Food for Thought

How do you extract a dataset from a URL, load it as a dataframe, and split it into training and test sets?

Fetching Data

Overview: Download a single compressed file (e.g. housing.tgz) which contains a CSV file (e.g. housing.csv)

```
import os
import tarfile
from six.moves import urllib

DOWNLOAD_ROOT = "https://raw.githubusercontent.com/ageron/handson-ml2/master/"
HOUSING_PATH = os.path.join("datasets", "housing")
HOUSING_URL = DOWNLOAD_ROOT + "datasets/housing/housing.tgz"

def fetch_housing_data(housing_url=HOUSING_URL, housing_path=HOUSING_PATH):
   if not os.path.isdir(housing_path):
      os.makedirs(housing_path)
   tgz_path = os.path.join(housing_path, "housing.tgz")
   urllib.request.urlretrieve(housing_url, tgz_path)
   housing_tgz = tarfile.open(tgz_path)
   housing_tgz.extractall(path=housing_path)
   housing_tgz.close()

fetch_housing_data()
```

Loading Data

Overview: Use the pandas library to convert the data into a readable, usable format.

```
import pandas as pd

def load_housing_data(housing_path=HOUSING_PATH):
   csv_path = os.path.join(housing_path, “housing.csv”)
   return pd.read_csv(csv_path)
```

Splitting Data (Train/Test)

Overview: Divide the data set into two subsets called the training set, a subset to train a model and a test set, a subset to test the trained model.

```
import numpy as np

def split_train_test(data, test_ratio):
   shuffled_indices = np.random.permutation(len(data))
   test_set_size = int(len(data) * test_ratio)
   test_indices = shuffled_indices[:test_set_size]
   train_indices = shuffled_indices[test_set_size:]
   return data.iloc[train_indices], data.iloc[test_indices]

housing = load_housing_data()
train_set, test_set = split_train_test(housing, 0.2)
```

Resources

Training and Test Sets: https://developers.google.com/machine-learning/recommendation
Video Tutorial: https://www.youtube.com/watch?v=AtkWpgJJHgQ