Web Scraping - Instagram

Nelson Chen

Agenda

- Introduction
- Web scraping
- Image Classification (Google and Clarifai's software)
- Exploratory Analysis of Data
- Conclusion
- Next steps

Introduction

- A picture is worth a thousand words!
- Mobile camera technology created surplus of image data
- Pictures are integral part of social media like **Instagram**
- Goal: develop workflow to analyze instagram accounts in hopes to optimize *likes* and *comments*

Web Scraping

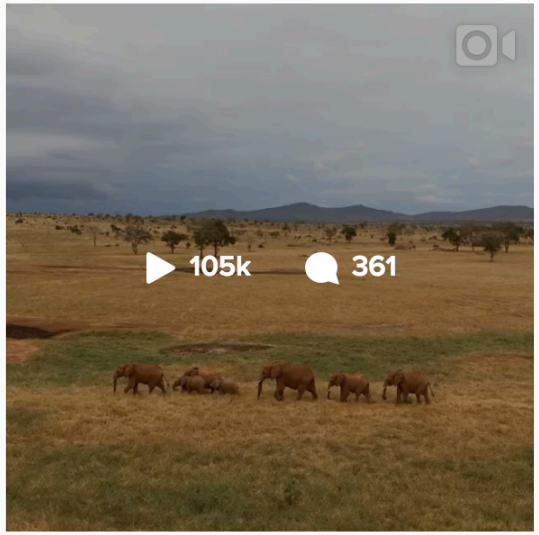
- Instagram API: complex and restricted to brands, advertisers, etc.
- Instagram official site is dynamically loaded, so instead scraped proxy site: **imgrum.net**

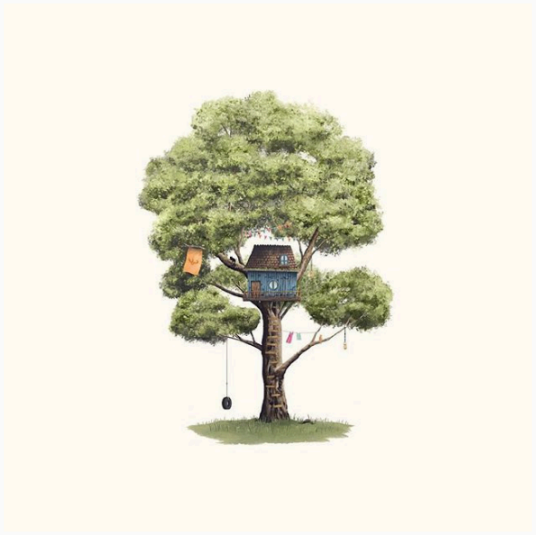
Instagram

instagram Follow

3,808 posts **199m** followers **191** following

Instagram Discovering — and telling — stories from around the world. Curated by Instagram's community team. **blog.instagram.com**





Web Scraping

- Instagram API: complex and restricted to brands, advertisers, etc.
- Instagram official site is dynamically loaded, so instead scraped proxy site: **imgrum.net**
- Imgrum loads instagram files from API and displays them in simple web format
- Used Scrapy to scrape @Instagram for all **image urls**, **number of likes**, and **number of comments**
- Downloaded images using python library urllib

Imgrum

imgrum | an online Instagram web viewer

Search the Instagram...

INSTAGRAM

@INSTAGRAM

Images by instagram

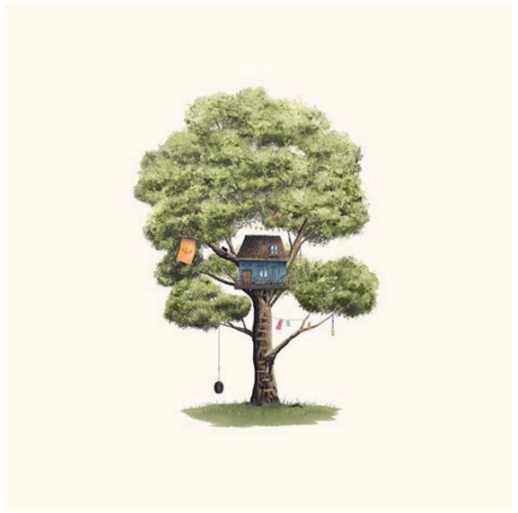
@instagram

A parade of elephants in lockstep — @koraninst's #WHPinmotion submission from Kenya captures the rhythm of life in the wild. 🐘 #Boomerang by @koraninst

351 ♡36044

@instagram

In a single frame, @misteur_z captured three simultaneous jumps that feel like separate stages of an airborne flip — no small feat for the photographer or the performers. The goal of #WHPinmotion was to capture movement in creative ways. Follow along as we feature more of our favorite submissions.



@instagram

It's the little things that Sam Lyne (@samlyne) enjoys most in his illustrations. "I hope to make pieces that people can explore," says the 27-year-old resident of Tasmania's capital, Hobart. "I really enjoy creating works that are sometimes minimal and simplistic in nature, but also jampacked with hidden details." He traces this

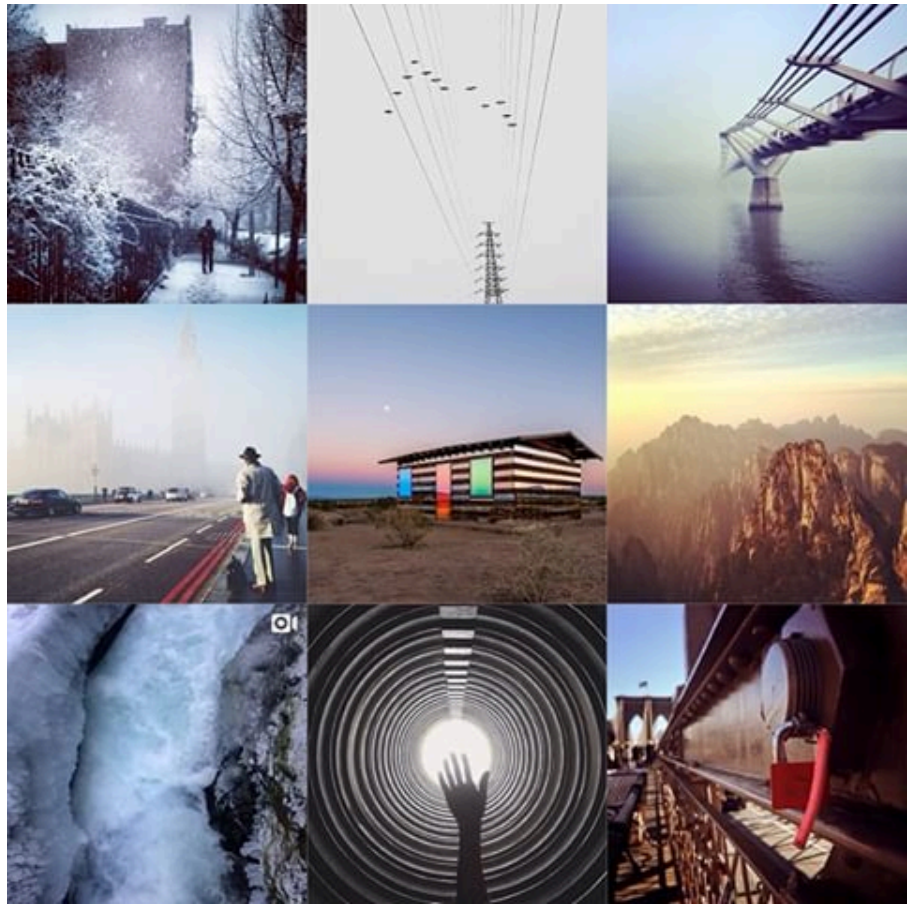
Prebuilt Image Classifiers

- Google's Inception v3
  - Image classifier from ImageNet competition released in 2015
  - Uses deep Convolutional Neural Networks (CNN)
  - Classifies images into 1000 categories trained on over 1 million images
- Clarifai's object/region classification
  - Uses Regions with Convolutional neural networks (R-CNN)
  - Separates image to regions/objects and classifies each individually with CNN

Google's Inception Image Classifier

- Lion, king of beats, panthera leo (score = 0.93634)
- Tiger, Panthera tigris (score = 0.00074)
- Sundial (score = 0.00070)
- Hartebeest (score = 0.00056)
- Impala, Aepyceros melampus (score = 0.00049)

Google's Inception Image Classifier



- Crane (score = 0.11851)
- Solar dish, solar collector, solar, furnace (score = 0.07567)
- Parachute, chute (score = 0.06488)
- Steel arch bridge (score = 0.02698)
- container ship, containership, container vessel (score = 0.01682)

Clarifai's Region Classifier

- Air (score = 0.99684)
- Balloon (score = 0.99685)
- Sky (score = 0.99579)
- Flying (score = 0.99501)
- Freedom (score = 0.99281)

Clarifai’s Region Classifier

- Gun (score = 0.99165)
- Lid (score = 0.98930)
- People (score = 0.98835)
- Man (score = 0.984703)
- Weapon (score = 0.98191)

Image Likes by Google Classifier

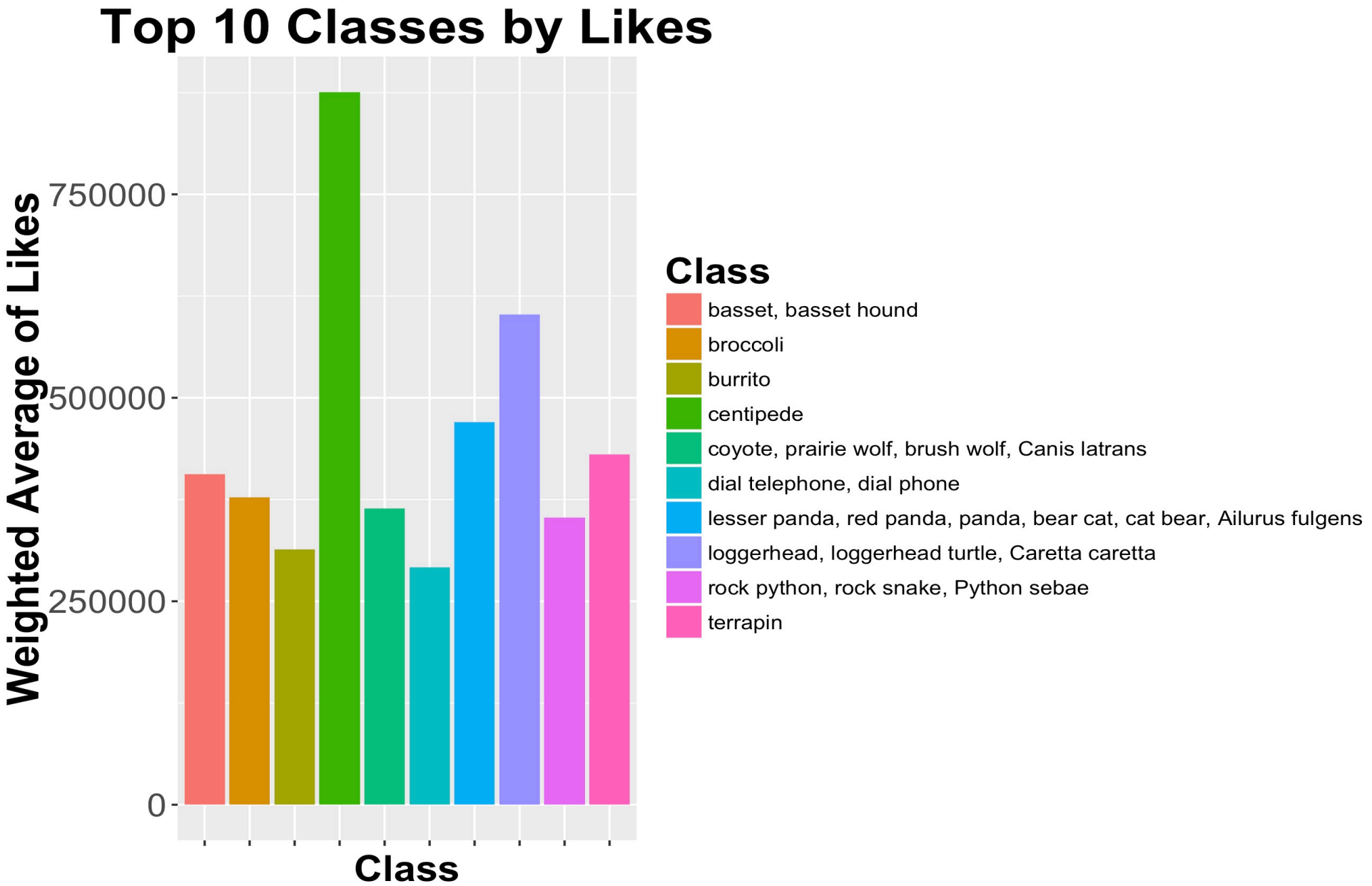


Image Likes by Clarifai Classifier

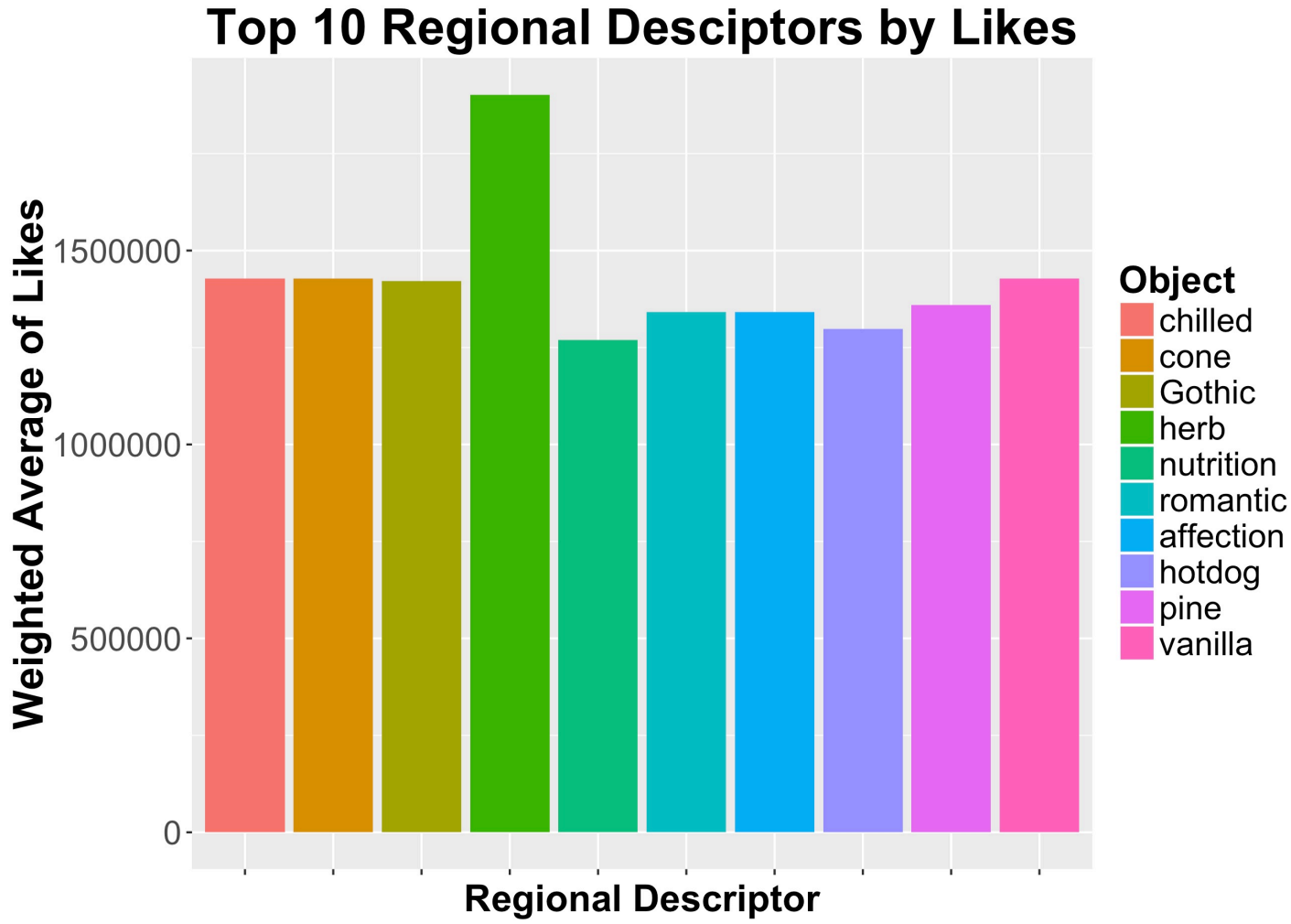


Image Comments by Google Classifier

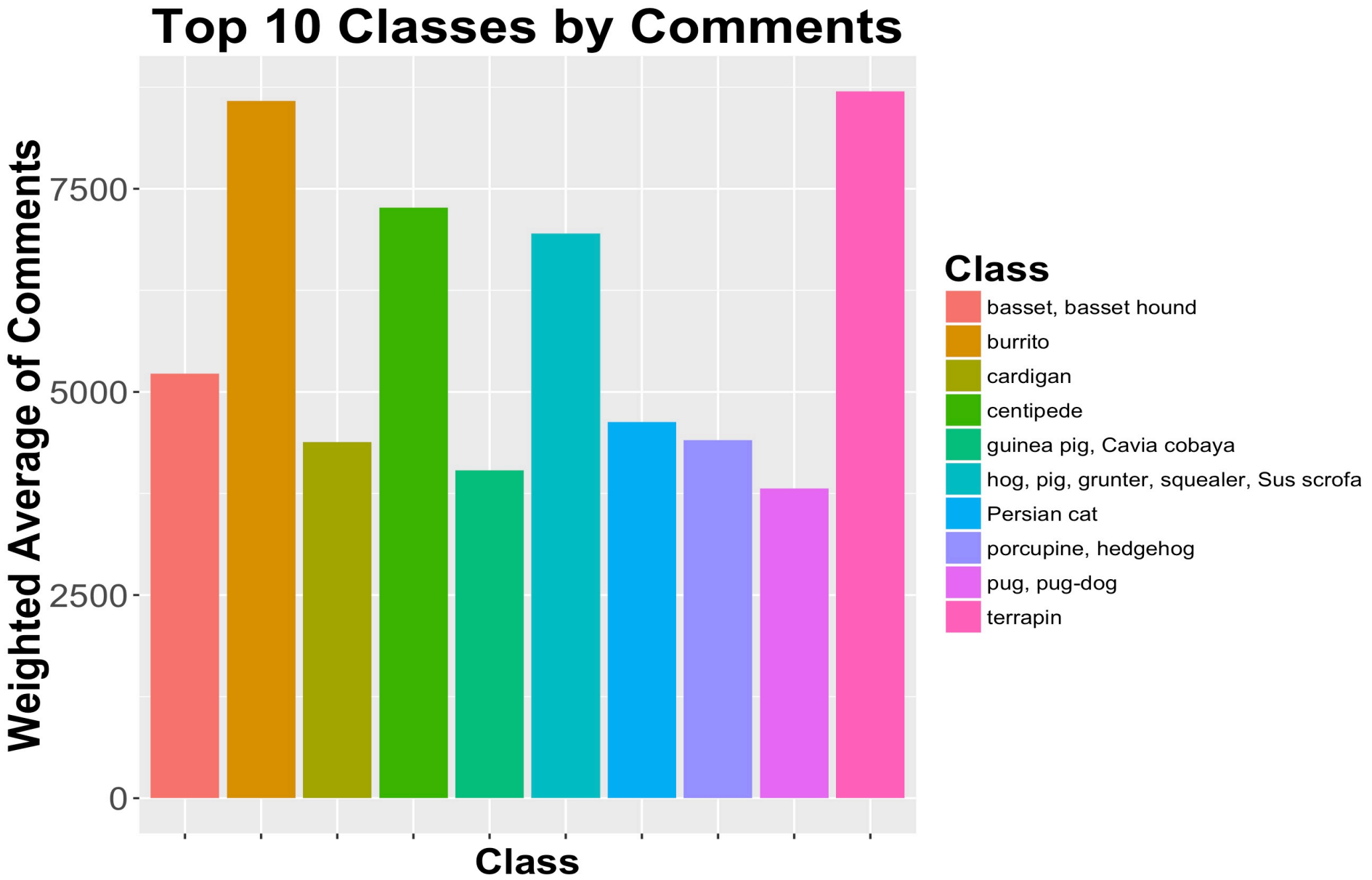


Image Comments by Clarifai Classifier

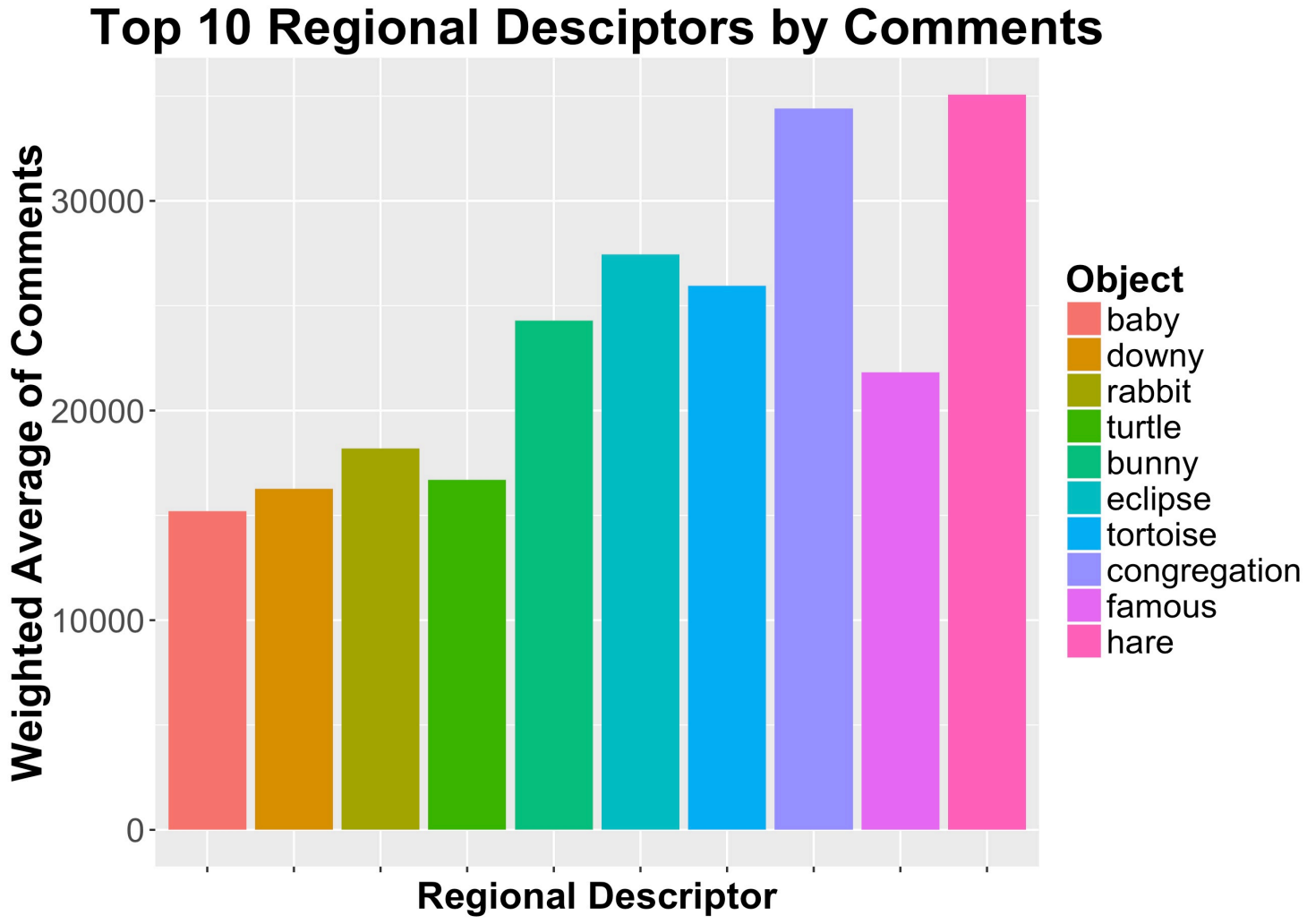


Image Counts by Google Classifier

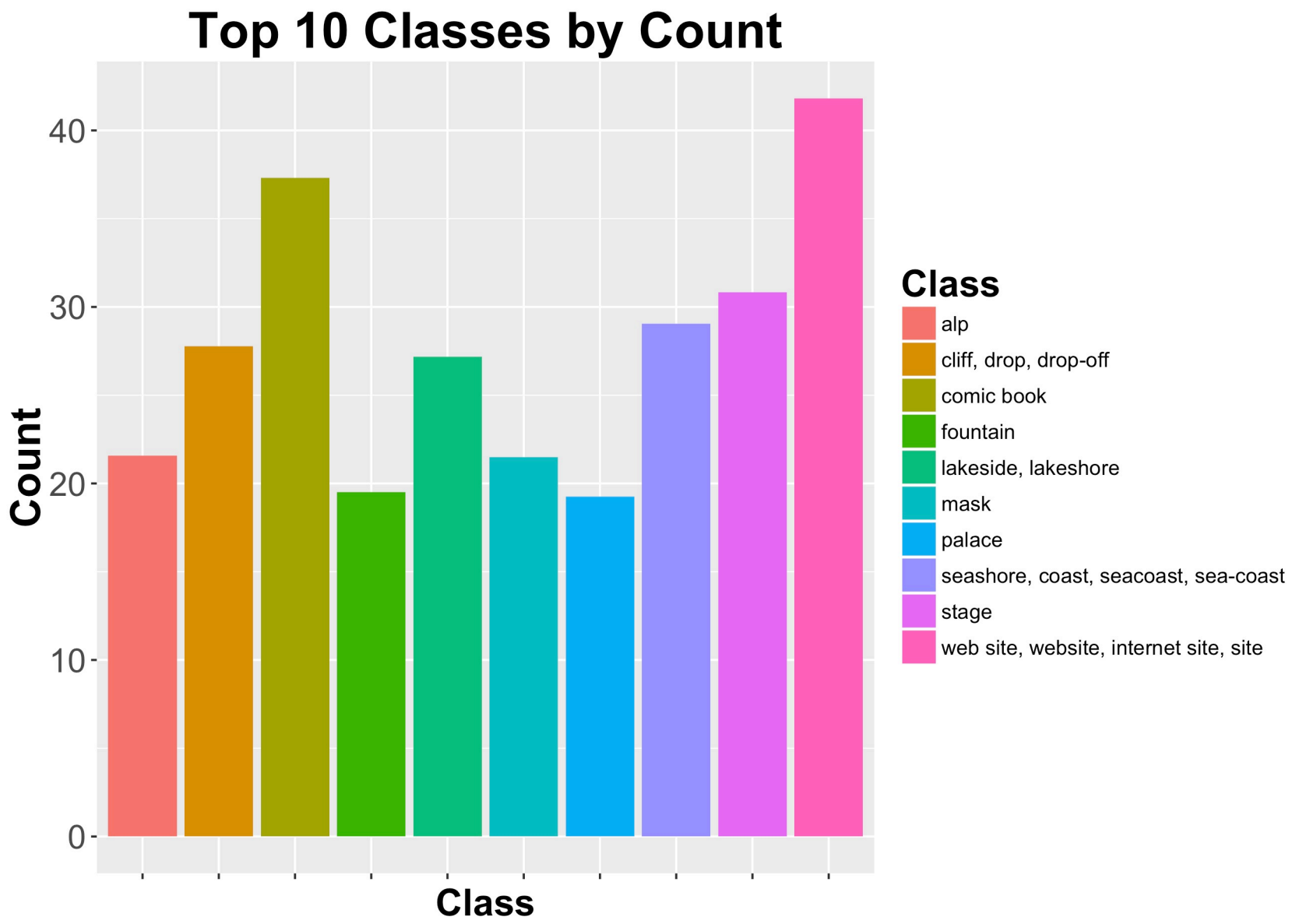
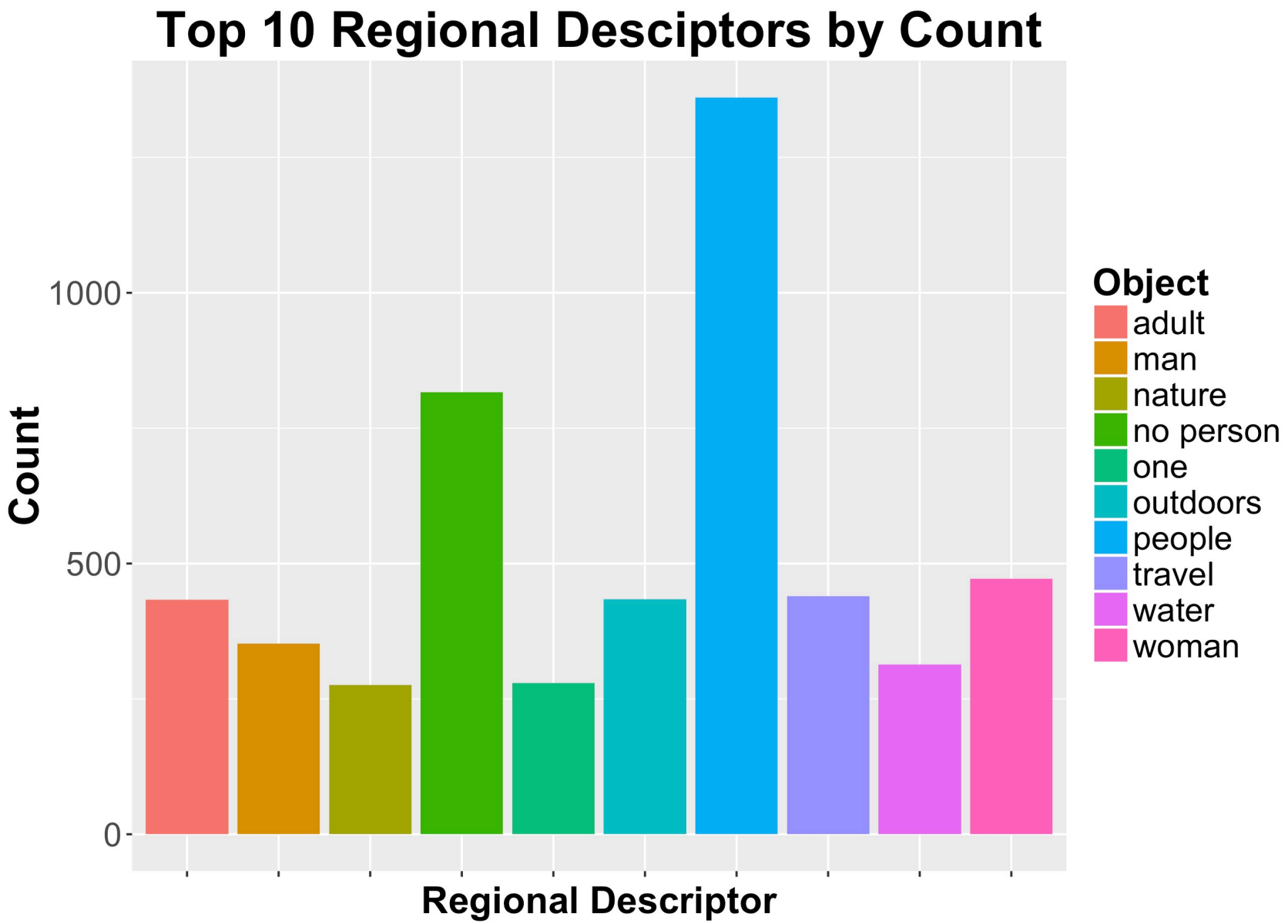


Image Counts by Clarifai Classifier



Conclusion

- Web scraping is messy and sometimes out your control (server problems)
- Computer vision is pretty awesome (but needs work)
- Not all data will be nicely distributed
- Clarifai's system seem to be better than the google classifier, but would cost money

Future Steps

- Scrape comments and do Natural Language Processing to gain more insight
- Use other computer vision models, i.e google's show and tell or caffe model zoo
- Train new model for specific category (i.e food) and perform similar analysis but on [food] instagram
- Fit machine learning model to predict likes/comments

Thank you! Questions?