Shopee Scraping

February 25, 2018

1 Scrape Chinese Companies' Ratings

from Shopee.com

```
In [3]: # Import Libraries
        from bs4 import BeautifulSoup
        import string
        import pandas as pd
        from googletrans import Translator
```

1.1 Try Oppo's Page on Shopee

Categories: 1. Response Rate (Persentase Chat Dibalas) 2. Number of Products (Produk) 3. Number of Followers (Pengikut) 4. Rating (of Products) (Penilaian)

```
In [4]: with open("shopee.html") as fp:
            shop_soup = BeautifulSoup(fp, 'lxml')

In [8]: shop_att = shop_soup.find_all('div', class_='section-seller-overview__item')

In [9]: # Getting all attributes
        att = []
        value = []
        title = shop_soup.title.text

        for container in shop_att:
            x = container.find('div', class_='section-seller-overview__item-text-name').text
            att.append(x)

            y = container.find('div', class_='section-seller-overview__item-text-value').text
            value.append(y)

In [10]: # Att has some weird string attached to it. Replace.
         for i in range(len(att)):
             att[i] = att[i].replace(':\xa0', '')
```

```
In [11]: # Convert to dataframe
         shop_df = pd.DataFrame({'shop name': title,
                                 att[0]: value[0],
                                 att[1]: value[1],
                                 att[2]: value[2],
                                 att[3]: value[3],
                                 att[4]: value[4],
                                 att[5]: value[5],
                                 att[6]: value[6],
                                 att[7]: value[7]}, index=[0])
         shop_df
```

```
Out[11]:   Persentase Chat Dibalas Waktu Chat Dibalas      bergabung masa pengemasan  \
         0                     92%       hitungan jam  15 bulan lalu        2-3 hari

           mengikuti pengikut                    penilaian produk  \
         0         2    56297  4.8 dari 5 (1145 penilaian)     19

                                                  shop name
         0  Toko Online OPPO Indonesia Official Store | Sh...
```

```
In [12]: # Translate to english
         ts = Translator()

         for i in range(len(att)):
             print(ts.translate(att[i]).text, ":", ts.translate(value[i]).text)
```

```
Maced Chat Percentage : 92%
product : 19
follow : 2
Chat Time Replied : hours count
join : 15 months ago
followers : 56297
assessment : 4.8 out of 5 (1145 ratings)
packaging time : 2-3 days
```

1.2 Generalizing to other company profiles

Find other Chinese companies in Shopee!

```
In [5]: def open_html(html_file):
            with open(html_file) as fp:
                soup = BeautifulSoup(fp, 'lxml')

            return soup

        vivo_shop = open_html("vivo_shopee.html")
```

1.3 Updating company lists

1. Oppo
2. Vivo
3. Xiaomi
4. TCL (The Creative Life)
5. Lenovo
6. Huawei

```
In [19]: # Try read filename names

         # html_list.txt contains the name of html files stored (Currently, OPPO, VIVO, and Xiao
         with open("html_list.txt", encoding="utf8") as f:
             data = f.readlines()

         soup = []

         for i in range(len(data)):
             data[i] = data[i].replace('\n', '')
             soup.append(open_html(data[i]))
```

```
In [7]: def shop_attributes(shop_soup):
            shop_att = shop_soup.find_all('div', class_='section-seller-overview__item')
            att = []
            value = []
            title = shop_soup.title.text

            for container in shop_att:
                x = container.find('div', class_='section-seller-overview__item-text-name').text
                att.append(x)

                y = container.find('div', class_='section-seller-overview__item-text-value').tex
                value.append(y)

            for i in range(len(att)):
                att[i] = att[i].replace(':\xa0', '')

            return att, value, title
```

```
In [20]: def create_df(soup):
             shop_name = []
             response_rate = []
             product = []
             following = []
             time_to_reply = []
             joined = []
             rating = []
             followers = []
```

```
    for i in range(len(soup)):
        att, value, title = shop_attributes(soup[i])
        shop_name.append(title)
        response_rate.append(value[0])
        product.append(value[1])
        following.append(value[2])
        time_to_reply.append(value[3])
        joined.append(value[4])
        followers.append(value[5])
        rating.append(value[6])

    # Create dataframe
    shop_df = pd.DataFrame({'shop name': shop_name,
                            att[0]: response_rate,
                            att[1]: product,
                            att[2]: following,
                            att[3]: time_to_reply,
                            att[4]: joined,
                            att[5]: followers,
                            att[6]: rating})

    return shop_df

shop_df = create_df(soup)
shop_df
```

Out[20]:

	Persentase Chat Dibalas	Waktu Chat Dibalas	bergabung	mengikuti	pengikut	penilaian produk	shop name
0	92%	hitungan jam	15 bulan lalu	2	56297	4.8 dari 5 (1145 penilaian)	19	Toko Online OPPO Indonesia Official Store | Sh...
1	65%	hitungan hari	6 bulan lalu	6	20680	4.4 dari 5 (576 penilaian)	9	Toko Online Vivo Mobile Official Store | Shope...
2	93%	hitungan jam	43 hari lalu	1	69668	4.8 dari 5 (8635 penilaian)	52	Toko Online Xiaomi Official Store | Shopee Ind...
3	19%	hitungan jam	6 bulan lalu	1	867	0 penilaian	11	Toko Online TCL Official Shop | Shopee Indonesia
4	57%	hitungan jam	9 bulan lalu	1	6529	4.5 dari 5 (46 penilaian)	21	Toko Online Lenovo Official Shop | Shopee Indo...
5	27%	hitungan jam	9 bulan lalu	1	2130	4.0 dari 5 (136 penilaian)	6	

```
5  Toko Online Huawei Mobile Broadband Official S...
```

In [17]:
```
# Export to Excel
writer = pd.ExcelWriter('CN_in_IDN_consumer_products.xlsx')
shop_df.to_excel(writer,sheet_name='Sheet1')
writer.save()
```