17/1/2017

* Install pandas. Use pandas to read CSV values from CSV file.
* Reading csv file directly using pandas results in memory error.
* Instead, read csv file chunk by chunk, then store it into a HDFStore, for convenient retrieval, so we don’t have to read the csv file each time.
* Result of reading csv file for first two rows: 2 rows and 848 (including probe IDs column) columns. Rows represent probe IDs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 3998567027\_R01C01 | 3998567027\_R02C01 | … | 3998567028\_R06C02 |
| cg00000029 | 0.395178 | 0.369714 | … | 0.432102 |
| cg00000108 | 0.885050 | 0.886108 | … | 0.901180 |

18/1/2017

* Use interrow and chunksize to read TXT file chunk by chunk and then counting the number of rows. Can use this pattern to find the number of rows/columns.
* Use pattern matching “disease\_status=1” to find how many control/patient cases there are.