Kaden Franklin

Dr. Ferrer

CSCI 270

10 May 2022

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

Social media has become a massive industry over the course of the past decade. With companies like Facebook, founded in 2004, and twitter, in 2006 quickly becoming industrial titans in their own regard, it is hard to compare their success in economic factors to other industries. Though most major social media corporations have been founded in the U.S. it has nonetheless become a worldwide phenomenon. Social medias prosperous growth can be attributed to one key factor, its price. Anyone with an internet connection and a cell phone can create a social media account. The number of iPhone users has grown swiftly to 1 billion since the launch of the product in 2007 (1). Thus, it is no surprise that this unbridled growth of social media continues. It costs no money for the user to create an account. The user typically pays in another way, in information. The company which develops said social media app does not need to charge users because they get paid by other companies to place advertisements on the app. This is typically a mutually beneficial relationship because the social media app can give the advertising company information on its users, and in turn that company can use personalized and advanced marketing tactics. The normal person does not care that they are being subjected to this because they are not losing anything tangible. Some have voiced their concerns for more regulations on the amount of data that social media companies can be allowed to collect. It is still a serious issue being debated today. There are very few people outside of these companies that know exactly what the company knows about its users, or how much they give to advertisers. But certainly everyone who uses social media is aware, that this is happening, it will become a problem, and very little is being done to stop it.

On a much more lighthearted note social media enhances the internets’ ability to connect people across the world. It allows for underrepresented individuals to easily find communities, and for new forms of media and entertainment to be shared. It is certainly not all bad, after all there is a reason people keep going back to it while being aware of what was stated before. Given this it is sometimes fun to see what one can deduce about a person given only their social media account. After all it may be one of the only ways for some individuals have the chance to connect during their lives. Given a person’s social media account is it possible to determine their best friend. Given a particularly special social media account, might you be able to deduce more about that person. Let’s say for example that someone you know was a connoisseur of a particularly popular form of pie. If they posted about this particular pie on a consistent basis might you be able to determine their favorite flavor of pie? Let’s see by attempting such an analysis on a local chef at Hendrix college, Eric the Pizza Guy.

The data collected for this analysis is of course from a social media account on Instagram, @eric\_thepizzaguy (2). There is a python library named Instaloader that can be easily used to fetch the images, text, and much of the metadata associated with an Instagram account. To get all posts from a single account is fairly simple: Graphical user interface, text

Description automatically generatedIn total there were 1108 posts collected, and a total of 4,432 files. For each post there are four associated files that Instaloader retrieves. 1 image/video, 1 text document which contains the users text associated with the post, 1 JSON file containing metadata such as number of likes or comments, and 1 JSON.xz file. The files are each labeled after the time of the post in the form, YYYY-DD-MM\_HH-MM-SS\_UTC. In total this dataset takes up a grand total of 405 MB.

The majority of this project relies on OpenCV. OpenCV is written in C++ but has APIs in Java, Python, and MATLAB. It is an open source computer vision library that was developed by Intel in 1999. Since then it has been updated by the technology company Willow Garage. The library contains a number of different applications and tools that serve purposes ranging from machine learning libraries to facial detection systems. OpenCV serves innumerable purposes, but for this project the main components used were the Haar Cascade Classifier object detection method. This method uses a collection of weak features to train a cascade classifier to detect an object. A Haar like feature is some aspect of a digital image that can be used in object recognition. The cascade classifier is some preset module that is trained on these Haar-like features using a set of positive images and negative images. Since these features individually do not carry much weight, it takes a large number of samples to train a classifier. To put it simply the when the cascade classifier is called applied to an image it checks different locations in the image and retrieves a value for the pixel intensities in that region. It compares this with other regions and this way cascade classifier can determine which sections contain the feature it is looking for. This method was developed by Paul Viola and Michael Jones in their 2001 paper “Rapid Object Detection using a boosted cascade of simple features.” (6)

The process for creating a new cascade classifier is a bit tedious. It requires the user to manually assemble a set of positive and negative images. There is a helper function called pizza\_pie() in main.py, made to assist with generating the positive and negative text files required for the first step of this process. The user then uses a few more OpenCV programs to turn this set of images into a classifier. These programs can be run from the command line, the cascadeutils.py file describes in more detail how to complete this process. Basically opencv\_annotation.exe lets the user mark up the set of images, and generates a text file of corresponding annotations. Opencv\_createsamples.exe makes a vector file, which is then used by opencv\_traincascade.exe to generate data for the cascade classifier (7). This process is straight forward once you understand the arguments each program takes in, but can take a variable amount of time depending on factors such as the number of stages the cascade classifier is being trained on. There were some complications along the development of these cascades. Initially there were going to be as many as 10 cascade classifiers for different ingredients. By the end of development there only 5 classifiers that were able to run without any errors. The ingredients detectable by this project are jalapenos, mushrooms, olives, pepperoni, and pickles.

After the cascade classifiers have been created, main.py does all the remaining work with a few helper functions. Dect\_face() checks multiple cascade classifiers made by OpenCV, to determine if an image contains a human face. If the image does contain a face, it gets copied to the /Person directory. Depending on what you train classifiers to detect, this can also be used as the negative image dataset. Dect\_pers() is a helper function, that detects human faces such that the values can be appended to list and compared with frequency of ingredients. Not\_comment() is a leftover function that attempted to extract metadata from the Json files associated with each post. Variables such as like count, post sentiment, and number or sentiment of comments were to be used to determine eric’s favorite person. It remains unfinished.

A picture containing chart

Description automatically generated The main algorithm has an underlying structure that depends on a list of lists, called pizza\_list. It loops through every image in the ‘#assets’ folder generated by instaloader. For each image the cascade classifier determines if the ingredient is contained in the image. If the image does not contain an ingredient, append the length of what is returned by the classifiers detectMultiScale() function to the corresponding list. This is the number of hits that the classifier got when searching for an ingredient in the image. If the image does not contain an ingredient, append zero. After looping through each image, the values stored in pizza\_list are then used to generate a frequency graph of ingredients. Each ingredient has a corresponding color. Jalapenos. yellow; Mushrooms, brown; Olives, black; Pepperoni, red; and pickles, green. Later the person category will also be added, the corresponding color is blue. Though this plot looked a little simple. It was clear that the olives were either all to frequent ingredient, or the cascade classifier was picking up multiple false positives. This is not surprising as they are one of the most plain looking ingredients, and black being as common a color as it is. After some consideration the number of stages in the ‘Olive’ cascade classifier was increased to reduces the number of false Chart

Description automatically generatedpositives, without much effect. There was also a variable added to account for people in eric’s posts.

The most popular kind of ingredient detect was still olive. The second most popular ingredient being pepperoni, this would be expected to be the most common ingredient, and may still be. Though this data is entirely based off one Instagram account, it reflects much more about the community he is an integral part of. Yes, making pizza is clearly Eric’s passion, he has been doing so, at least publicly for five years now. His Instagram account is more directly tied to his connection with the Hendrix community. Just scrolling through his page or even when using somewhat questionable tools, it is evident that the main purpose of Eric’s Instagram is not to display what is for lunch today. But instead to show who is having lunch today. It could be a stranger, it could be your best friend, it could be someone having their first or last meal at Hendrix. In any event, Eric is always there happy to be a compassionate listener, and dutiful community member. He has never once denied a person a shoutout or a slice. Clearly Eric is a shining beacon of how we can use social media to spread positivity.

All tools necessary to replicate this project can be found at:

<https://github.com/KadenFranklin/pizza_pasta>

References

1. <https://backlinko.com/iphone-users#iphone-key-stats>
2. <https://www.instagram.com/eric_thepizzaguy/?hl=en>
3. <https://instaloader.github.io/as-module.html>
4. <https://docs.opencv.org/4.x/d6/d00/tutorial_py_root.html>
5. <http://man.hubwiz.com/docset/OpenCV.docset/Contents/Resources/Documents/d7/d8b/tutorial_py_face_detection.html>
6. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.6807
7. <https://manpages.ubuntu.com/manpages/impish/man1/opencv_createsamples.1.html>