

Kulvir Singh

19BCE2074

Digital Assignment 1

Question 1

Problem Statement :

Write a program (using nltk toolkit in python environment) to tokenize

- a) Sentence
- b) A paragraph

Procedure :

Install the nltk toolkit in python environment using the pip install command. Import nltk to the code file. Download the class stopwords and punkt from nltk toolkit. Import stopwords from nltk.corpus and word_tokenize from nltk.tokenize. Open a file using open method and store its contents in a variable using read method. Create a variable which stores the stopwords that are fetched from nltk. Use the tokenize method to create tokens of the document/file read. Loop through the tokens and filter out the stopwords. Display the tokens and filtered paragraph.

Code :

```
!pip install nltk

import nltk

nltk.download('stopwords')

nltk.download('punkt')

from nltk.corpus import stopwords

from nltk.tokenize import word_tokenize

f = open('demo.txt','r')

paragraph = f.read()

stop_words = set(stopwords.words('english'))

words_tokens = word_tokenize(paragraph)
```

```
filtered_paragraph = [w for w in words_tokens if not w in stop_words]

print(words_tokens)

print(filtered_paragraph)
```

Code Screenshot:

```
!pip install nltk
import nltk
nltk.download('stopwords')
nltk.download('punkt')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

f = open('demo.txt', 'r')
paragraph = f.read()

stop_words = set(stopwords.words('english'))

words_tokens = word_tokenize(paragraph)
filtered_paragraph = [w for w in words_tokens if not w in stop_words]

#alternate way to filter paragraph
# for w in words_tokens:
#     if w not in stop_words:
#         filtered_paragraph.append(w)
print(words_tokens)
print(filtered_paragraph)
```

Output Screenshots :

```
Requirement already satisfied: nltk in /usr/local/lib/python3.6/dist-packages (3.2.5)
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from nltk) (1.15.0)
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
['Hello', '!', 'Welcome', 'to', 'demofile.txt', 'This', 'file', 'is', 'for', 'testing', 'purposes', '.', 'Good',
'Hello', '!', 'Welcome', 'demofile.txt', 'This', 'file', 'testing', 'purposes', '.', 'Good', 'Luck', '!']
```

Question 2

Problem Statement :

Use scrapy to crawl any one of the E-commerce websites of your choice and perform the same. The following information needs to be extracted from the page: (Choose any one product: e.g. laptop, smartphone ... etc.)

- a)Product name
- b)Product price
- c)Product discount
- d)Product image

Procedure :

Tools Used: Scrapy packages, python, ide to run the program, terminal/cmd (scrappy shell) for execution of various commands

Steps for Implementation:

1. Install the scrapy packages (pip install scrapy)
2. Create a new scrapy project (scrapy startproject [projectname])
3. Generate a spider to crawl through the web pages
4. Configure/edit the settings.py and the spider files as per requirement.
5. Run the program using ide and cmd to obtain results.

Code :

items.py

```
import scrapy

class ProductsItem(scrapy.Item):

    # define the fields for your item here like:

    product_name = scrapy.Field()

    product_price = scrapy.Field()

    product_discount = scrapy.Field()

    product_image = scrapy.Field()
```

flipkart_spider.py

```
import scrapy

from ..items import ProductsItem

class pro(scrapy.Spider):

    name = "pro"

    start_urls = ["https://www.flipkart.com/samsung-galaxy-m21-midnight-blue-128-
gb/p/itm0cec19c31b3cb?pid=MOBFSF85ZMVH3ZMG&lid=LSTMObFSF85ZMVH3ZMGHVRMDR&marketp
lace=FLIPKART&fm=neo%2Fmerchandising&iid=M_ab883bcf-9673-4bf2-adc9-
fccdcf6198a4_1_1BUWY8OBA8L9_MC.MOBFSF85ZMVH3ZMG&ppt=clp&ppn=samsung-mobile-
store&ssid=0lzd5i7qgw0000001596104962790&otracker=clp_pmu_v2_Latest%2BSamsung%2Bmobiles
%2B_1_1.productCard.PMU_V2_Latest%2BSamsung%2Bmobiles%2B_samsung-mobile-
store_MOBFSF85ZMVH3ZMG_neo%2Fmerchandising_0&otracker1=clp_pmu_v2_PINNED_neo%2Fmerc
handising_Latest%2BSamsung%2Bmobiles%2B_LIST_productCard_cc_1_NA_view-
all&cid=MOBFSF85ZMVH3ZMG"]

    def parse(self, response):

        product_name=response.css("span.B_NuCl::text").extract()

        product_price=response.css("div._30jeq3._16Jk6d::text").extract()

        product_discount=response.css("div._3Ay6Sb._31Dcoz").css("span::text").extract()

        product_image=response.css("div.q6DCIP::attr(style)").extract()

        items=ProductsItem()

        items["product_name"]=product_name

        items["product_price"]=product_price

        items["product_discount"]=product_discount

        items["product_image"]=product_image

        yield items
```

Code Screenshot:

items.py

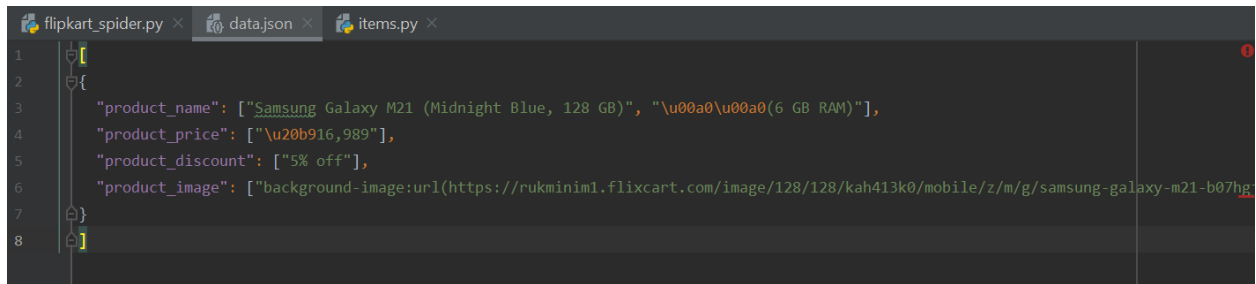
```
flipkart_spider.py × datajson × items.py ×
1
2 import scrapy
3
4 class ProductsItem(scrapy.Item):
5     # define the fields for your item here like:
6     product_name = scrapy.Field()
7     product_price = scrapy.Field()
8     product_discount = scrapy.Field()
9     product_image = scrapy.Field()
```

flipkart_spider.py

```
flipkart_spider.py × datajson × items.py ×
1
2 import scrapy
3 from ..items import ProductsItem
4 class pro(scrapy.Spider):
5     name = "pro"
6     start_urls = ["https://www.flipkart.com/samsung-galaxy-m21-midnight-blue-128-gb/p/itm0cec19c31b3cb?pid=MOBESF85ZMVH3ZMG&lid=LSTMOR"]
7     def parse(self, response):
8         product_name=response.css("span.B_NuCI::text").extract()
9         product_price=response.css("div._30jeq3._16Jk6d::text").extract()
10        product_discount=response.css("div._3Ay6Sb._31Dcoz").css("span::text").extract()
11        product_image=response.css("div.q6DC1P::attr(style)").extract()
12        items=ProductsItem()
13        items["product_name"]=product_name
14        items["product_price"]=product_price
15        items["product_discount"]=product_discount
16        items["product_image"]=product_image
17        yield items
```

Output Screenshots :

data.json



```
1 [{"product_name": ["Samsung Galaxy M21 (Midnight Blue, 128 GB)", "\u00a0\u00a0(6 GB RAM)"],
2   "product_price": ["\u20b916,989"],
3   "product_discount": ["5% off"],
4   "product_image": ["background-image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-b07hgj55il-original-imafsfewggf3dqwc.jpeg?q=70)", "background-
5   image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-b07hgj55il-original-imafsfewptm9vmsb.jpeg?q=70)", "background-
6   image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-b07hgj55il-original-imafsfewxhmhptvw.jpeg?q=70)", "background-
7   image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-b07hgj55il-original-imafsfewyghchvfy.jpeg?q=70)", "background-
8   image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-b07hgj55il-original-imafsfgtjwa5vmgf.jpeg?q=70)", "background-
9   image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/a/w/samsung-galaxy-m21-b07hgj559-original-imafsfgtgxhpszpfr.jpeg?q=70)", "background-
10  image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-b07hgj55il-original-imafsfgtx2gty9tu.jpeg?q=70)"]}]
```

data.json:

```
[
{
  "product_name": ["Samsung Galaxy M21 (Midnight Blue, 128 GB)", "\u00a0\u00a0(6 GB RAM)"],
  "product_price": ["\u20b916,989"],
  "product_discount": ["5% off"],
  "product_image": ["background-
image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-
b07hgj55il-original-imafsfewggf3dqwc.jpeg?q=70)", "background-
image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-
b07hgj55il-original-imafsfewptm9vmsb.jpeg?q=70)", "background-
image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-
b07hgj55il-original-imafsfewxhmhptvw.jpeg?q=70)", "background-
image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-
b07hgj55il-original-imafsfewyghchvfy.jpeg?q=70)", "background-
image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-
b07hgj55il-original-imafsfgtjwa5vmgf.jpeg?q=70)", "background-
image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/a/w/samsung-galaxy-m21-
b07hgj559-original-imafsfgtgxhpszpfr.jpeg?q=70)", "background-
image:url(https://rukminim1.flixcart.com/image/128/128/kah413k0/mobile/z/m/g/samsung-galaxy-m21-
b07hgj55il-original-imafsfgtx2gty9tu.jpeg?q=70)"],
}
]
```