Stock Price Outlier Detection

Steps

Use the `yahoo-finance` api to collect stock price information for each company. The NYSE symbols for each company are required to obtain the stock price data.

```
from yahoo_finance import Share

symbols = ['LNKD', 'GOOG', 'FB', 'MSFT', 'AMZN']
company_names = ['LinkedIn', 'Google', 'Facebook', 'Microsoft', 'Amazon']
companies = []
for symbol in symbols:
    companies.append(Share(symbol))
```

Obtain company stock price in intervals of one minute for 30 minutes. Store these values on a new line in separate files. These files are stored in the directory `stock-values/`

```
def get_current_price(companies, company_names):
  for i, company in enumerate(companies):
      company.refresh()
      price = company.get_price()

      # Store the price in separate text files.
      with open('Stock-Prices/' + company_names[i] + '_shares.txt', 'a') as f:
          f.write(price + '\n')

start_time = time.time()
count = 0

while count < 30:
  get_current_price(companies, company_names)
  count += 1
  # intervals of 1 min
  time.sleep(60)
```

With the collected stock data, we create an RDD of stock prices for every company. Since the prices are stored as strings in the `.txt` files, we need to convert them into floating point numbers.

```
from os import listdir

files = [ f for f in listdir('./stock-values') if '.txt' in f ]
for f in files:
  rdd = sc.textFile('stock-values/' + f)
```

```
    prices = rdd.map(lambda s : float(s))
```

We now find the mean and standard deviation of the stock prices for each company using the `.stats()` function of RDD.

```
stats = prices.stats()
mean = stats.mean()
stdev = stats.stdev()
```

To find the outliers we find values which are greater than 2 standard deviations away from the mean.

```
outliers = prices.filter(lambda p: abs(p - mean) > 2 * stdev)
```

And finally we print the values which are outliers in the terminal itself.

```
Outliers by Company:
Amazon_shares.txt: [507.9, 507.9]
Facebook_shares.txt: []
Google_shares.txt: []
LinkedIn_shares.txt: [100.62, 100.62, 100.62]
Microsoft_shares.txt: [49.1]
```