Kaden Franklin

Dr. Ferrer

CSCI 270

15 April 2022

Pasta, Pizza, Person. An analysis of digital presence.

Research Question

Given a large dataset of images and text that are somewhat personal, such as a social media account, and some tools such as OpenCV. Is it possible to determine the favorite aspect of someone’s life? Such as their favorite kind of pizza? It might not be too far a stretch for a pizza connoisseur, but might we be able to delve deeper. In my project I will analyze posts from Eric\_thepizzaguy on Instagram to try and categorize some of his favorite things. He also frequently has posts dedicated to students, coworkers, and even visitors sometimes. In this project I will use Harr-cascade detection to find Eric’s favorite person by popularity on posts. There are already modules pre trained for facial recognition, also available under OpenCV. I will attempt to have the module produce frequency count if it can distinguish a person between different images. I will also attempt to train a Harr-cascade classifier to recognize different kinds of pizza. In order to do this I will need to have a set of positive samples (images with features you want to detect) and negative samples (images with features you do not want to detect). If I can successfully train the algorithm to detect different features of the of different pizzas. Then I can assemble a frequency chart for pizza over time, and adequately determine Eric’s favorite kind of pizza.

Data Set

At first I tried using a chrome extension to download images in bulk. Or maybe a python library that is runnable from the command line to download some images from a specific page. And you know these things exist, it’s just Instagram keeps disabling their features or updating code to not allow the downloading of their images. Which is fair but makes it a little harder to finish my project. So I just wrote some code that manipulates the web browser and mouse and keyboard in such a way that it iterates through every post that eric has made, opens it in a new tab and then saves the image or images associated with this post. Either of the former methods probably would have been faster and easier on their servers, but that’s their fault not mine.

The data set I will be using is the collection of posts from eric the pizza guys Instagram (eric\_thepizzaguy, <https://www.instagram.com/eric_thepizzaguy/>). Fortunately I was able to use a python library called instaloader, that just lets you download posts in bulk. (for free!) After a bunch of dumb errors I had to run it from the command line using:

python

import instaloader

L = instaloader.Instaloader()

for post in Post.from\_username(L.context, ‘eric\_thepizzaguy’).get\_posts():

L.download\_post(post, target = ‘#assets’)

The posts are labeled in chronological order, with JSON files containing post metadata, and the images associated with each post. In the JSON file there is post metadata such as the like count, and comments associated with this. I may go as far as to account for the amount of likes on a post to determine how popular someone is. In total there are 4,431 files, 1709 of which are images, with a slight 17 videos and the entire data set is 405 MB. I would gather that this is sufficient data for my purposes.

References

<https://docs.opencv.org/4.x/d6/d00/tutorial_py_root.html>

<https://instaloader.github.io/as-module.html>

<https://www.instagram.com/eric_thepizzaguy/?hl=en>