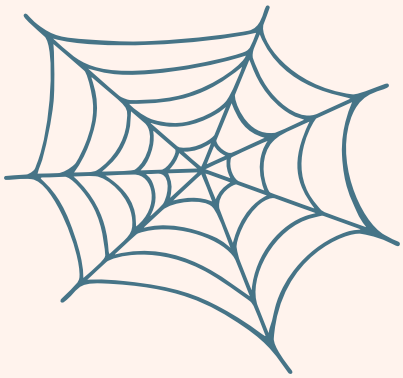


NOVEMBER 2023

EXERCISE 2

M.Amin HosseinNiya

Presented to:
Dr. Teymourpour

دانشگاه تربیت مدرس
Tarbiat Modares University

اسکرپ‌کردن لیست پایان‌نامه‌ها و تگ‌های هر استاد

با استفاده از scrapy اسپایدری ساختم که روی صفحات مربوط به پایان‌نامه‌های دکتر **امیر البدوی** بخزد و اطلاعات مورد نیاز ما را (شامل عنوان پایان‌نامه‌ها و تگ‌های هر پایان‌نامه) استخراج کند.

```
import scrapy
import requests
import json

from ..items import ex2Item

class GanjscrapspiderSpider(scrapy.Spider):
  name = "spider"
  allowed_domains = ["ganj.irandoc.ac.ir"]
                start_urls = ['https://ganj.irandoc.ac.ir/api/v1/search/main?
basicscope=1&keywords=%D8%A7%D9%85%DB%8C%D8%B1+%D8%A7%D9%84
%D8%A8%D8%AF%D9%88%DB%8C&page=1']

                url_format = 'https://ganj.irandoc.ac.ir/api/v1/search/main?
basicscope=1&keywords=%D8%A7%D9%85%DB%8C%D8%B1+%D8%A7%D9%84
%D8%A8%D8%AF%D9%88%DB%8C&page={page_number}'
  tags_url_format = "https://ganj.irandoc.ac.ir/api/v1/articles/{uuid}/show_tags"

  def parse(self, response):
    if "page=" not in response.url:
      page = 1
    else:
      page = int(response.url.split("page=")[-1])

    items = ex2Item()
```

اسکرپ‌کردن لیست پایان‌نامه‌ها و تگ‌های هر استاد

```
raw = json.loads(response.body)
data = raw['results']
number_of_pages = raw['total_pages']

for i in data:
items['title'] = i['title']
if i['keywords_status']:
doc_id = i['uuid']
tags_url = self.tags_url_format.format(uuid = doc_id)
r = requests.get(tags_url)
if r.status_code == 200:
tags = json.loads(r.content)
tag = []
for t in tags['tags']:
tag.append(t['title_fa'])
items['tags'] = tag

yield items

if page <= number_of_pages:
  yield scrapy.Request(self.url_format.format(page_number = page + 1), callback=self.parse)
```

اسکرپ‌کردن لیست پایان‌نامه‌ها و تگ‌های هر استاد

```
raw = json.loads(response.body)
data = raw['results']
number_of_pages = raw['total_pages']

for i in data:
items['title'] = i['title']
if i['keywords_status']:
doc_id = i['uuid']
tags_url = self.tags_url_format.format(uuid = doc_id)
r = requests.get(tags_url)
if r.status_code == 200:
tags = json.loads(r.content)
tag = []
for t in tags['tags']:
tag.append(t['title_fa'])
items['tags'] = tag

yield items

if page <= number_of_pages:
  yield scrapy.Request(self.url_format.format(page_number = page + 1), callback=self.parse)
```

بدی‌ترتیب خروجی در یک فایل json در اختیارم قرار گرفت.

اسکرپ‌کردن لیست پایان‌نامه‌ها و تگ‌های هر استاد

محتویات فایل جیسون را استخراج کردم:

```
## To read the json file and it's context:
import json

with open("E:\Master 01 Semester\Complex Networks\ex2-
HosseinNiya\Scraping\ex2\output3.json", "r", encoding='utf-8') as file:
    data = json.load(file)
file.close()
data_copy = data.copy
for item in data:
    if "tags" in item.keys():
        item["tags"] = [tag.strip("12345()") for tag in item["tags"]]
all_tags = list()

for item in data:
    if "tags" in item.keys():
        [all_tags.append(tag) for tag in item["tags"]]
all_titles = list()

for item in data:
        all_titles.append(item["title"])

len(all_titles)
all_tags = [element.strip("12345()") for element in all_tags]
len(all_tags)
unique_tags = list(set(all_tags))
len(unique_tags)
```

و همچنین نیم‌فاصله‌ها را با فاصله جایگزین کردم:

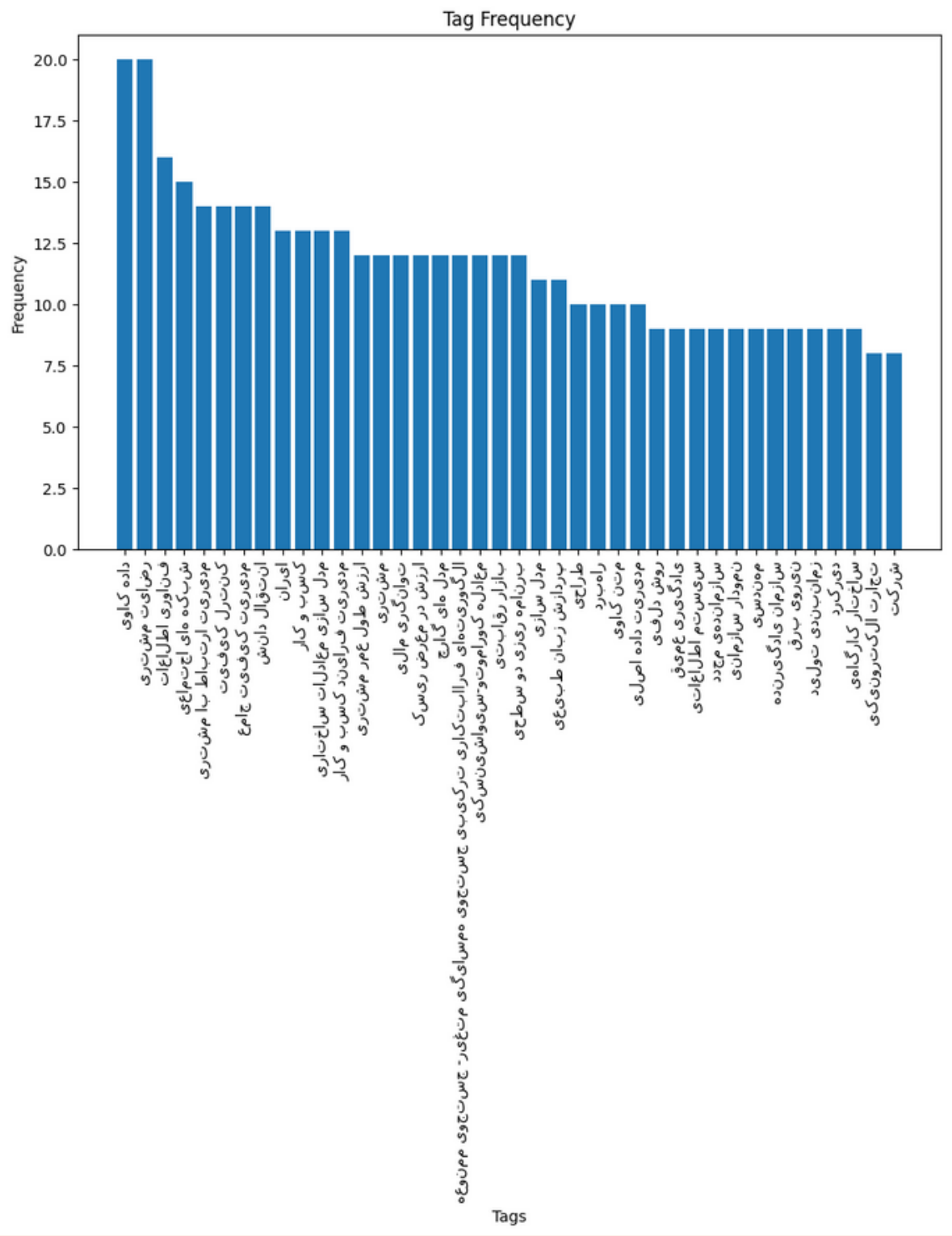
```
unique_tags = [i.replace("\u200c", " ") for i in unique_tags]
```

رسم هیستوگرام تگ‌های دکتر البدوی

```
import matplotlib.pyplot as plt
from collections import Counter
# from bidi.algorithm import get_display
# from persian_matplotlib import PersianMatplotlib

tags = all_tags
tag_counts = Counter(tags)

unique_tags = list(tag_counts.keys())
tag_frequencies = list(tag_counts.values())

sorted_indices = sorted(range(len(tag_frequencies)), key=lambda k:
tag_frequencies[k], reverse=True)
unique_tags = [unique_tags[i] for i in sorted_indices]
tag_frequencies = [tag_frequencies[i] for i in sorted_indices]

persian_xticks = [tag[::-1] for tag in unique_tags]

plt.figure(figsize=(10, 6))
plt.bar(unique_tags[:40], tag_frequencies[:40])
# with PersianMatplotlib():
  # plt.bar(range(len(unique_tags)), tag_frequencies)
plt.xlabel('Tags')
plt.ylabel('Frequency')
plt.title('Tag Frequency')
# plt.xticks(rotation=90)
plt.xticks(range(len(unique_tags[:40])), persian_xticks[:40], rotation=90)

plt.show()
```

رسم هیستوگرام تگ‌های دکتر البدوی

به دلیل تعداد بالای تگ‌ها و تراکم بیش‌اندازه‌ی نمودار، تنها چهل مورد پرتکرارتر را در نمودار آوردم:



رسم هیستوگرام تگ‌های دکتر البدوی

همانطور که مشاهده می‌شود، موضوعات داده‌کاوی، رضایت مشتری و فناوری اطلاعات، سه موضوع مورد علاقه‌ی دکتر البدوی هستند.