Bonus Assignment (deadline: April 17, 2017, 8:00 AM)

Task goal: page https://plants.usda.gov/java/factSheet has 1091 plants information, take this page as the starting page, followed by two levels of hyperlinks and write all information extracted from each plant description page to the MySQL database.

Each plant has a profile page, for example: https://plants.usda.gov/core/profile?Symbol=ACRU

Profile page of all the bold tag (e.g. Symbol, Group, Family, Duration, Growth Habit, Native Status, etc.) should be extracted as a field name and fill the corresponding content.

Corresponding fact sheet and Plant Guide PDF files in boldface keyword should be extracted as the database fields, the value of the text as the value of the column.

Each plant has a page Characteristics, such as: https://plants.usda.gov/java/charProfile?Symbol=ACRU

The Characteristics of each plant should also be extracted and put in different column (e.g. Active Growth Period, CaCO3 how Fruit/Seed Period End, Fodder Product etc.) here field may exceed $166 or more, before data collection completely, it is difficult to determine how many data fields could be.

Database table name should be defined as usdaplant, column name should use the underscore _ replacing Spaces (alphabet and underscores only). Field extracted from Fact Sheet documents shall increase fs_ prefix, as a field from the Plant extract field Guide documents shall increase pg_ prefix as fields.

Version, milestone, code, documentation, mind mapping, etc., the iteration/update/branch must be submitted to github.

TABLE structure can be dynamically created while data processing (using the CREATE TABLE to create the initial TABLE structure and ALTER TABLE to add fields or adjust field). Table cardinality should be around 1091, degree might be more than 200 columns.

Duplicated columns must be eliminated. The quality of your table design and data will be evaluated.

Below is an incomplete table structure:

Symbol	Scientific_Name	Common_Name	Group	Growth_Habit	Height_at_20_Years_Maximum_feet	fs_Alternative_Names	fs_Uses	pg_Alternate_common_names	pg_Uses	……
ACRU	**Acer rubrum**	**red maple**	Dicot	Tree	35	swamp maple	Erosion control: Red maple is available in quantity for revegetation work……	Carolina red maple, Drummond red maple, scarlet maple, soft maple, swamp……	Red maple has long been valued as an ornamental tree (shade, specimen……	……

liyu@bjtu.edu.cn

数字签名者：liyu@bjtu.edu.cn
DN：cn=liyu@bjtu.edu.cn,
email=liyu@bjtu.edu.cn
日期：2017.04.09 14:22:34
+08'00'