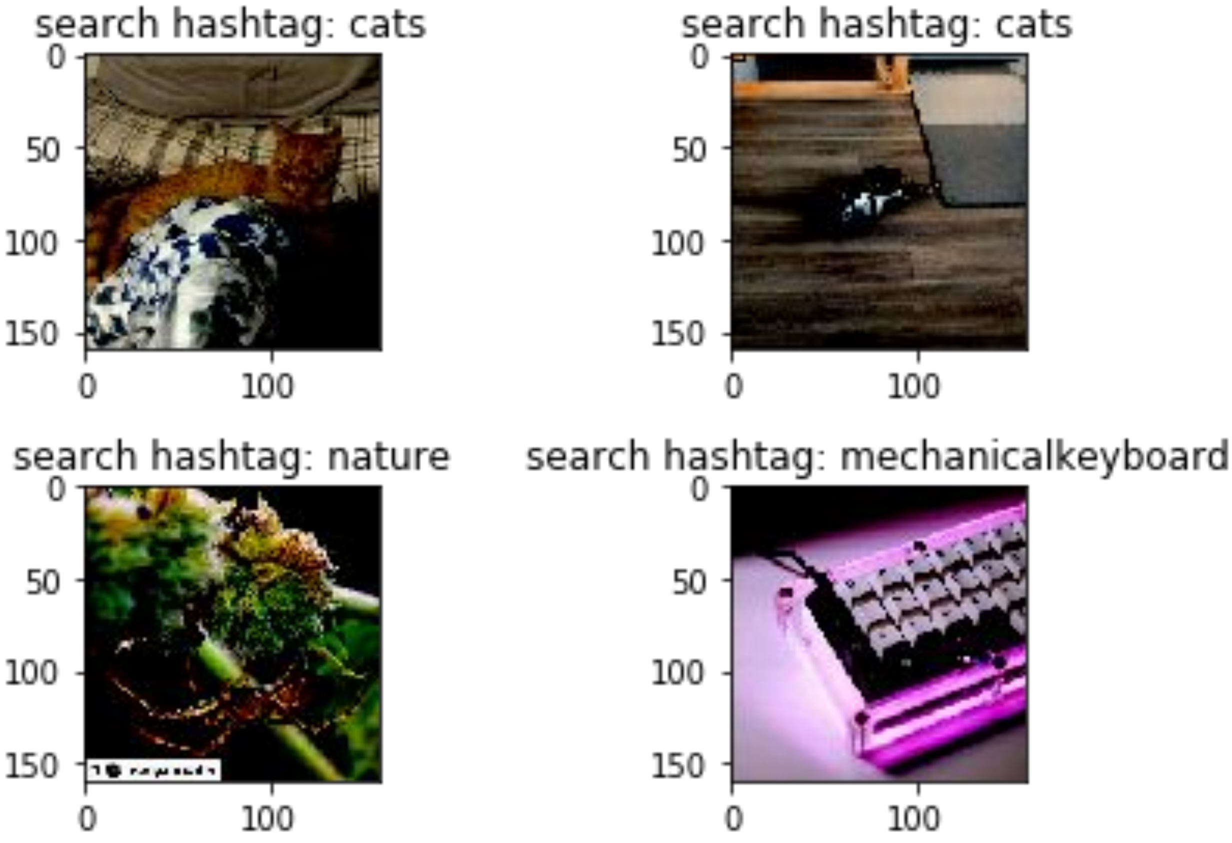
Hashtags

- Posts that use more hashtags are the only ones that tend to get >500 likes on a post
- Posts that use 30 hashtags can still get a very low number of likes though



Post pictures

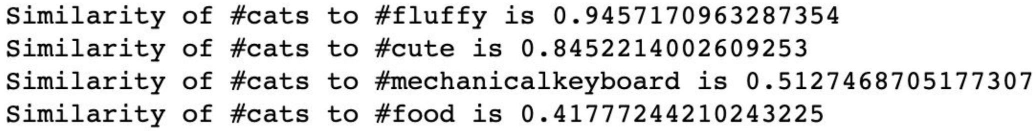
- The search hashtags seem pretty accurate in what they depict
- There are many instances though in categories such as #fitness and #tech where the photos are not at all related
- This is not a problem for us though



Modeling: Recommender System

Continuous Bag of Words (CBOW)

- Predicts target word by looking at words around it, for example:
    - “Today is a [...] day”
- Find top 10 hashtags most similar to #cats:
    - Catsofinstagram
    - Cat
    - Catlover
    - Kitten
    - Kitty
    - Catlife
    - Catlovers
    - Meow
    - Instacat
    - Catstagram

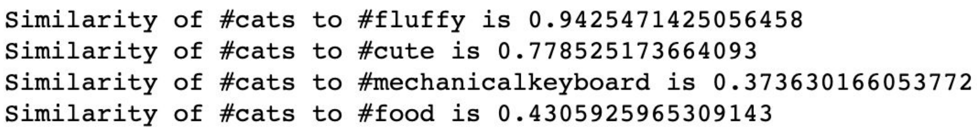


```
Similarity of #cats to #fluffy is 0.9457170963287354
Similarity of #cats to #cute is 0.845221400260925 3
Similarity of #cats to #mechanicalkeyboard is 0.5127468705177307
Similarity of #cats to #food is 0.4177724421024322 5
```

- Scores #cats and #fluffy as very similar!
- Scores #cats and #cute as very similar as well
- Recognizes that #food and #mechanicalkeyboard are not similar to #cats

Skip-gram

- Predicts words that would be around a target word
- Find top 10 hashtags most similar to #cats:
    - Gato
    - Instacat
    - Meow
    - Catlover
    - Catlovers
    - Catstagram
    - Catlove
    - Cutecat
    - Kittens
    - Tabbycat



```
Similarity of #cats to #fluffy is 0.9425471425056458
Similarity of #cats to #cute is 0.778525173664093
Similarity of #cats to #mechanicalkeyboard is 0.373630166053772
Similarity of #cats to #food is 0.4305925965309143
```

- Still predicts #cute and #fluffy as very similar to #cats
- Scores #mechanicalkeyboard and #food even lower than the CBOW model

Basic recommender system

- Find the top 20 most similar hashtags for each and every unique hashtag in our dataset
- Store in dataframe
- This allows a user to input any hashtag from our dataset and find the most similar tags very quickly, as they are all pre-calculated

Downside:

- Requires starting point hashtag
- Not very personalized to the post

Example:

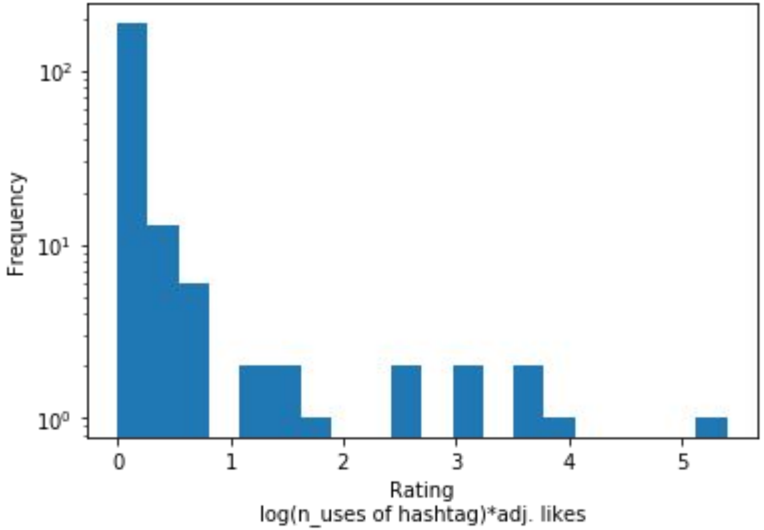
hashtag: 35mm

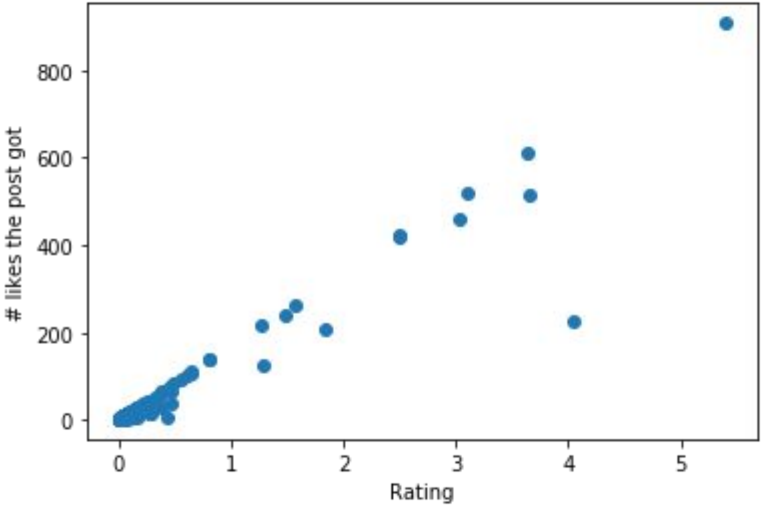
top 10 recommended hashtags: ['capture', 'artofvisuals', 'canonphotography', 'nikonphotography', 'botanical', 'photographylovers', 'mobilephotography', 'flowerphotography', 'birds_captures', 'natgeo']

Skip-gram seems promising, but could use improvement...

- Since it uses surrounding words for context, order matters
  - Randomly shuffle the order of hashtags listed on a post
- Optimize recommendations to suggest hashtags that tend to get the most likes
  - Use “weights” from the adjusted likes (previously calculated)
- Incorporate photo-identification
  - Use deep-features

Rate each use of each hashtag

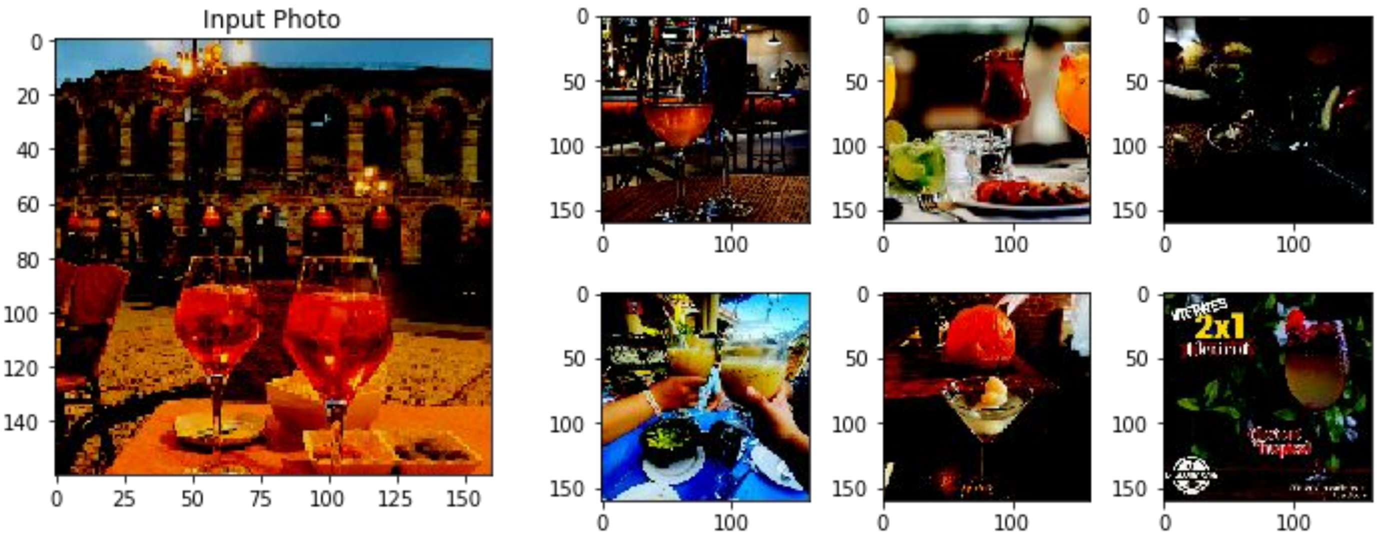




- Take the log of the number of times the hashtag is used in the dataset
    - Hashtags only used once will end up with scores of 0
- Multiply by the adjusted like score

- #cats_of_instagram
- Strong correlation with the number of likes a post gets and the rating

K-Nearest Neighbors Model



```
('quedateencasa', 0.9701948165893555), ('patio', 0.9639737606048584), ('dulces', 0.9609004259109497),
('event', 0.9601226449012756), ('trevlighelg', 0.9599580764770508), ('chips', 0.9584953784942627),
('election', 0.9576938152313232), ('saturday', 0.9557901620864868), ('fooddelivery', 0.9557449817657471),
('charcuterie', 0.9556083679199219)
```

- It looks like the hashtags identified are similar to the picture. Travel, patio, event, chips, saturday, charcuterie, and fooddelivery all seem relavent enough.

Compare Weighted Recommendations

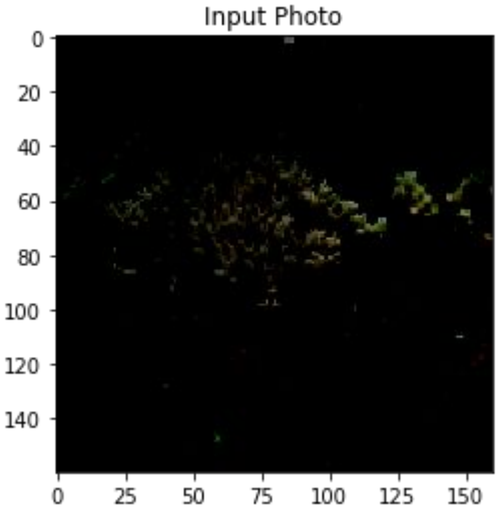
Similarity

patio, median likes: 19.0
event, median likes: 9.5
election, median likes: 22.0
charcuterie, median likes: 14.0
dulces, median likes: 7.0
chips, median likes: 11.0
fooddelivery, median likes: 6.5
saturday, median likes: 12.0
quedateencasa, median likes: 9.0
trevlighelg, median likes: 12.0
The average number of median likes
these hashtags get is: 12.2

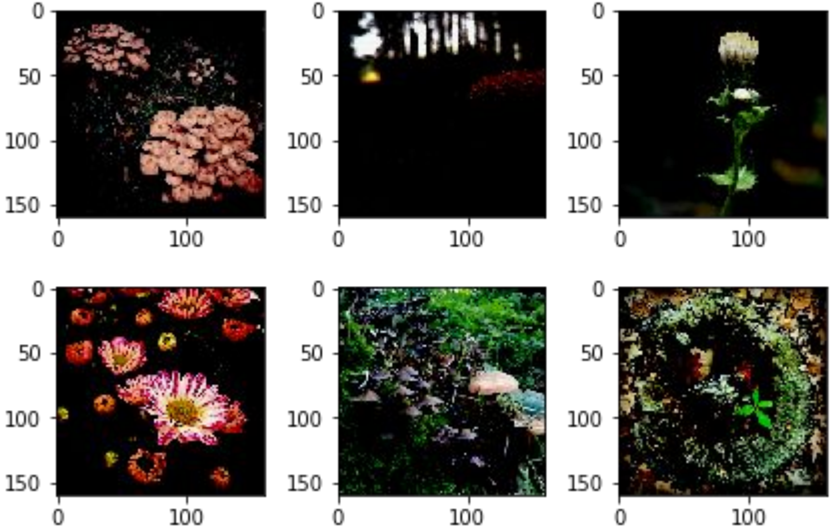
Optimized for likes

livemusic, median likes: 14.0
aperitivo, median likes: 15.5
patio, median likes: 19.0
カレー, median likes: 11.0
manchester, median likes: 7.0
eventos, median likes: 26.0
bar, median likes: 8.0
event, median likes: 9.5
yemek, median likes: 55.0
churrasco, median likes: 22.0
The average number of median likes
these hashtags get is: 18.7

Nature example



6 most similar photos from dataset



The deep features have done well in recognizing similar green/floral photos

Optimized tags perform better

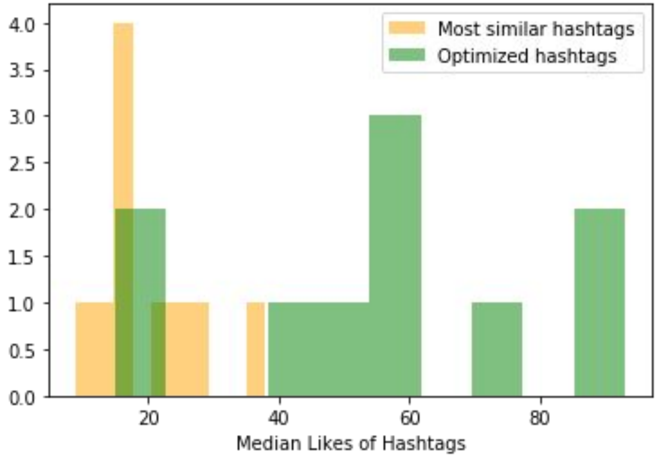
Similarity

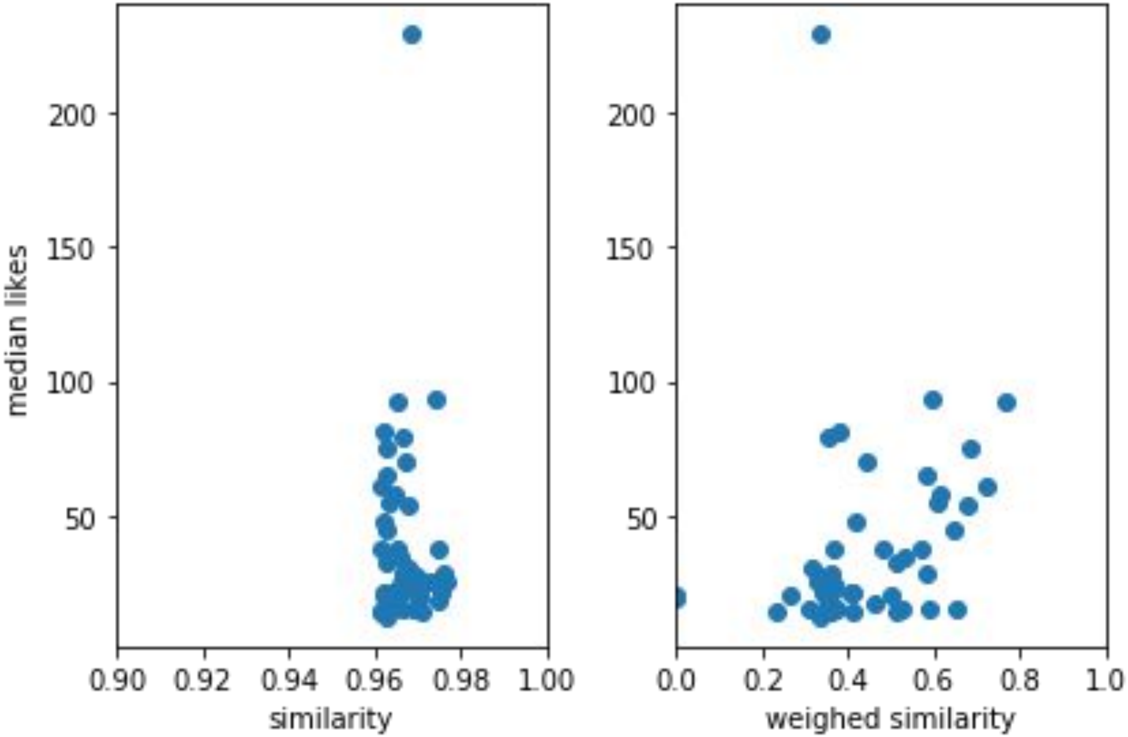
findyourpark, median likes: 22.0
syksy, median likes: 16.0
tatry, median likes: 28.5
countryliving, median likes: 13.5
tb, median likes: 38.0
visit_serbia, median likes: 9.0
naturalbeauty, median likes: 24.5
valokuva, median likes: 16.0
instaphotography, median likes: 15.0
oceanlover, median likes: 17.0
The average number of median likes these hashtags get is: 19.95

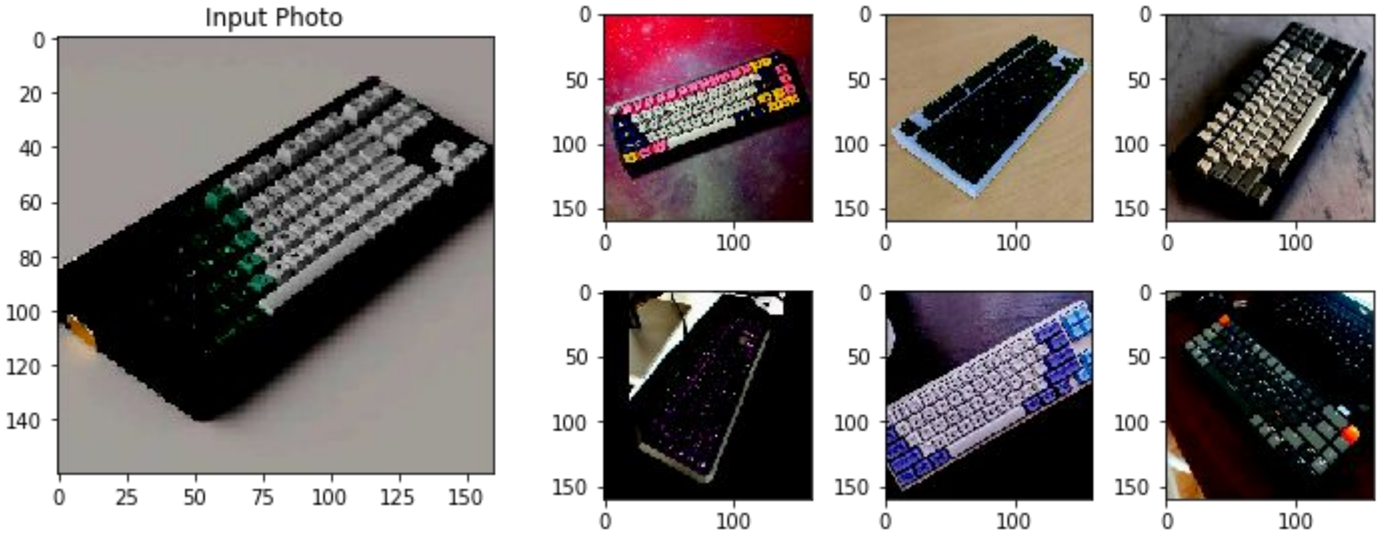
Optimized for likes

natureshot, median likes: 92.0
discoverearth, median likes: 61.0
island, median likes: 75.0
nationalgeographic, median likes: 53.5
countryside, median likes: 15.0
roamtheplanet, median likes: 45.0
deutschland, median likes: 58.0
sonnenuntergang, median likes: 55.0
landscape_captures, median likes: 93.0
tourism, median likes: 15.0
The average number of median likes these hashtags get is: 56.25

backup


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
40
60
80
Median Likes of Hashtags
Most similar hashtags
Optimized hashtags


median likes
200
150
100
50
0.90
0.92
0.94
0.96
0.98
1.00
similarity
200
150
100
50
0.0
0.2
0.4
0.6
0.8
1.0
weighed similarity



```
('gamersofinstagram', 0.9750460386276245), ('kbd', 0.9723407626152039), ('mecha', 0.9714288711547852), ('engineeringjobs',
0.9694418907165527), ('desksetup', 0.9689429402351379), ('redswitch', 0.9681743383407593), ('womenintech', 0.9680366516113281),
('cnc', 0.9677553772926331), ('gamingkeyboards', 0.9671439528465271), ('duckykeyboard', 0.9661998152732849)
```

Similarity

redswitch, median likes: 21.5
mecha, median likes: 54.0
cnc, median likes: 49.0
kbd, median likes: 9.0
desksetup, median likes: 29.0
gamingkeyboards, median likes: 14.0
gamersofinstagram, median likes: 19.5
duckykeyboard, median likes: 16.0
womenintech, median likes: 13.5
engineeringjobs, median likes: 16.0
The average number of median likes these hashtags get is: 24.15

Optimized

keyboardcable, median likes: 35.5
groupbuy, median likes: 43.0
redswitch, median likes: 21.5
audiophile, median likes: 10.5
battlestations, median likes: 63.0
customcables, median likes: 32.5
mecha, median likes: 54.0
geekcable, median likes: 40.0
novelkeys, median likes: 18.5
cnc, median likes: 49.0
The average number of median likes these hashtags get is: 36.75