Choose Your “Dog”stiny

How many times have you been online trying to find the perfect house, your next car, what restaurant to go to, etc.? We are used to visiting websites that help us find these items by typing in specific criteria and then generating a list of results based on this. What happens if we could utilize this same idea to help people find a furry best friend?

A lot of the pet finder websites rely on someone to select a breed and then generating a compilation of characteristics that pertain to that breed. What I am attempting to do is to reverse that look up. I will be working to create a project in which people can go in and pick criteria and then receive a list of breeds that relate to that criteria. For example, if I am active and looking for a dog that will do well in an apartment, but also go for long walks, I would select those characteristics. Based on what I submit, the search engine could come back and tell me a Jack Russell Terrier would be a great match.

This would be very valuable in providing the searcher with breeds that would do well with their environment, home, lifestyle, etc. It would serve to educate people on breeds as well. A lot of times people get dogs that they think sound awesome, only to not be able to care for them or give them the environment that the dog needs. For example, when “Game of Thrones” first started, people fell in love with the Direwolves and started to get Huskies. However, Huskies are very vocal and require a lot of exercise. A lot of people did not realize this and unfortunately, a lot of the huskies ended up in shelters or given back to breeders.

I will be getting my data source from an existing website called dogbreedinfo.com. This site provides comprehensive descriptions of 100s of dog breeds. I plan on scraping each dog breed once and then saving the data into a csv. There is a lot of data, so data cleaning will be a must. Techniques such as Beautiful Soup will work nicely because it will be helpful with web scraping.

I am very interested in NLP and think that this would be an excellent project to really dive into the details. Techniques such as text extraction and text summarization would be ideal to use in this project. As the majority of the data in this project is text based, I will be utilizing unsupervised learning techniques. Clustering methods (K-Means and Mean Shift) will be useful. Clustering will allow me to generate features for a model, as well as understanding if my data points have common threads. For example, let’s say I want to find breeds that are similar to the Siberian Husky, clustering would help me to identify what common key words exist within similar breeds based on how the clustering looks. Word2Vec and PCA will be utilized for feature generation. The data will be cleaned by removing stop words, creating lists, tokenization, etc. There are numerous pandas packages that can be used to do this. Things like punctuation and numbers will be removed as well. Tfid Vectorizers will allow me to transform text into feature vectors.

I see two big challenges that I will need to work through. The data source itself has gaps in it as well – while the dog breeds do have all the different sections, not all the sections are filled out. I will have dog breeds that do not contain any information. Even with this challenge, I am confident that I can create a project that will produce the results I am looking for. Another challenge is that we will have a lot of common words and uncommon words. It will be important to work through this to find the right balance.

There is a lot of potential with the data. The key to success will be to not get lost in all the data points, but to use those data points to create picklists that make sense without adding unnecessary complexity.