COURSE PROJECTS (MID-TERM)

NAME: SHAHAB HASHMI

ROLL NO.: AP18110010240

SECTION: CSE-D

INTRODUCTION

This assignment was made as part of the Information Retrieval Course. Our aim was to design a simple web crawler to get the latest headlines from the Entertainment section of the Times of India website, store the content of each section in a text file with first 3 letters of headline as its filename.

Methodology

- **Get the initial URL:**

```
f = requests.get(url)
url = 'https://timesofindia.indiatimes.com/'
```

- **Fetch the HTML content of the page, parse it to get the URLs of the entertainment section.**

```
data = s.find('div',{'class':"_1A86C"}).find_all('a',{'class':'linktype1'})
```

- **Put uri response in a variable:**

```
data = s.find('div',{'class':"_1A86C"}).find_all('a',{'class':'linktype1'})
```

- **Loop through the queue, get the data from each href using the above steps:**

```
for i in data:
req = requests.get(i['href'])
s = BeautifulSoup(req.content, 'lxml')
newData = s.find('div', {'class':'article_content clearfix'}).find_all('div',
{'class': 'Normal'})
```

- **Store first 3 words of the headline in a variable:**

```
fName = ''.join(i.text.split()[:3])
```

- **Retrieve the data from each href and store in a text file with its name from previous step:**

```
for j in newData:
print(j.text, end="\n")
with open('%s.txt'%(fName), 'w') as f:
f.write(j.text)
```

Tools and Software

Conclusion

- Content from each href was retrieved and stored in a text with its name as the headline successfully.
- The size of the data was 14.

References:

- https://www.topcoder.com/thrive/articles/web-crawler-in-python
- https://timesofindia.indiatimes.com/defaultinterstitial.cms
- https://scrapy.org/
- https://www.datacamp.com/community/tutorials/making-web-crawlers-scrapy-python