

Why Python is a good tool for data mining

Xiaolu Xiong
CS548

Why Python is a ~~great~~ **better** tool for data mining

Xiaolu Xiong
CS548

References

- Python (programming language): [http://en.wikipedia.org/wiki/Python_\(programming_language\)](http://en.wikipedia.org/wiki/Python_(programming_language))
- When You Write Your Essays in Programming Languages: <http://www.somethingofthatilk.com/index.php?id=135>
- sklearn.cluster.KMeans: <http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans>
- Pricing and Licensing: <http://www.mathworks.com/pricing-licensing/>
- Pricing Options: <http://rapidminer.com/pricing/>
- PythonForArtificialIntelligence: <https://wiki.python.org/moin/PythonForArtificialIntelligence>
- How To Read And Parse CSV File In Java: <http://www.mkyong.com/java/how-to-read-and-parse-csv-file-in-java/>

Tools for data mining

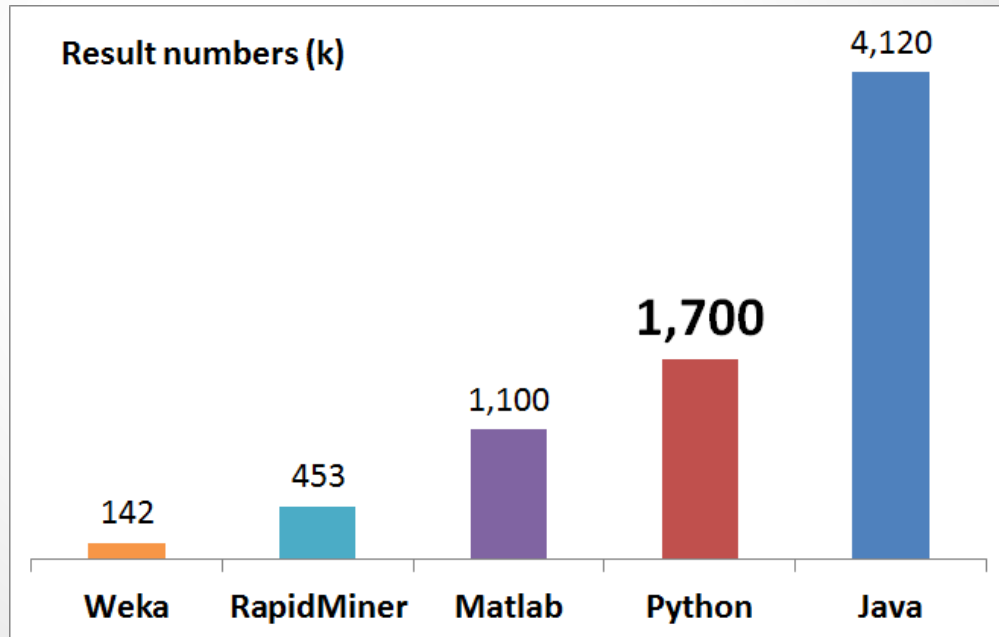
We have learned to use a lot tools

Weka, RapidMiner, Matlab ...

What tools are being used in the real word?

Tools for data mining?

Ask Google: *data mining job requirements* +
<*tool name*>

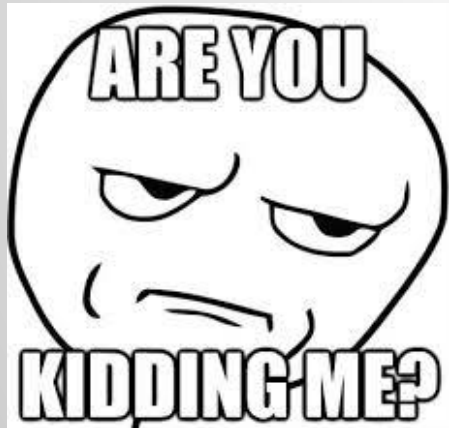


Why I think Python is a better tool?

Free!

Matlab: \$50 - \$2600

RapidMiner: \$999/yr - \$2999/yr!!



Why I think Python is a better tool?

Simple: better than Java

Readability is the core philosophy

```
# load data from a csv file
data = array([map(float, row) \
              for row in reader(open("ads.csv", "r"))])
```

VS

```
public class ReadCSV {

    public static void main(String[] args) {

        ReadCSV obj = new ReadCSV();
        obj.run();

    }

    public void run() {

        String csvFile = "/Users/mkyong/Downloads/GeoIPCountryWhois.csv";
        BufferedReader br = null;
        String line = "";
        String cvsSplitBy = ",";

        try {

            br = new BufferedReader(new FileReader(csvFile));
            while ((line = br.readLine()) != null) {

                // use comma as separator
                String[] country = line.split(cvsSplitBy);

                System.out.println("Country [code= " + country[4]
                    + " , name=" + country[5] + "]);

            }

        }

    }

}
```

Why I think Python is a better tool?

Powerful: better than many

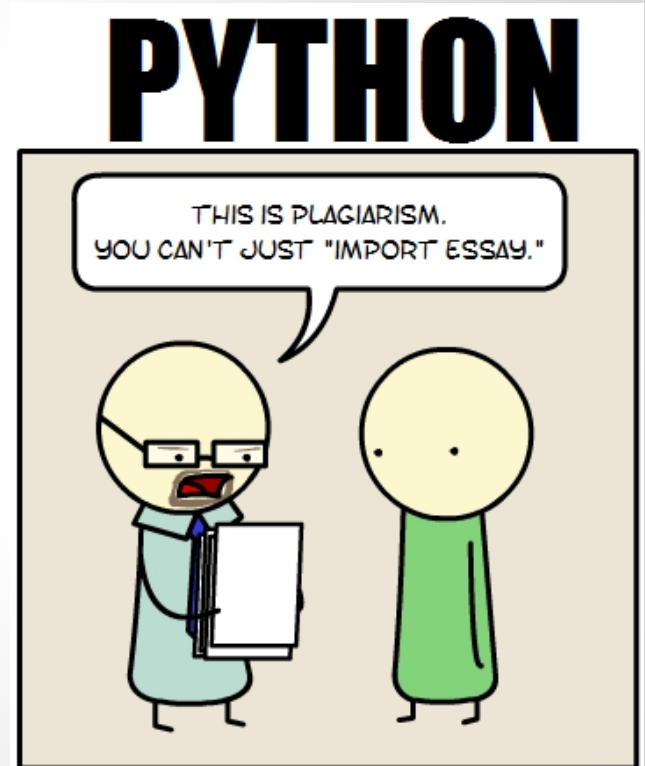
Anything you can think of, there is a library for that - over 20 libraries for Data mining

Web programming

Large scale data processing for big data

Image/Audio processing

...



Why I think Python is a better tool?

Object Oriented:

Because it is a real programming language

Open Source:

So you can help to improve it

Example: K-means using Python

1. Load data from a csv file
2. Define K-means clustering
3. Fit the data to K-means
4. Plot the clusters in a 3D space

Example: K-means using Python

Load csv data:

```
# load data from a csv file
data = array([map(float, row) for row in reader(open("ads.csv", "r"))
```

Define K-means clustering:

```
# define kmeans parameters
k_means = KMeans(init='k-means++', n_clusters=2, n_init=10)
```

Example: K-means using Python

Fit the data to K-means:

```
# fit the data with kmeans clustering  
k_means.fit(data)
```

Return the labels of every points

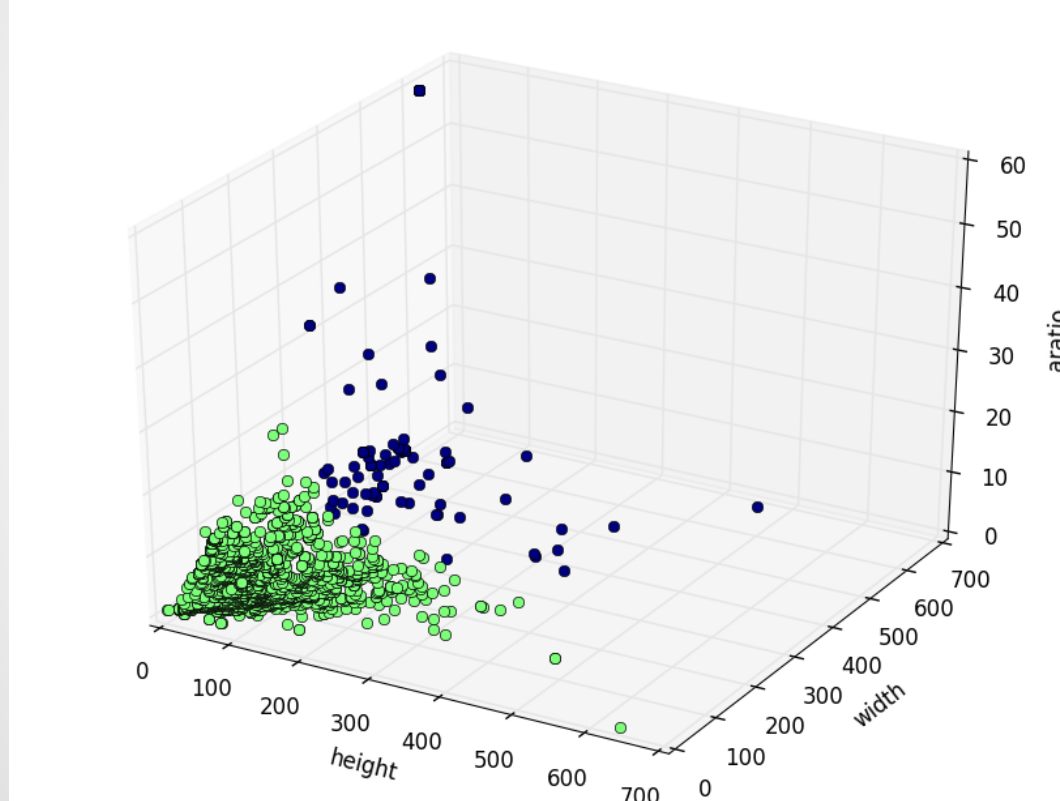
```
# get the cluster labels  
label = k_means.labels_
```

Example: K-means using Python

Plot the clusters in a 3D space

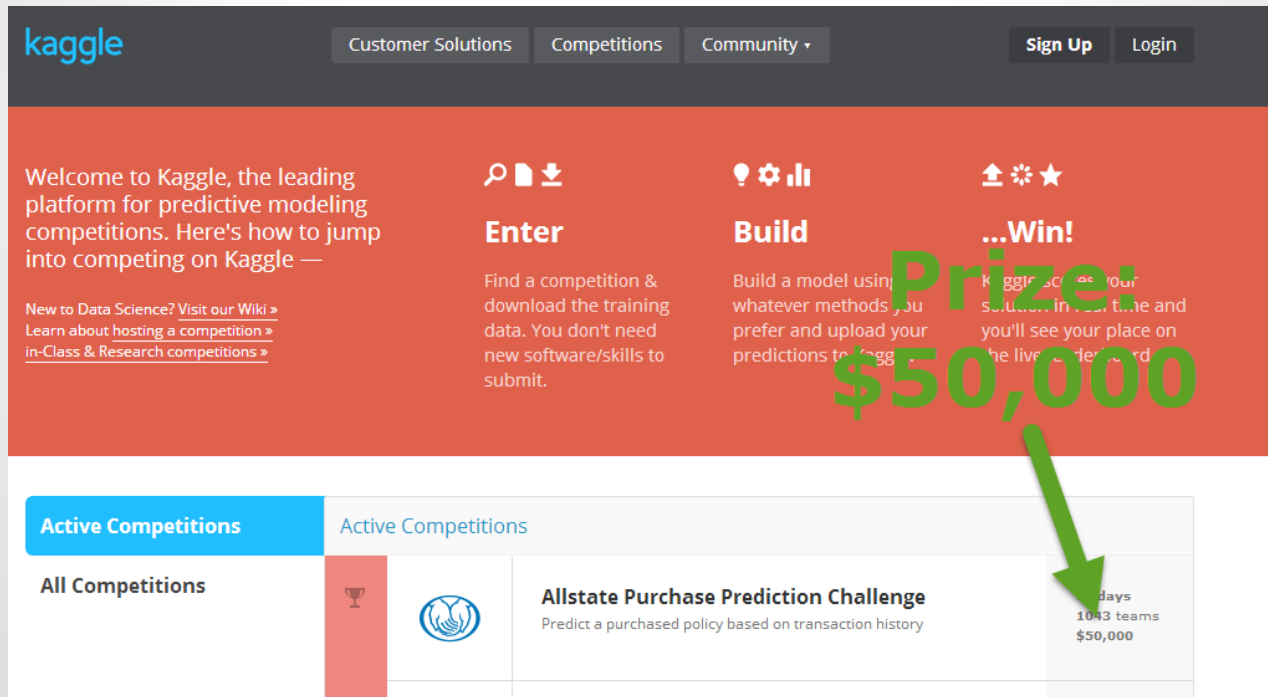
```
# plot the clusters in a 3d space
fig = plt.figure()
ax = plt.Axes3D(fig)
for l in np.unique(label):
    ax.plot3D(data[label == l, 0], data[label == l, 1], data[label == l, 2],
              'o', color=plt.cm.jet(np.float(l) / np.max(label + 1)))
```

Example: K-means using Python



Thank you and one more thing

Need a dataset? Try: <https://www.kaggle.com/>



The image shows a screenshot of the Kaggle website homepage. The top navigation bar includes the Kaggle logo, links for 'Customer Solutions', 'Competitions', and 'Community', and buttons for 'Sign Up' and 'Login'. The main content area is divided into three columns: 'Enter', 'Build', and '...Win!'. The 'Enter' column describes finding competitions and downloading data. The 'Build' column describes building models and uploading predictions. The '...Win!' column describes seeing one's place on a leaderboard. A large green arrow points from the text '\$50,000' to a competition listing in the 'Active Competitions' section. The listing is for the 'Allstate Purchase Prediction Challenge', which has 1043 teams and a prize of \$50,000.

Welcome to Kaggle, the leading platform for predictive modeling competitions. Here's how to jump into competing on Kaggle —

New to Data Science? Visit our Wiki »
Learn about hosting a competition »
In-Class & Research competitions »



Enter
Find a competition & download the training data. You don't need new software/skills to submit.

Build
Build a model using whatever methods you prefer and upload your predictions to Kaggle.

...Win!
Kaggle scores your solution. In real time and you'll see your place on the live leaderboard.

Prize: \$50,000

Active Competitions

Active Competitions	Active Competitions
All Competitions	
	
	Allstate Purchase Prediction Challenge Predict a purchased policy based on transaction history
	Days 1043 teams \$50,000