Python - Analiza danych z modułem PANDAS

www.udemy.com (http://www.udemy.com) (R)

LAB - S04-L005-Operacje na kolumnach tekstowych

1. Zaimportuj moduł pandas i numpy, nadaj im standardowe aliasy. Do zmiennej **fortune** wczytaj zawartość pliku **Fortune_500_2017.csv**. Pobierz tylko następujące kolumny: **'Rank','Title','Industry','Hqlocation','Sector'** Wyświetl nagłówek obiektu data frame.
2. Zmień zawartość kolumny **Sector** tak, aby była zapisana wielkimi literami. Wyświetl nagłówek **fortune**
3. Wyszukaj te firmy, które w kolumnie **Industry** zawierały napis **comp**. Zadbaj o to aby odnalezione były dane niezależnie od tego czy w kolumnie tekst jest napisany małymi czy wielkimi literami.
4. Dodaj do **fortune** dwie nowe kolumny **'City','State'**, które powstaną przez rozbicie kolumny **Hqlocation** ze względu na przecinek. Wyświetl nagłówek **fortune**
5. Chcemy do obiektu **frame** dodać kolumnę **IndustryShort**, która powstanie wskutek połączenia pierwszych liter wyrazów znajdujących się w kolumnie **Industry** (nie przejmuj się nawiasami, przecinkami itp. W tym celu:
    - Stwórz funkcje **BuildShortcut**, która jako parametr przyjmie wiersz
    - Funkcja ma pobrać wartość z kolumny **Industry** i zapisać ją w zmiennej **industry**
    - Zadeklarować pusty napis **result**
    - Dla każdego napisu znajdującego się na liście powstałej w skutek rozbicia **industry** ze względu na spacje
    - Pobrać pierwszą literę słowa i dodać ją do napisu **result**
    - Zwrócić napis **result**
6. Sprawdź działanie funkcji dla słownika: **{'Industry':'Factory Under Newspaper'}**
7. Dodaj do **fortune** kolumnę **IndustryShort**, która zawierać będzie skrót wyznaczony przez funkcję **BuildShortcut** dla każdego wiersza. Wyświetl nagłówek **fortune**.

Rozwiązania:

Poniżej znajdują się propozycje rozwiązań zadań. Prawdopodobnie istnieje wiele dobrych rozwiązań, dlatego jeżeli rozwiązujesz zadania samodzielnie, to najprawdopodobniej zrobisz to inaczej, może nawet lepiej :) Możesz pochwalić się swoimi rozwiązaniami w sekcji Q&A

In [1]:
```
import pandas as pd
import numpy as np
fortune = pd.read_csv("Fortune_500_2017.csv",usecols=['Rank','Title','Industry',
                                                      'Hqlocation','Sector'])
fortune.head(5)
```

Out[1]:

	Rank	Title	Sector	Industry	Hqlocation
0	1	Walmart	Retailing	General Merchandisers	Bentonville, AR
1	2	Berkshire Hathaway	Financials	Insurance: Property and Casualty (Stock)	Omaha, NE
2	3	Apple	Technology	Computers, Office Equipment	Cupertino, CA
3	4	Exxon Mobil	Energy	Petroleum Refining	Irving, TX
4	5	McKesson	Wholesalers	Wholesalers: Health Care	San Francisco, CA

In [2]:
```
fortune["Sector"] = fortune["Sector"].str.upper()
fortune.head()
```

Out[2]:

	Rank	Title	Sector	Industry	Hqlocation
0	1	Walmart	RETAILING	General Merchandisers	Bentonville, AR
1	2	Berkshire Hathaway	FINANCIALS	Insurance: Property and Casualty (Stock)	Omaha, NE
2	3	Apple	TECHNOLOGY	Computers, Office Equipment	Cupertino, CA
3	4	Exxon Mobil	ENERGY	Petroleum Refining	Irving, TX
4	5	McKesson	WHOLESALERS	Wholesalers: Health Care	San Francisco, CA

In [3]:
```
fortune[fortune["Industry"].str.lower().str.contains('comp')].head()
```

Out[3]:

	Rank	Title	Sector	Industry	Hqlocation
2	3	Apple	TECHNOLOGY	Computers, Office Equipment	Cupertino, CA
27	28	Microsoft	TECHNOLOGY	Computer Software	Redmond, WA
40	41	Dell Technologies	TECHNOLOGY	Computers, Office Equipment	Round Rock, TX
46	47	Intel	TECHNOLOGY	Semiconductors and Other Electronic Components	Santa Clara, CA
60	61	HP	TECHNOLOGY	Computers, Office Equipment	Palo Alto, CA

In [4]:
```
fortune[['City','State']] = fortune["Hqlocation"].str.split(",",expand=True)
fortune.head()
```

Out[4]:

	Rank	Title	Sector	Industry	Hqlocation	City	State
0	1	Walmart	RETAILING	General Merchandisers	Bentonville, AR	Bentonville	AR
1	2	Berkshire Hathaway	FINANCIALS	Insurance: Property and Casualty (Stock)	Omaha, NE	Omaha	NE
2	3	Apple	TECHNOLOGY	Computers, Office Equipment	Cupertino, CA	Cupertino	CA
3	4	Exxon Mobil	ENERGY	Petroleum Refining	Irving, TX	Irving	TX
4	5	McKesson	WHOLESALERS	Wholesalers: Health Care	San Francisco, CA	San Francisco	CA

In [5]:
```
def BuildShortcut(row):
    industry = row["Industry"]
    result = ''
    for i in industry.split(' '):
        result += i[0]
    return result
```

In [6]:
```
BuildShortcut({'Industry':'Factory Under Newspaper'})
```

Out[6]:
```
'FUN'
```

In [7]:
```
fortune['IndustryShort'] = fortune.apply(BuildShortcut,axis=1)
fortune.head()
```

Out[7]:

	Rank	Title	Sector	Industry	Hqlocation	City	State	IndustryShort
0	1	Walmart	RETAILING	General Merchandisers	Bentonville, AR	Bentonville	AR	GM
1	2	Berkshire Hathaway	FINANCIALS	Insurance: Property and Casualty (Stock)	Omaha, NE	Omaha	NE	IPaC(
2	3	Apple	TECHNOLOGY	Computers, Office Equipment	Cupertino, CA	Cupertino	CA	COE
3	4	Exxon Mobil	ENERGY	Petroleum Refining	Irving, TX	Irving	TX	PR

	Rank	Title	Sector	Industry	Hqlocation	City	State	IndustryShort
				Wholesalers: Health	San	San		

In [ ]: