**Data Bootcamp Capstone Project ML Model Notes**

**High Level Project Overview**

Diagram

Description automatically generated

**ML Model Outline**

**1. Web Scraping / Inserting Data to PostgresSQL DB**

* For our final deliverable we plan on creating a program with BeautifulSoup, Splinter and py-postgresql that will scrape each Annual Sunshine List into a table within a PostgresSQL database
* The following is a screenshot displaying Sunshine List records from 1996 and the columns for reference: Graphical user interface, application, Word

  Description automatically generated

**2. Data Cleansing**

* Once the data has been scraped and inserted to our PostgresSQL DB we will begin cleaning the data for ML and analysis using SQLAlchemy and Pandas
* Basic cleansing will be performed on the entire dataset including:
  + Optimal/desired column rearrangement
  + Desired column titles
  + Dealing with Null values
  + Etc.
* Then, the ‘First Name’ column will be used to create a new Pandas series where only the unique names will be kept
* Lastly, a new column will be added to this series--making it a DF—called ‘Gender’
  + We are going to use our ML model to predict the gender of each Sunshine List record

**3A Machine Learning Model – NLTK Gender Identification**

* Natural Language Toolkit (NLTK)
* The NLTK Python library uses supervised classification to determine a gender class (output) for each input (name) it is given
* When the library is installed it will automatically import two text files, ‘male.txt’ and ‘female.txt’ which contain a large list of male and female names respectively **(could use other names source … TBD)**
* From there we will prepare a labelled class list using the following code: Text

  Description automatically generated
* Then we will randomly shuffle the above list and divide the resulting list into a training set and a testing set
* From there, we instantiate a naive Bayes classification model (classifier) and train it with the training set
* Assess accuracy **(specific methods TBD)**

**3B Machine Learning Model – Public Sentiment Analysis (Twitter Data)**