Json Structure:

Annotations:

{'genus\_id': 1, 'institution\_id': 0, 'category\_id': 0, 'image\_id': '00000\_\_001'}

Categories:

{'category\_id': 0, 'scientificName': 'Abies amabilis (Douglas ex Loudon) J.Forbes', 'family': 'Pinaceae', 'genus': 'Abies', 'species': 'amabilis', 'authors': '(Douglas ex Loudon) J.Forbes'}

Contain three levels of hierarchical structure, family - genus - species, from the highest rank to the lowest rank. Species is only unique under its parent node genus, which means that we can find multiple categories with the same species name under different genus names. This is due to the taxonomic nature that plants are named after, and the genus-species pair is always unique in our data.

Genera:

{'genus\_id': 1, 'genus': 'Abies'}

Images:

{'image\_id': '00000\_\_001', 'file\_name': '000/00/00000\_\_001.jpg', 'license': 0}

Distances:

{'genus\_id\_x': 2508, 'genus\_id\_y': 2576, 'distance': 0.0325632}

A set of pairwise phylogenetic distances among genera

License:

{'id': 0, 'name': 'Public Domain Dedication', 'url': 'http://creativecommons.org/publicdomain/zero/1.0/'}

Institutions:

{'institution\_id': 0, 'collectionCode': 'A'}

Conclusion:

So from information above, we can get ‘image\_id’ from Annotations to find image source in Images and get ‘category\_id’ from Annotations to find family – genus – species. Besides, the genus-species pair is unique, so I think dependent variable is genus-species pair and I will try to use traditional classification method like knn at first. Then I will try new method to build model.