Hort 503 Assignment 15

11530998 Chun-Peng Chen

One of my project in summer is to build a web platform that allow plants breeders to access their data in a friendly way. In crop breeding, it's always easy to generate tons of data from the field work and sequencing companies. However, it's also a pain for breeders whenever they attempt to fetch meaningful information from those random files. My idea is to apply the knowledges we've learned in the class on genotype and phenotype files, such as data wrangling by Python/Pandas, Python/NumPy and Python/SciPy. After have the back-end work done, I can visualize the summary information by Python/Matplotlib, or even Python/Bokeh if I want to have the plots become interactively. Finally, I would be able to wrap up the whole framework with Python/Django. Thanks to Python is an Object-oriented programming language, I think the whole work would make more sense than it would be in R in terms of readability and encapsulation.

Another part I've enjoyed in this class is to perform some simple tasks on text files by "awk" or "sed". Often time when I need to make some changes on a text file, I would do it in a programming interface like R or Python. Usually it works fine. But when it comes to a big matrix (say it has 10 million lines and 100 thousand fields), it would take quite a while for the interface to load files in. With "awk" or "sed", we can easily do text tasks on large files without spending time on loading, and the script is also tidy enough to use. It particularly useful to me in organizing files from different breeding programs on servers or HPC, since there's no dependency requirement for awk and sed.