Assignment 2

Jiarong Ye

October 10, 2018

Import packages

```
In [117]: import redis
          from urllib.request import urlopen
          import pandas as pd
          import re
          from bs4 import BeautifulSoup
```

scrape car retail information from a chosen car dealer website

Titles

```
In [245]: url = 'http://www.statecollege.com/auto/dealers/stocker-chevrolet,9803/'
          page = urlopen(url=url)
          soup = BeautifulSoup(page, 'html.parser')
          titles = soup.findAll('strong', {"class": "title"})
          titles_df = pd.DataFrame(list(map(lambda x: x.getText(), titles)), columns=['Title'])
```

Embedded urls

```
In [246]: urls_df = pd.DataFrame(['http://www.statecollege.com/'+ a['href'] for i in soup.findAl
```

Addresses, phones of sellers & descriptions of cars

```
In [247]: addresses = []
          phones = []
          descriptions = []
          for url in urls_df.values:
              page = urlopen(url=url[0])
              soup = BeautifulSoup(page, 'html.parser')
              for address in soup.findAll('p', {'class': 'business_address_address'}):
                  addresses.append(re.sub('\s+', ' ', address.getText().replace('\n', '')))
              for phone in soup.find('p', {'class': 'business_address_phone'}):
                  phones.append(phone)
              for i, c in enumerate(soup.findAll('p')):
                  if i==5:
                      descriptions.append(c.getText())
          address_df = pd.DataFrame(addresses, columns=['address'])
          phones_df = pd.DataFrame(phones, columns=['phone'])
          descriptions_df = pd.DataFrame(descriptions, columns=['description'])
```

Other attributes

```
In [248]: tables = []
          for url in urls_df.values:
              page = urlopen(url=url[0])
              soup = BeautifulSoup(page, 'html.parser')
              for table in soup.findAll('table', {'class': 'auto_detail_data_table'}):
                  table = pd.read_html(str(table))
                  tables.append(table)
```

```
In [249]: attributes = ['Year', 'Mileage', 'Make', 'Model', 'Trim', 'Style', 'Engine', 'Exterior
          attributes_dict = {}
          for i in attributes:
              attributes_dict[i] = []
          for table in tables:
              table = table[0]
              for attribute in attributes:
                  attributes_dict[attribute].append(table[table.iloc[:, 0]==attribute+':'].iloc[
          dataset = pd.concat([titles_df, address_df, phones_df, pd.DataFrame(attributes_dict),
```

Combine all attributes

```
In [257]: dataset
```

Out[257]:

	Title	address	phone	Year	Mileage	Make	Model	Trim
0	2014 Subaru Outback	701 Benner Pike State College PA, 16801	(866) 235-0270	2014	75655	Subaru	Outback	2.5i Premium
1	2013 Subaru XV Crosstrek	701 Benner Pike State College PA, 16801	(866) 235-0270	2013	69331	Subaru	XV Crosstrek	Premium
2	2014 Chevrolet Cruze	701 Benner Pike State College PA, 16801	(866) 235-0270	2014	61147	Chevrolet	Cruze	LS
3	2015 Chevrolet Malibu	701 Benner Pike State College PA, 16801	(866) 235-0270	2015	25757	Chevrolet	Malibu	LS
4	2015 Chevrolet Silverado 1500	701 Benner Pike State College PA, 16801	(866) 235-0270	2015	41869	Chevrolet	Silverado 1500	High Country
5	2016 Chevrolet Malibu Limited	701 Benner Pike State College PA, 16801	(866) 235-0270	2016	52582	Chevrolet	Malibu Limited	LS
6	2016 Subaru Forester	701 Benner Pike State College PA, 16801	(866) 235-0270	2016	16994	Subaru	Forester	2.5i
7	2015 Chevrolet Silverado 1500	701 Benner Pike State College PA, 16801	(866) 235-0270	2015	49503	Chevrolet	Silverado 1500	LT
8	2017 Subaru Legacy	701 Benner Pike State College PA, 16801	(866) 235-0270	2017	2	Subaru	Legacy	Limited
9	2017 Subaru Outback	701 Benner Pike State College PA, 16801	(866) 235-0270	2017	3	Subaru	Outback	Limited

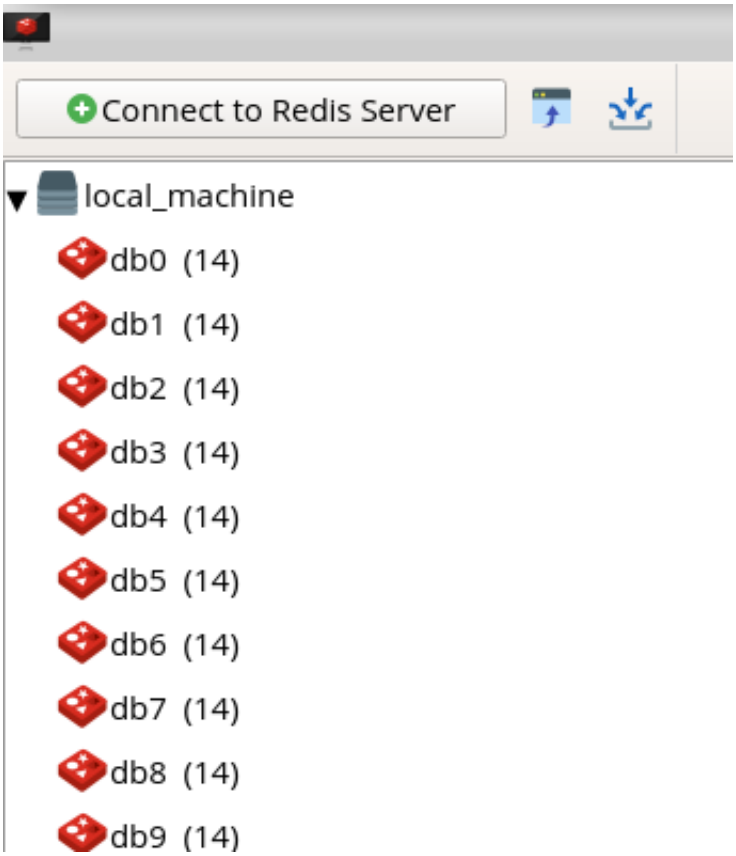
	Style	Engine	Exterior Color	Interior Color

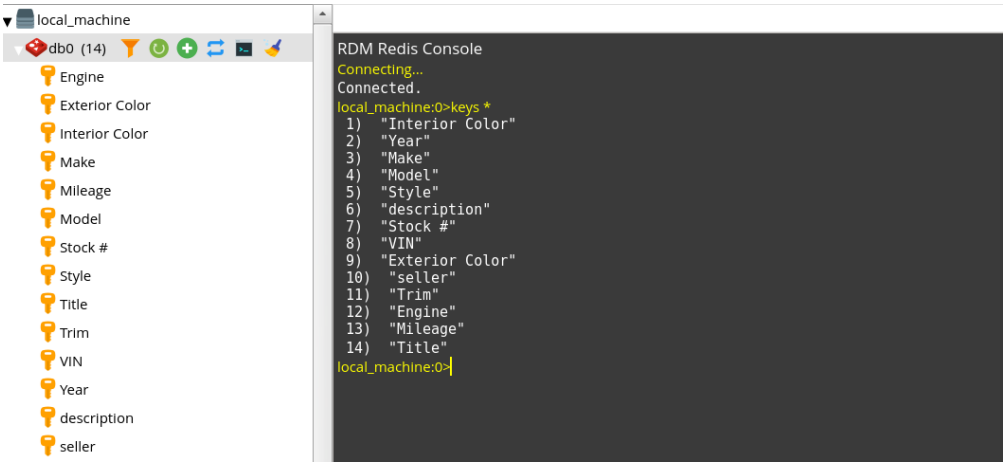
```
0    Sport Utility    4 -  Twilight Blue Metall                 Black
1    Station Wagon    4 -  Tangerine Orange Pea                 Black
2          4dr Car    4 -  Atlantis Blue Metall  Jet Black/Medium Tit
3          4dr Car    4 -   Ashen Gray Metallic    Jet Black/Titanium
4  Crew Cab Pickup    8 -    Deep Ruby Metallic                Saddle
5          4dr Car    4 -  Champagne Silver Met    Jet Black/Titanium
6    Sport Utility    4 -   Ice Silver Metallic                 Black
7  Crew Cab Pickup    8 -          Summit White             Jet Black
8          4dr Car    6 -     Tungsten Metallic            Warm Ivory
9    Sport Utility    4 -  Twilight Blue Metall           Slate Black

                 VIN  Stock #  \
0  4S4BRBCC4E3219562  606614A
1  JF2GPACCXD1843307  204842B
2  1G1PA5SG4E7343131  204275A
3  1G11B5SL2FF313357  204597A
4  3GCUKTEC9FG401065  204778A
5  1G11B5SAXGF135777   15498A
6  JF2SJAAC3GG451169  606499A
7  3GCUKREC9FG459437  204882A
8  4S3BNEN6XH3029939   604992
9  4S4BSANC0H3391975   605565

                                          description
0  CARFAX One-Owner. Clean CARFAX. Blue 2014 Suba...
1  Clean CARFAX. Orange 2013 Subaru XV Crosstrek ...
2  CARFAX One-Owner. Clean CARFAX. Blue 2014 Chev...
3  CARFAX One-Owner. Clean CARFAX. Grey 2015 Chev...
4  CARFAX One-Owner. Clean CARFAX. Burgundy 2015 ...
5  CARFAX One-Owner. Gold 2016 Chevrolet Malibu L...
6  CARFAX One-Owner. Clean CARFAX. Ice Silver Met...
7  CARFAX One-Owner. Clean CARFAX. White 2015 Che...
8  $750 off MSRP!We at Stocker Chevrolet apprecia...
9  We at Stocker Chevrolet appreciate your time a...
```

Insert data into Redis

```
In [260]: seller_address=''
          seller_phone=''
          for idx, t in enumerate(dataset.values):
              r = redis.StrictRedis(host='localhost', port=6379, db=idx)
              for i, attribute in enumerate(dataset.columns):
                  if attribute=='address':
                      seller_address = t[i]
                  elif attribute=='phone':
                      seller_phone = t[i]
                  else:
                      r.set(attribute, t[i])
```

```
r.hmset('seller', {
    'address':seller_address,
    'phone':seller_phone
})
```





Data model design

- Attributes
    * Title
    * Year
    * Engine
    * Exterior Color
    * Interior Color
    * Make
    * Mileage
    * Model
    * Stock
    * Style
    * Trim
    * VIN
    * Description
    * Seller
        * address
        * phone

The data model covers all basic attributes of a car that could be found on the retail website and that buyers should be aware of.

Display sample results

```
In [279]: for idx in range(3):
              r = redis.StrictRedis(host='localhost', port=6379, db=idx)
              for key in r.keys():
                  key = key.decode("utf-8")
                  try:
                      value = r.get(key)
                  except Exception as e:
                      value = r.hgetall(key)
                  print(key,' : ', value)
              print('\n')
```

```
Interior Color  :  b'Black'
Year  :  b'2014'
Make  :  b'Subaru'
Model  :  b'Outback'
Style  :  b'Sport Utility'
description  :  b"CARFAX One-Owner. Clean CARFAX. Blue 2014 Subaru Outback 2.5i Premium AWD 6-Sp
Stock #  :  b'606614A'
VIN  :  b'4S4BRBCC4E3219562'
Exterior Color  :  b'Twilight Blue Metall'
seller  :  {b'address': b' 701 Benner Pike State College PA, 16801 ', b'phone': b'(866) 235-0270
Trim  :  b'2.5i Premium'
Engine  :  b'4 -'
```

```
Mileage  :  b'75655'
Title  :  b'2014 Subaru Outback'


Interior Color  :  b'Black'
Year  :  b'2013'
Make  :  b'Subaru'
Model  :  b'XV Crosstrek'
Style  :  b'Station Wagon'
description  :  b"Clean CARFAX. Orange 2013 Subaru XV Crosstrek 2.0i Premium AWD 5-Speed Manual
Stock #  :  b'204842B'
VIN  :  b'JF2GPACCXD1843307'
Exterior Color  :  b'Tangerine Orange Pea'
seller  :  {b'address': b' 701 Benner Pike State College PA, 16801 ', b'phone': b'(866) 235-0270
Trim  :  b'Premium'
Engine  :  b'4 -'
Mileage  :  b'69331'
Title  :  b'2013 Subaru XV Crosstrek'


Interior Color  :  b'Jet Black/Medium Tit'
Year  :  b'2014'
Make  :  b'Chevrolet'
Model  :  b'Cruze'
Style  :  b'4dr Car'
description  :  b"CARFAX One-Owner. Clean CARFAX. Blue 2014 Chevrolet Cruze LS FWD 6-Speed Autom
Stock #  :  b'204275A'
VIN  :  b'1G1PA5SG4E7343131'
Exterior Color  :  b'Atlantis Blue Metall'
seller  :  {b'address': b' 701 Benner Pike State College PA, 16801 ', b'phone': b'(866) 235-0270
Trim  :  b'LS'
Engine  :  b'4 -'
Mileage  :  b'61147'
Title  :  b'2014 Chevrolet Cruze'
```