PANDAS: AN OPEN SOURCE TOOL FOR DATA SCIENCE

LEVI BRIGGS, MATTHEW ERDNER, ALLYSON LESHER, AVI MILLER, JILL SIRIGNANO
COLLEGE OF CHARLESTON DEPARTMENT OF COMPUTER SCIENCE

DATA 495. CAPSTONE PROJECT. APRIL 2020

ABSTRACT

Pandas is an open source software library written for Python that helps simplify data structures and data analysis. It is one of the most popular libraries for data manipulation and is widely used in the field of Data Science. The main theme surrounding our project is to provide Pandas with information they can reflect on their website. The Pandas website has very little information about what someone can do using their package. Our goal was to provide information and visualizations about the big ideas surroundings what Pandas can do and how people in the business and academic world can use it to benefit their project.

Project Highlights

Importing Data

The main use of the pandas package is to bring data from excel, csv, text, json, html, sql, SAS, SPSS, or other IO tools into an interface that has the capability to preform complex data analysis. Given a path to a file or URL, pandas will read in the textual or numerical data and store it in a data frame. When bringing data in from an outside source, you can set a delimiter, specify column data types, name columns, handle dates, and work with indexes in a file in order to make sure that the data you are preparing for analysis is in its proper form.

Data Cleansing

Once you have your imported data in a pandas data frame structure, you can add features to the data frame in order to cleanse the data you are working with. Some of the basic operations that the pandas package has includes but is not limited to adding or selecting indexes (column names), merge, change how missing values are stored and presented, perform basic math operations, and categorize data.

Data Manipulation

You can also add columns and data to an existing data frame. With pandas operations, you can join or concatenate data from two different data frames. Pandas also allows you to reshape the data you brought in from a file if the structure needs to change before data analysis can occur. Finally, pandas will allow you to add a column to a data frame if needed.

> “**pandas** is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.”

ECOSYSTEM

Python

Python is a user-friendly language for beginners and experienced programmers. It is famous for its simple and readable syntax. Python has a large set of intensive libraries making the language versatile and usable for a variety of projects. Second, only to R, Python is the most used language for data science due to its use of big data and cloud computing solutions. You can use python across many domains – websites, apps, hardware, desktop applications, etc. No matter where or how you use Python, it will serve as a reliable and efficient environment for all your programming needs.

Plots

Matplotlib, Bokeh, Altair are three of the plotting tools that can help you to visualize a pandas data frame. Visualization of a data frame can help with exploratory data analysis.

Matplotlib has features within its package for you to customize plots and make unique visualizations from your pandas data frame.

The Bokeh package will allow you to make interactive visualizations from your pandas data frame. Bokeh plots are ideal for displaying on a webpage.

Altair is built on a simple and user friendly API. Altair can also be used to create interactive visualizations.

Distributing

Large datasets and complex computations outside the bandwidth power of pandas can be performed using Dask, a large parallel data frame composed of multiple smaller pandas data frames split along the index.

Interface

Juptyer Notebook / Juptyer Lab, Spyder, and IPython are all examples of user interfaces that work with the pandas packages. Jupyter is an open source web-based environment designed for interactive computing across multiple programming languages. Spyder and IPhython is an interface designed specifically for data analysis in the python language.

Conda

Conda is an open source package manager. Using Conda, you can install and update packages on your computer efficiently. Conda works with any language including python.

NumPy

The NumPy package allows a programmer to perform scientific computing with Python. NumPy can be used with pandas data frames for fast array computing.

xarray

The xarray package makes working with multideminsional arrays efficient. Xarray objects can easily be converted to and from pandas objects.

In 2019, Trivago wanted to specify images that appear next to hotel listings according to keywords in user search. A positive user experience is really important to this company and many users know exactly what they want when searching for available hotels. Specifically for amenities, many users will search for hotels with pools, spas, gyms, rooftops, etc.

Trivago was able to use Pandas in order to have those specific search related images appear; this way users could easily compare amenities and find their ideal hotel. Originally, the team thought about handpicking images to solve this issue, but they soon realized they would have to manually pick an image for every topic someone could possibly search for. This would then have to be applied to every hotel on their site which consists of 3+ million hotels and 100+ million images linked to those listings. They were first able to collect training data from ImageNet, which is real-world data that humans contributed to by answering questions such as "Is this a pool?", "Is this a hot tub?", etc. Correlating a specific model to 'Spa and Wellness', it was able to reach an average precision and recall of about 85-89% for each class (bedroom, gym, hot tub, massage, non spa, pool, sauna, and yoga). This analysis and model allowed images to be placed in categories based on class and whether or not it was a True-Positive or a False-Positive image.

Creators of Zillow were mind blown that no website had the capabilities to provide an accurate selling price estimate.

The main reason Zillow is preferred over most online real estate marketplaces is because of the 'Zestimate' and its capabilities to accurately estimate a home's market value. The company needed some sort of advantage to make them stand out from competitors, especially since the launch of their company occurred during the Great Recession, one of the most severe recessions in the U.S.

Zillow used Pandas to create this algorithm as well as provide a visual guide to all users. The founders of the company found it unacceptable that when they were looking to purchase new homes, they had to crunch the numbers themselves to find these costs, and soon realized they weren't the only ones struggling through this. Pandas helped create a starting point to this idea, which is not an exact valuation, but instead an estimate to start out with. They were able to apply this method to every house in a neighborhood, raising much competition between home sellers and buyers.

Many other companies, such as Trulia, began to follow in their footsteps by creating their own algorithms, but Zillow still comes out on top with it being one of their most engaging features. While continuing to perfect the algorithm, as of now the Zestimate can compute accuracy within 10% of the actual value. This feature continues to develop as more information is added over time. Today, they provide Zestimates for 110 million+ homes in the U.S. along with rental Zestimates for 100 million+ rental homes.

Doordash needed to find the optimal path between the number of products to be delivered, the number of drivers, and the number of stops in between - The Routing Problem’.

Doordash used pandas to solve an optimization problem. Which driver should pick up a delivery while considering the amount of time the driver will have to wait for the delivery, the amount of food within the delivery, the time to the business, and finally the time to the resident. The solution to the routing problem allows Doordash to optimize the time it takes for food to get your door – as well as everyone elses.