Pandas

Neil Muller

July 7, 2012

Pandas 0.8.0

- Pandas - "Powerful Python Data Analysis Toolkit"
  - "goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language."
- Provides data structures for working with labeled / relational data
  - Data doesn't have to be homogenously typed
  - Well suited to statistical data
  - Handles missing data
  - Support merging, data alignment, grouping, fancy slicing operations, reshaping and much more
- With 0.8.0, now ow completely replaces scikits.timeseries
- Good integration with statsmodels (will become a hard dependancy in future statsmodels releases)
- Limited set of analysis and visualisation tools - statsmodels, etc should handle this

Series

- Basic building block of pandas
  - Homogenous list of values with labels
  - Supports a handful of types - will upcast to the most inclusive type for mixed type input
- Supports vector operations
- Array interface, so generally "just works" with numpy methods
- Automatic alignment working with unaligned arrays
- 0.8 allows non-unique indexes
  - in previous versions this worked for some operations due to way indexes are evaluated, but wasn't actually supported
  - Various operations not supported for non-unique indexes, though

DataFrame

- "Series of Series"
- Doesn't mimic 2d ndarrays - supports hetrogenous data, etc.
  - If data is numeric, operations & some np methods will work
- Various internal conversion options (to_dict, to_string, etc.)
- IO support - save, load (pickle), to_csv, read_csv, etc.
- 3D extension - "Panel"
  - Designed around econmetrics applications, so behaves differently in some cases
  - Hasn't seen much work on the last couple of releases, and will probably be reworked significantly soon, so we'll skip it

Hierarchical Indexing

- The other way of doing 3D (or higher dimensional) data in pandas
- MultiIndex
- Assume sorting for partial indexing and other trickery, but this is not enforced
  - Helper method sortlevel
- Arise naturally out of the grouping functions
- Can be re-ordered (swaplevel, reorder_levels)
- MultiIndexes (and Indexes) can be created from data already in a DataFrame

Time Series

- Special case of series - index is time-related
  - Two types of time index - DatetimeIndex (timestamps - fixed samples in time) & PeriodIndex
- Lots of options for generating sequences, resampling, etc.
- DateOffset objects for useful data manipulations