Sentiment analysis On the companies reviews

Natural Language Processing

Project Work

Jayanth Shivanandappa (3000530)

31.08.2022

Professor:
Prof.Dr. Winfried Bantel

Hochschule Aalen

Studying Machine learning and Data analysis

The topic of "Natural Language Processing" describes a large part of the field of "Machine Learning & Data Analytics". Natural Language Processing is used, for example, in large data processing to analyze large data sets with linguistic approaches such as Pattern matching, word embedding, tf-idf, stemming or lemmatization. The examination in the lecture "Natural Language Processing", which is held by Prof. Dr. Winfried Bentel, requires a project work Students were allowed to choose any data set according to their preferences.

I have taken job website called Ambition box (https://www.ambitionbox.com/list-of-companies?page=1) and I have web scraped the data using beautiful soup. I have scraped about 200 pages of data.

```
: 1 headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 6.3; Win64; x64)AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.
  2
  3 webpage=requests.get ('https://www.ambitionbox.com/list-of-companies?page=1',headers=headers).text
```

Commands to fetch the data from the URL

```
soup = BeautifulSoup(webpage,'lxml')
```

The parser used is lxml parser

```
company = soup.find_all('div',class_='company-content-wrapper')
```

Main class of the wrapper

```
1  name =[]
2  rating =[]
3  reviews = []
4  ctype = []
5  hq = []
6  old = []
7  employees=[]
8  for i in company:
9
10     name.append(i.find('h2').text.strip())
11     rating.append(i.find('p',class_='rating').text.strip())
12     reviews.append(i.find('a',class_='review-count').text.strip())
13     ctype.append(i.find_all('p',class_='infoEntity')[0].text.strip())
14     hq.append(i.find_all('p', class_='infoEntity')[0].text.strip())
15     old.append(i. find_all('p', class_='infoEntity')[0].text.strip())
16     employees.append(i.find_all('p', class_='infoEntity')[0].text.strip())
17 #d={' name': name, 'rating': rating, 'reviews': reviews, 'type':ctype, 'hq':hq, 'old' :old, 'employees':
18 d={'name':name,'rating':rating,'reviews':reviews,'type':ctype,'hq':hq,'old':old,'employees':employees}
19
20 df=pd.DataFrame(d)
```

In the Beginning I started the scrape the name, rating, reviews, type, hq,old,employees for the single page.

```
final = pd.DataFrame()

for j in range(1,180):

    url = 'https://www.ambitionbox.com/list-of-companies?page={}'.format(j)

    headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 6.3; Win64; x64)AppleWebKit/537.36 (KHTML, like Gecko) Chrome
    webpage=requests.get (url,headers=headers).text

    soup = BeautifulSoup(webpage,'lxml')
    company = soup.find_all('div',class_='company-content-wrapper')

    name =[]
    rating =[]
    reviews = []
    ctype = []
    hq = []
    old = []
    employees=[]
    for i in company:

        name.append(i.find('h2').text.strip())
        rating.append(i.find('p',class_='rating').text.strip())
        reviews.append(i.find('a',class_='review-count').text.strip())
        ctype.append(i.find_all('p',class_='infoEntity')[0].text.strip())
        hq.append(i.find_all('p', class_='infoEntity')[0].text.strip())
        old.append(i. find_all('p', class_='infoEntity')[0].text.strip())
        employees.append(i.find_all('p', class_='infoEntity')[0].text.strip())
     #d={' name': name, 'rating': rating, 'reviews': reviews, 'type':ctype, 'hq':hq, 'old' :old, 'employees':
    d={'name':name,'rating':rating,'reviews':reviews,'type':ctype,'hq':hq,'old':old,'employees':employees}

    df=pd.DataFrame(d)

    final = final.append(df)
```

Later I did pagination and I scraped about 200 pages of data. With the scraped data with reviews and rating I am going to build model for sentiment Analysis.