Progress Report I: Are Adoption Postings Visually Subjective? Comparing the Results of a Purebred Dog Breed Image Classification Model Against Mixed-Breed Petfinder Listings

Allison Aprile, April 2019

Abstract

The first week of work paves the road for the remainder of the project. While there has been some success, underlying issues in the original plan have called for re-evaluation and implementation of methods. This document reports on the project progress from the period of **April 15th-22nd, 2019**.

Work Proposed

The data collection and annotations, in addition to the model selection, were planned to be completed by the end of this time period. Specifically, I had intended to use the Petfinder API in conjunction with the PetPy Python wrapper to filter and extract dog breed listings. Then, I would visit the pages of all 100-200 listings and save the images with their respective annotations in a file. Finally, I would select and modify the best Dog Breed Image Classification Model to prepare for the next week's implementation.

Complications

My data collection method was dependent on the functionality of the PetPy wrapper. In search for an efficient implementation, I came across iPython notebook code that used PetPy to quickly download images from the database. After setting up a Petfinder developer account and installing the wrapper, I revised the code to collect 1,000 dog images and their breed classifications. Despite other API tests and no mention of issues in the documentation, I received JSONDecodeErrors.

I initially attributed this to an error with the JSON installation in my iPython environment. I exhausted over three hours attempting to reconfigure the environment and JSON. Towards the end of this period, I performed several other JSON trial runs, including with the Twitter API and raw JSON files. The codes performed without issue, indicating that the problem did not lie with JSON.

I then compared the PetPy source code loaded from the installation to that on the documentation. There were no discrepancies, leading me to probe issues with the Petfinder database communication. After discussing my procedure by phone with a Petfinder technology representative, Joe Cabanilla (see **Acknowledgements**), we concluded that the issue lied in the authentication and authorization of my developer account. He instructed me to perform a few tests in Jupyter Notebook using the `requests` module, which confirmed the issue. He then kindly assisted me with the Petfinder API authentication procedure, involving cURL commands to OAuth 2.0 through command prompt. After receiving an access token, I was able to retrieve data successfully directly using cURL commands. However, as I had never worked with cURL before, I was intimidated by the process and imprudently decided to devote more time to PetPy.

After doing some research, I found that the JSONDecodeErrors were indicative of the inability to connect to the database, rather than an issue with the JSON parser. I read through the documentation again and discovered that Petfinder had recently released a second version of the API, including the authorization process not required in the first version and thus, not considered in PetPy. I was confident that I could revise the PetPy source code to process the access token in addition to the API key and secret. However, despite following several online guides, my attempts were fruitless. The access token is time-sensitive, overall hindering my ability to request large quantities of data through PetPy as intended.

After approximately ten hours of fighting with PetPy, I decided to try webscraping with BeautifulSoup. I followed the instructions of several online tutorials in addition to Lab 4. Although I was able to scrap some data (including pet name and location), I had difficulty extracting the primary breed of each animal due to the HTML structure of the listings webpage. I ultimately decided to seek an alternative method as the webscraping was not yielding the correct data with the desired efficiency.

I dedicated some time to thoroughly reading the Petfinder API documentation, which confirmed that direct calls through cURL would be the most efficient and organized method of collection. As I was inexperienced with the procedure, I enlisted the advice of my peer Tommy Tan (see **Acknowledgements**). His help was fundamental to successful data collection and the continuation of my project.

Work Complete

Data Collection: The Petfinder API documentation provides detailed instructions on data requests using cURL commands. With the assistance of both Joe and Tommy, I was able to extract 1,600 dog listings in JSON format. The procedure including the following call:

```
curl -H "Authorization: Bearer eyJ0eXA..."
https://api.petfinder.com/v2/animals\?type\=dog\&limit=100 -> json_test(call#).json
```

The command restricts the request to listings of "dogs". Additionally, to speed up the process, I requested the maximum limit (100 elements). Tommy helped me with the call suffix that allowed me to export each output to a JSON file, which I labelled per the call number (1-16).

I collected 1,600 elements to ensure a diverse selection of dog breeds as well as to account for any missing information or broken image links. The cURL collection took approximately 20 minutes and yielded excellent results, the JSON containing multiple image links per listing in addition to a unique identification number, primary breed, and other not as relevant information.

Data Cleaning: To gain a better picture of my dataset, I converted all of the JSON files to CSV using an online program. The raw formatting contained blank lines and overwritten data so I dedicated approximately four hours to carefully deleting empty lines and any listings with overflowing cells. I then spent another hour combining the 16 individual files into a master Excel worksheet and inspecting for any other issues. I remained with about 1,500 elements.

Data Implementation: Because I may plan to use the master worksheet to perform EDA on the general Petfinder trends, I did not want to delete any irrelevant features directly in Excel. Thus, I loaded the CSV in Jupyter Notebook and created a DataFrame. I narrowed the data down to four columns: ID, Primary Breed, Mixed Breed (Boolean), and Small Photo (URL). Then, I randomly selected 150 elements with the Mixed Breed = True filter and saved that in a separate DataFrame.

Time Approximation: 15 hours

Work Incomplete

Due to unexpected complications with PetPy, I fell behind in my progress. Therefore, I will first need to dedicate about five or six additional hours to the data annotation process. This will involve visiting individual image links, saving the image as the unique animal ID, and annotating it with its corresponding primary breed in a CSV file following the breed annotations of the Stanford Dogs Dataset.

Next, I will need to select and modify one of the CNN Dog Breed Image Classification models. I will first train and test it on the purebred dataset until I reach a desired accuracy, then perform several tests with the mixed-breed testing dataset. I intend to have this complete by the end of the next work period (**April 23-29, 2019**).

I am scheduled to present my project on **Tuesday, April 23rd**, so I will also need to complete the accompanying PowerPoint before the evening of Monday, April 22nd.

The final project will summarize my findings regarding the visual subjectivity of the Petfinder breed classifications as well as the possible addition of general Petfinder Exploratory Data Analysis. This is to be completed by **May 7th**, followed by an individual project discussion on **May 8th**.

Reflection

I faced many detriments throughout my data collection process that compromised my intended timeline. However, despite the discouraging week, I gained valuable knowledge and application of cURL commands. The cURL collection surprisingly enriched my dataset and has allowed me to analyze attributes not available through the PetPy wrapper. I feel confident entering the next work period and hope to make significant progress.

Acknowledgements

- **Joe Cabanilla** – *Petfinder Technology Operations Specialist*
- **Tommy Tan** – *Student, Peer (Principles of Data Science)*