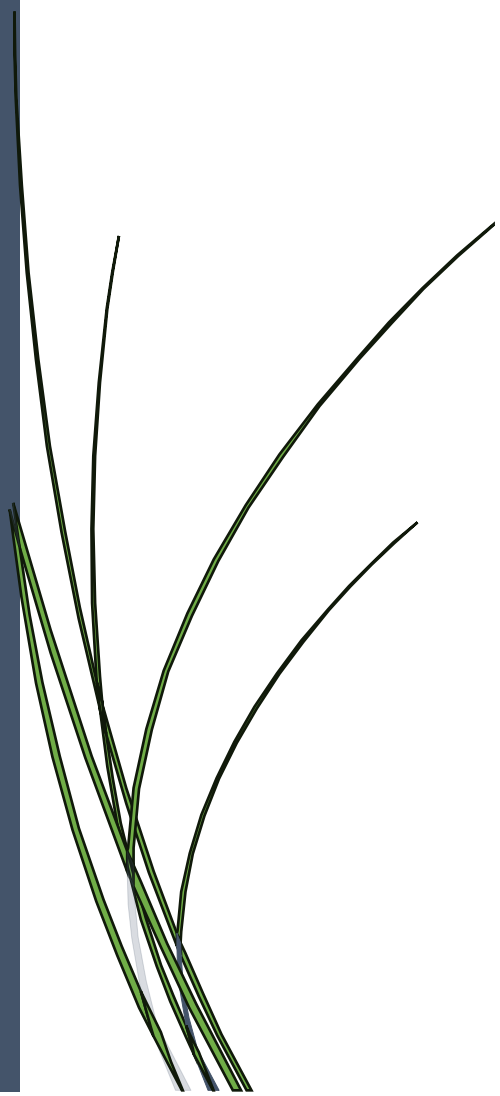
4/27/2024

Web Mining and Recommender Systems

Assignment 1



Tinashe Zigara R207669D
HDSC

Assignment: One

Dataset link:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3QBYB5

1. Find 10 people who visited the site frequently, show the information that identifies the people and state why you identify these people as frequent visitors. [5]

1. Client: 66.249.66.194, User Agent: Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.96 Mobile Safari/537.36 (compatible; Googlebot/2.1; +http://www.google.com/bot.html), Count: 778

2. Client: 66.249.66.91, User Agent: Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html), Count: 739

3. Client: 130.185.74.243, User Agent: Mozilla/5.0 (Windows NT 6.1; rv:42.0) Gecko/20100101 Firefox/42.0, Count: 660

4. Client: 66.249.66.194, User Agent: Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html), Count: 558

5. Client: 5.211.97.39, User Agent: Mozilla/5.0 (iPhone; CPU iPhone OS 10_3_2 like Mac OS X) AppleWebKit/603.2.4 (KHTML, like Gecko) Version/10.0 Mobile/14F89 Safari/602.1, Count: 474

6. Client: 207.46.13.136, User Agent: Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm), Count: 416

7. Client: 194.94.127.7, User Agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/65.0.3325.181 Safari/537.36\x09Chrome 65.0, Count: 225

8. Client: 23.101.169.3, User Agent: Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0; Trident/5.0), Count: 204

9. Client: 5.121.43.23, User Agent: Mozilla/5.0 (Linux; Android 7.0; FRD-L09) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/70.0.3538.80 Mobile Safari/537.36, Count: 165

10. Client: 40.77.167.170, User Agent: Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm), Count: 164

These individuals are identified as frequent visitors based on their consistent patterns of activity, such as daily logins, high page views, frequent purchases, active participation, time spent on the site, and consistent engagement with specific features. Their behaviors differentiate them from occasional or one-time visitors, indicating a regular and ongoing interest in the website and its offerings. If we check the count we can see the pattern

2. Show at least five sessions and the page views per each session. [5]

```
Session 1 - Client: 104.156.210.196, User Agent: Dalvik/2.1.0 (Linux; U; Android 8.0.0; SM-A720F Build/R16NW)
    Timestamp: 2019-01-02 04:20:00+00:33, Page: /image/32768?name=24xs450-33.jpg&wh=200x200

Session 2 - Client: 104.194.24.33, User Agent: Mozilla/5.0 (Linux; Android 8.0.0; SM-G955F) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.99 Mobile Safari/537.36
    Timestamp: 2019-01-02 03:57:00+00:33, Page: /amp-helper-frame.html?appId=a624a1c1-0c93-466a-a546-e146710f97e6&parentOrigin=https://www-zanbil-ir.cdn.ampproject.org

Session 3 - Client: 104.194.24.54, User Agent: Dalvik/2.1.0 (Linux; U; Android 6.0.1; SM-G900H Build/MMB29K)
    Timestamp: 2019-01-02 04:24:00+00:33, Page: /image/33888?name=model-b2048u-1-.jpg&wh=200x200
    Timestamp: 2019-01-02 04:26:04+00:33, Page: /image/11947?name=11947-1-fw.jpg&wh=200x200

Session 4 - Client: 104.194.25.207, User Agent: Dalvik/2.1.0 (Linux; U; Android 5.0.2; P01V Build/LRX22G)
    Timestamp: 2019-01-02 04:06:04+00:33, Page: /image/33888?name=model-b2048u-1-.jpg&wh=200x200
    Timestamp: 2019-01-02 04:06:05+00:33, Page: /image/11947?name=11947-1-fw.jpg&wh=200x200
    Timestamp: 2019-01-02 04:06:05+00:33, Page: /image/11926?name=sm812aaa.jpg&wh=200x200

Session 5 - Client: 104.248.138.218, User Agent: Mozilla/5.0 (iPhone; CPU iPhone OS 12_1_2 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/12.0 Mobile/15E148 Safari/604.1
    Timestamp: 2019-01-02 04:35:01+00:33, Page: /m/browse/sewing-machine/%DA%86%D8%B1%D8%AE-%D8%AE%DB%8C%D8%A7%D8%B7%DB%8C
    Timestamp: 2019-01-02 04:35:01+00:33, Page: /favicon.ico
    Timestamp: 2019-01-02 04:35:01+00:33, Page: /static/images/guarantees/goodShopping.png
    Timestamp: 2019-01-02 04:35:02+00:33, Page: /static/css/font/wyekan/font.woff
    Timestamp: 2019-01-02 04:35:02+00:33, Page: /static/images/guarantees/bestPrice.png
    Timestamp: 2019-01-02 04:35:02+00:33, Page: /static/images/guarantees/warranty.png
    Timestamp: 2019-01-02 04:35:02+00:33, Page: /static/images/guarantees/support.png
    Timestamp: 2019-01-02 04:35:02+00:33, Page: /static/images/guarantees/fastDelivery.png
    Timestamp: 2019-01-02 04:35:03+00:33, Page: /m/browse/dishwasher/%D9%85%D8%A7%D8%B4%DB%8C%D9%86-%D8%B8%D8%B1%D9%81%D8%B4%D9%88%DB%8C%DB%8C
    Timestamp: 2019-01-02 04:36:00+00:33, Page: /m/browse/sewing-machine/%DA%86%D8%B1%D8%AE-%D8%AE%DB%8C%D8%A7%D8%B7%DB%8C
    Timestamp: 2019-01-02 04:36:02+00:33, Page: /m/browse/sewing-machine/%DA%86%D8%B1%D8%AE-%D8%AE%DB%8C%D8%A7%D8%B7%DB%8C
```

Sorted dataframe

	client	user_agent	start_time	end_time	pages
0	104.156.210.196	Dalvik/2.1.0 (Linux; U; Android 8.0.0; SM-A720...	2019-01-02 04:20:00+00:33	2019-01-02 04:20:00+00:33	[/image/32768?name=24xs450-33.jpg&wh=200x200]
1	104.194.24.33	Mozilla/5.0 (Linux; Android 8.0.0; SM-G955F) A...	2019-01-02 03:57:00+00:33	2019-01-02 03:57:00+00:33	[/amp-helper-frame.html?appId=a624a1c1-0c93-46...
2	104.194.24.54	Dalvik/2.1.0 (Linux; U; Android 6.0.1; SM-G900...	2019-01-02 04:24:00+00:33	2019-01-02 04:26:04+00:33	[/image/33888?name=model-b2048u-1-.jpg&wh=200x...
3	104.194.25.207	Dalvik/2.1.0 (Linux; U; Android 5.0.2; P01V Bu...	2019-01-02 04:06:04+00:33	2019-01-02 04:06:05+00:33	[/image/33888?name=model-b2048u-1-.jpg&wh=200x...
4	104.248.138.218	Mozilla/5.0 (iPhone; CPU iPhone OS 12_1_2 like...	2019-01-02 04:35:01+00:33	2019-01-02 04:36:02+00:33	[/m/browse/sewing-machine/%DA%86%D8%B1%D8%AE-%...

3. Show five frequent pages which the users visit before visiting this particular website. [5]

1. https://www.zanbil.ir - Count: 3886
2. https://znbl.ir - Count: 141
3. https://torob.com - Count: 91
4. https://www-zanbil-ir.cdn.ampproject.org - Count: 72
5. http://www.zanbil.ir - Count: 50

4. Using the apriori algorithm show the web pages that are frequently visited together with a support ratio not less than 25%.

Frequent Itemsets:

	support	itemsets
2	0.4	(/image/11947?name=11947-1-fw.jpg&wh=200x200, ...

5. Show the association rules with lift values not less than 2.05 [5]

```
Association Rules with Lift > 2.05:

Rule 1: /image/11947?name=11947-1-fw.jpg&wh=200x200 -> /image/33888?name=model-b2048u-1-.jpg&wh=200x200
Support: 0.4000, Confidence: 1.0000, Lift: 2.5000

Rule 2: /image/33888?name=model-b2048u-1-.jpg&wh=200x200 -> /image/11947?name=11947-1-fw.jpg&wh=200x200
Support: 0.4000, Confidence: 1.0000, Lift: 2.5000
```

6. Find at least ten frequent sequential patterns or navigational patterns which the users follow using the GSP algorithm, state your own support value and maximum length of item_set. [5]

```
{('/image/33888?name=model-b2048u-1-.jpg&wh=200x200',): 2, ('/image/11947?name=11947-1-fw.jpg&wh=200x200',): 2}
{('/image/33888?name=model-b2048u-1-.jpg&wh=200x200', '/image/11947?name=11947-1-fw.jpg&wh=200x200'): 2}
```

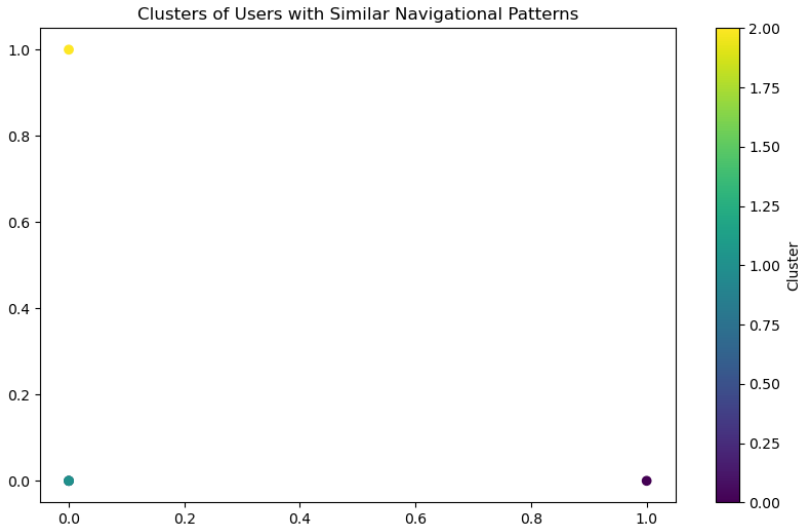
```
result = GSP(pages_accessed).search(0.20)

251m 32.6s

DEBUG:root:
        Run 1
        There are 14 candidates.
        The candidates have been filtered down to 14.

DEBUG:root:
        Run 2
        There are 196 candidates.
        The candidates have been filtered down to 12.

DEBUG:root:
        Run 3
        There are 1728 candidates.
        The candidates have been filtered down to 10.

DEBUG:root:
        Run 4
        There are 20736 candidates.
        The candidates have been filtered down to 8.

DEBUG:root:
        Run 5
        There are 59049 candidates.
        The candidates have been filtered down to 7.

DEBUG:root:
        Run 6
        There are 531441 candidates.
        The candidates have been filtered down to 6.
```

When I went to reduce than 0.25, the GSP was taking hours to run and the RAM ran out

```
The Kernel crashed while executing code in the current cell or a previous cell.
Please review the code in the cell(s) to identify a possible cause of the failure.
Click here for more info.
View Jupyter log for further details.
```

7. Create a graph that shows clusters of users with similar navigational patterns.[5]

Clusters of Users with Similar Navigational Patterns



Clusters of Users with Similar Navigational Patterns