Prateek Prakash

EZ7543

31 October 2015

**CSC 5991 – Final Project Proposal**

1. **OBJECTIVE**

The objective of this project to be able to create a classifier that is successful in classifying unknown fossil data in the proper taxonomy. The classifier should implement machine learning methods and should be implemented in Python. The objective is to try out at least two different machine learning methods to implement two different classifiers, and then compare the result of both.

1. **DATASET DESCRIPTION**

Our dataset will be composed to two separate sets. Our first set of data will be our training data. This will be used to train each classifier respectively. Our second set of data will be our testing data. This will be used to test the accuracy of each classifier.

* 1. **Training Data**

The training data consists of text data. Data for fossils have been collected into a reputable database, and this is the data I will be using. The data consists classification such as family, suborder, etcetera. The data will also have descriptions for each type of fossil. The descriptions, and titles such as family and suborder shall be used to train our classifier. The classifier should be able to take descriptions of fossils and output their classification.

* 1. **Testing Data**

The testing data will have to be gathered. It shall be gathered from publicly available sets online. Descriptions of different fossils shall be gather and put into a text file. These descriptions will be classified through the both classifiers.

1. **METHODS (MODELS)**

The two models that I found are part of in SciKit-Learn. The first one that I shall be implementing is the “Decision Tree Classifier”. The second one that I shall be implementing is the “Multinomial Naïve Bayes”.

* 1. **“Decision Tree Classifier”**

Reference: <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier>

Source: <https://github.com/scikit-learn/scikit-learn/blob/a95203b/sklearn/tree/tree.py#L401>

* 1. **“Multinomial Naïve Bayes”**

Reference: <http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB>

Source: <https://github.com/scikit-learn/scikit-learn/blob/a95203b/sklearn/naive_bayes.py#L516>

1. **STEPS**

The first step is essentially picking machine learning methods to use for the classifier. These as you know, were done already and were described above. The next step involves preprocessing all the data. This consists of removing noise from the dataset and labeling all the data. After getting the data preprocessed, we’ll have to use our methods in order to train the classifier. This will involve using using the training data described above and applying them to the methods, also described above. After training our classifier, we shall test the classifier on the same dataset used to train the data. After this step, we will need to collect testing data, which was described above. We will then test our classifier on this new dataset to test the successfulness of our classifiers. Finally, we will report the results. The results will include: preprocessing steps, methods and tools used, plots to graphically portray our results, and some simple comparisons between our two models.

1. **EXPECTED OUTCOME**

Our expected outcome is to have at least one, if not two, solid classifiers implemented in Python which are as close to 100% accurate in classifying unknown fossil into their proper taxonomies.