CS 5010 Homework: Python and Web Scraper
Kip McCharen (cam7cu)

Virginia Dendrology Dataset

Overview/Reasons

- What are all the berries that grow in virginia?
- Which are good to eat fresh?
- Which of them are poisonous?
- Are there any other uses for trees and bushes that grow in Virginia?

If you're like me, you definitely don't know the answer to these questions off the top of your head. However, by gathering data from available sources online, we can construct thorough and detailed answers to all of these questions! I did just that by leveraging academic, governmental, and private information sources in combination to produce a thorough and detailed dataset for each tree and bush growing in Virginia, although the module can be used to gather plants native to any state in the US.

Approach/Algorithm/Libraries/Data

In general my approach was to simply gather as much data as possible, then crunch it, then munch it. Within each module, I took the following steps:

1. Gather the html using **requests**
2. Leverage **BeautifulSoup** to navigate the html
3. Collect the appropriate sections using varieties of BS4's **find** function
4. Iterate through a collection of tags using **for**...**in** loops and list comprehensions
5. Munge the appropriate data using regular expressions (**re** library) and a truly monstrous number of **if** statements!
6. Collect data in a dictionary for each plant, then form a list of all the dictionaries

The overall order of actions was:

1. Scrape **http://dendro.cnre.vt.edu/dendrology/data_results.cfm?state=VA**
    a. Virginia Tech's dendrology website provides a list of plants growing in each state
2. Scrape Va Tech's dendrology **page for each specific plant**
    a. Available info includes latin name, similar-looking plants, pictures, detailed information about leaves, flowers, fruit, twigs, bark, and general form
3. Scrape **pfaf.org** for more detailed practical information about each species
    a. Plants For A Future (PFAF) is an online free-to-use information database and associated website for those interested in edible and useful plants.
    b. Available info includes hardiness zones, known hazards (is this poisonous), habitats, propagation advice, edible uses, medicinal uses
4. Get the official taxonomy ID # (TSN) and scrape the taxonomy data from the **ITIS API**
    a. The USGS runs the Integrated Taxonomic Information System (ITIS) which has a free unsecured API to get taxonomy info by species.
    b. The results of these websites/requests were in pure XML, and I had to use **xml.etree.ElementTree** to navigate the tree of items

    c. Available info includes official TSN, kingdom, phyla, class, order, family, genus, species, varieties, and many more classifications I had never heard of!
5. Once a dictionary for each plant was populated into a list, **Pandas** natively accepted that to create a dataframe, which was easily output to a csv.

Data Analysis

The questions asked above were:

- **What are all the berries that grow in virginia?**
    - I flagged all instances of “berry” in the name or biological fruit detail, then filtered the dataframe to show them!
    - I had no idea that there’s a “**Himalaya blackberry**” or “**Elliott’s blueberry**”
- **Which VA berries are good to eat fresh?**
    - Edible data was in sort of a garbage form, so I hacked a function to iterate through items and place them in the appropriate “box”. Then I flagged all instances of “Fruit - raw” in the Edible Uses column for a new column “FreshFruit”. Then out of the dataframe of berries above, found fresh ones!
    - I didn’t know you could eat **barberry** or **chokeberry**!
- **Which VA berries are poisonous?**
    - I flagged all instances of “poison” or “toxic” in the “Known Hazards” column, then applied that data to the berry data above
    - I was surprised to learn that **strawberry bushes**, **honeysuckle**, and **teaberry** are all poisonous in different ways
- **Are there any other uses for trees and bushes that grow in Virginia?**
    - The “Other Uses” column needed to be evaluated literally (using the ***ast*** package), then items in the lists were joined together, filtered out false values like “None known” or “Special Uses”, and then I used the ***Collections*** function ***Counter*** to show a list of the most common uses of VA trees and shrubs!
    - I had no idea that trees were a source of **tannins**, **soap**, or **dye**?

Future Extensions

In the future, the columns could be better refined, and the data could be used to answer:

- How closely related are two species
- What species are most easily confused

Extra Credit

Instead of scraping one page, 384 entries x 4 pages for each entry = 1,536 pages were scraped by this project and assembled neatly into 1 table. In total, this program gathered 17,116 cells of data. On my home computer it took exactly 34 minutes to complete. The scraping techniques are intricate for each different page, using both BeautifulSoup and xml.etree.ElementTree to successfully extract data.
I was able to answer not just the same sorts of questions about berries, but also my question about the wide variety of all uses for trees and shrubs that grow in Virginia. The potential of this wide accumulation of data is worth continuing to explore regardless of this assignment and I expect to use this in the coming years to better understand the woods and wildlife around me that I love so much.