



Name: Farhan Shikalgar	Course Name: SMA Lab
Class: BECO	Roll no: 19CO56

Experiment No: 3

Aim: Data Cleaning and Storage- Preprocess, filter and store social media data for business (Using Python, MongoDB, R, etc).

Theory:

Social Media Scraping of Myntra using Python:

1. Instagram:

```
2. # Instagram
3. from instagrampy import InstagramUser
4. user = InstagramUser("Myntra")
5. print(f"Username: {user.fullname}")
6. print(f"Biography: {user.biography}")
7. print(f"Verified User: {user.is_verified}")
8. print(f"Website: {user.website}")
9. print(f"Followers: {user.number_of_followers}")
10. print(f"Following: {user.number_of_followings}")
11. print(f"No. Of Posts: {user.number_of_posts}")
12. posts = user.posts
13. print(posts[0])
14. instaPosts = []
15. for i in range(10):
16.     post = {}
17.     post["Likes"] = posts[i].likes
18.     post["Comments"] = posts[i].comments
19.     post["post_source"] = posts[i].post_source
20.     post["post_url"] = posts[i].post_url
21.     post["time"] = posts[i].taken_at_timestamp
22.     instaPosts.append(post)
23. insta_df = pd.DataFrame(instaPosts)
24. insta_df.info()
```

```

25.print(insta_df.isna())
26.print(f"Latest posts:\n {insta_df}")

```

OUTPUT

```

python Exp3-SMA.py
Username: MYNTRA
Biography: If you never try, you'll never know ❤
👉 #MyMyntraLook get featured on Myntra Studio
Visit Myntra Studio for more fashion & style advice.
Verified User: True
Website: https://myntra.onelink.me/eWcy/gnozbjm9
Followers: 3168336
Following: 262
No. Of Posts: 10804
Post(Likes=283, comments=9, caption=None, is_video=True, timestamp=1675679094, location={'id': '109524955741121', 'has_public_page': True, 'name': 'India', 'slug': 'india'}, shortcode='CoUUCXvJzT6', post_url='https://www.instagram.com/p/CoUUCXvJzT6/', display_url='https://instagram.fbom56-1.fna.fbcdn.net/v/t51.2885-15/329543412_1342790036515421_8823910709516276393_n.jpg?stp=dst-jpg_e15&nc_ht=instagram.fbom56-1.fna.fbcdn.net&nc_cat=1&nc_ohc=sEZNfq2LUdkAX_mCWQ&edm=ABfd0MgBAAAA&ccb=7-5&oh=00_AfAurulbFXhitwlnp0VLfbppHaUgxzq70FfNuvlo9Up58w&oe=63E2D626&nc_sid=7bff83', video_url='https://instagram.fbom56-1.fna.fbcdn.net/v/t66.30100-16/54642810_578501280529533_5147923975634763362_n.mp4?nc_ht=instagram.fbom56-1.fna.fbcdn.net&nc_cat=109&nc_ohc=rKgHihplNFIAX_mxMkr&edm=ABfd0MgBAAAA&ccb=7-5&oh=00_AfBJTQ0v6StFY_MRMGFFz3XpgEmJhP9rYSz38cuKYINb-w&oe=63E274A7&nc_sid=7bff83', video_view_count=4455, post_source='https://instagram.fbom56-1.fna.fbcdn.net/v/t66.30100-16/54642810_578501280529533_5147923975634763362_n.mp4?nc_ht=instagram.fbom56-1.fna.fbcdn.net&nc_cat=109&nc_ohc=rKgHihplNFIAX_mxMkr&edm=ABfd0MgBAAAA&ccb=7-5&oh=00_AfBJTQ0v6StFY_MRMGFFz3XpgEmJhP9rYSz38cuKYINb-w&oe=63E274A7&nc_sid=7bff83', taken_at_timestamp=datetime.datetime(2023, 2, 6, 15, 54, 54))
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   Likes        10 non-null    int64
1   Comments     10 non-null    int64

```

```

4   time      10 non-null    datetime64[ns]
dtypes: datetime64[ns](1), int64(2), object(2)
memory usage: 384.0+ bytes
   Likes  Comments  post_source  post_url  time
0  False     False      False     False  False
1  False     False      False     False  False
2  False     False      False     False  False
3  False     False      False     False  False
4  False     False      False     False  False
5  False     False      False     False  False
6  False     False      False     False  False
7  False     False      False     False  False
8  False     False      False     False  False
9  False     False      False     False  False
Latest posts:
   Likes  Comments  ...  post_url  time
0    283      9    ...  https://www.instagram.com/p/CoUUCXvJzT6/  2023-02-06 15:54:54
1    790     15    ...  https://www.instagram.com/p/CoTuf19D-xu/  2023-02-06 10:27:51
2    327     10    ...  https://www.instagram.com/p/CoURNTXJ9zQ/  2023-02-06 15:29:32
3    505     30    ...  https://www.instagram.com/p/CoUam3nP22C/  2023-02-06 16:48:33
4    652      7    ...  https://www.instagram.com/p/CoUWZRxJCGx/  2023-02-06 16:20:11
5   2156     12    ...  https://www.instagram.com/p/CoUVMPxs5La/  2023-02-06 16:01:13
6    876      9    ...  https://www.instagram.com/p/CoSwXmrjct9/  2023-02-06 01:23:10
7    796     21    ...  https://www.instagram.com/p/CoSEvupvhTz/  2023-02-05 18:59:02
8   1320     55    ...  https://www.instagram.com/p/CoR3iSesDLg/  2023-02-05 17:03:36
9   1092     23    ...  https://www.instagram.com/p/CoRj7jrpNPI/  2023-02-05 14:12:17

[10 rows x 5 columns]

```

27. Facebook:

```
28. # Facebook
29. from facebook_scraper import get_posts
30. fbPosts = []
31. for post in get_posts('Myntra', pages=10):
32.     fbPosts.append(post)
33. print(fbPosts)
34. df = pd.DataFrame(fbPosts)
35. df.info()
36. df = df[['post_id', 'text', 'post_url', 'page_id', 'comments', 'likes',
            'time']]
37. print(df.isna())
38. print(print(f"Latest posts:\n {df}"))
```

OUTPUT

```
pwh in Python
[{'post_id': '10161169058823221', 'text': 'YAAASSS, we can't believe our eyes either😭😭😭\nStay tuned for something extraordinary dropping soon with our favourite trendsetter, @kiaraaliaadvani! 🤔\n\n#BeExtraordinaryEveryday #MyntraBeExtraordinaryEveryday #Myntra #KiaraAdvaniXMynttra', 'post_text': 'YAAASSS, we can't believe our eyes either😭😭😭\nStay tuned for something extraordinary dropping soon with our favourite trendsetter, @kiaraaliaadvani! 🤔\n\n#BeExtraordinaryEveryday #MyntraBeExtraordinaryEveryday #Myntra #KiaraAdvaniXMynttra', 'shared_text': '', 'original_text': None, 'time': datetime.datetime(2023, 2, 6, 16, 20, 21), 'timestamp': 1675680621, 'image': None, 'image_lowquality': 'https://scontent.fbom56-1.fna.fbcdn.net/v/t51.36329-10/329046477_605564214685874_8797136335172886712_n.jpg?stp=cp0_dst=jpg_e15_p320x320_q65&nc_cat=1&ccb=1-7&nc_sid=ccf8b3&nc_ohc=P1zGo84mtJgAX8TFifX&nc_ht=scontent.fbom56-1.fna&oh=00_AfASBb9u8inoEKqaqEYwVNGtWrKaMFf4LcuxvuzK8Su0zg&oe=63E6B6E2', 'images': [], 'images_description': [], 'images_lowquality': ['https://scontent.fbom56-1.fna.fbcdn.net/v/t51.36329-10/329046477_605564214685874_8797136335172886712_n.jpg?stp=cp0_dst=jpg_e15_p320x320_q65&nc_cat=1&ccb=1-7&nc_sid=ccf8b3&nc_ohc=P1zGo84mtJgAX8TFifX&nc_ht=scontent.fbom56-1.fna&oh=00_AfASBb9u8inoEKqaqEYwVNGtWrKaMFf4LcuxvuzK8Su0zg&oe=63E6B6E2'], 'images_lowquality_description': [None], 'video': 'https://video.fbom56-1.fna.fbcdn.net/v/t50.33967-2/329735113_212601647945025_5364783907066637839_n.mp4?_nc_cat=101&ccb=1-7&nc_sid=985c63&efg=eyJ2ZW5jb2RlX3RhZyI6IWhwZl9zZl9wcm9ncmVzc2l2ZS9J&nc_ohc=Hswv35rfVwAX92msim&nc_rml=0&nc_ht=video.fbom56-1.fna&oh=00_AfAt4TKSsZz7yVukImozpYfDnjH51MdiYzABmeBeDu_Dw&oe=63E196DC', 'video_duration_seconds': None, 'video_height': None, 'video_id': '1140172910007339', 'video_quality': None, 'video_size_MB': None, 'video_thumbnail': 'https://scontent.fbom56-1.fna.fbcdn.net/v/t51.36329-10/329046477_605564214685874_8797136335172886712_n.jpg?stp=cp0_dst=jpg_e15_p320x320_q65&nc_cat=1&ccb=1-7&nc_sid=ccf8b3&nc_ohc=P1zGo84mtJgAX8TFifX&nc_ht=scontent.fbom56-1.fna&oh=00_AfASBb9u8inoEKqaqEYwVNGtWrKaMFf4LcuxvuzK8Su0zg&oe=63E6B6E2', 'video_watches': None, 'video_width': None, 'likes': 26, 'comments': 6, 'shares': 0, 'post_url': 'https://facebook.com/myntra/posts/10161169058823221', 'link': None, 'links': [{'link': '/hashtag/beextraordinaryeveryday?_tn=%2As-R', 'text': '#BeExtraordinaryEveryday'}, {'link': '/hashtag/myntra?_tn=%2As-R', 'text': '#Myntra'}, {'link': '/hashtag/kiaraadvaniXmyntra?_tn=%2As-R', 'text': '#KiaraAdvaniXMynttra'}, {'link': '/story.php?story_fbid=pfbid02NcHHEFyygWDMRHhaUhcVvKZ54Ec6funFX5myA1AhMawWw9tJnUzSxre4dM9wHiyoGL&id=6466648220&stream_source=timeline&_tn=%2As%2As-R', 'text': ''}], 'user_id': '6466648220', 'username': 'Myntra', 'user_url': 'https://facebook.com/myntra/?_tn=C-R', 'is_live': False, 'factcheck': None, 'shared_post_id': None, 'shared_time': None, 'shared_user_id': None, 'shared_username': None, 'shared_post_url': None, 'available': True, 'comments_full': None, 'reactors': None, 'w3_fb_url': None, 'reactions': None, 'reaction_count': 26, 'with': None, 'page_id': '6466648220', 'sharers': None, 'image_id': None, 'image_ids': [], 'was_live': False}, {'post_id': '10161169030738221', 'text': 'YAAASSS, we can't believe our eyes either😭😭😭\nStay tuned for something extraordinary dr
```

```

31 False False False False False False False
32 False False False False False False False
33 False False False False False False False
34 False False False False False False False
35 False False False False False False False
Latest posts:
      post_id      text  ...  likes      time
0  10161169058823221  YAAASSS, we can't believe our eyes either😂😂😂\n...  ...  26.0  2023-02-06 16:20:21
1  10161169030738221  YAAASSS, we can't believe our eyes either😂😂😂\n...  ...  10.0  2023-02-06 15:55:05
2  10161168987678221  YAAASSS, we can't believe our eyes either😂😂😂\n...  ...   9.0  2023-02-06 15:29:42
3  10161168767013221  YAAASSS, we can't believe our eyes either😂😂😂\n...  ...  11.0  2023-02-06 12:31:15
4  10161168690158221  YAAASSS, we can't believe our eyes either😂😂😂\n...  ...  14.0  2023-02-06 11:15:04
5  10161167893718221  This Pink monochrome look is ❤️❤️\n\n👉 @simran ...  ...  29.0  2023-02-06 01:23:22
6  10161166649103221  Whether you're going on a date, or giving your ...  ...  82.0  2023-02-05 14:12:22
7  10161165440628221  Mark your calendar & set your reminder 'cause ...  ...  28.0  2023-02-04 23:38:35
8  10161162761178221  Jump into an unparalleled listening experience ...  ...  29.0  2023-02-04 19:46:55
9  10161164921723221  Shhhh...something extraordinary coming your way, ...  ...  70.0  2023-02-04 18:18:43
10 10161164580113221  Shhhh...something extraordinary coming your way, ...  ...  24.0  2023-02-04 13:36:57
11 10161164466943221  Shhhh...something extraordinary coming your way, ...  ...  20.0  2023-02-04 11:55:11
12 10161164422793221  Shhhh...something extraordinary coming your way, ...  ...  27.0  2023-02-04 11:07:25
13 10161163521578221  Denim game strong with pantaloons!\nCheck out ...  ...  34.0  2023-02-03 23:55:34
14 10161163439843221  A race before the actual race!\nThe Early Bird ...  ...  29.0  2023-02-03 23:09:04
15 10161163209568221  Keep swiping 🤳 ...  ... 121.0  2023-02-03 21:12:09
16 10161161257373221  Vibe check ❤️\n👉 @phoolfilledvish\n#Reels #Fash ...  ...  62.0  2023-02-03 00 :12:56
17 10161161191548221  ❤️\n👉 @thankgod_itsfashion\n#Reels #FashionReel ...  ...  19.0  2023-02-02 23 :41:28
18 10161161063803221  The clock is ticking! 🕒\nYou don't wanna miss ...  ...  45.0  2023-02-02 22:38:48
19 10161160787473221  Pastel magic!\nCheck out👉Myntra Studio 👉for fa ...  ...  60.0  2023-02-02 20:30:45
20 10161158857943221  Cute outfit inspo👉\nCheck out👉Myntra Studio 👉f ...  ...   NaN  2023-02-02 00:01:45
21 10161158772288221  Tea party ready!👉\nCheck out our stunning rang ...  ...  81.0  2023-02-01 23:23:33

```

39. Twitter:

```

40. # Twitter
41. import twint
42. c = twint.Config()
43. c.Lang = "en"
44. c.Username = "Myntra"
45. c.Pandas = True
46. # Run
47. twint.run.Search(c)
48. Tweets_df = twint.storage.panda.Tweets_df
49. Tweets_df.info()
50. print(Tweets_df.isna())
51. Tweets_df = Tweets_df[Tweets_df.language == "en"]
52. Tweets_df = Tweets_df[['id', 'date', 'nlikes', 'language', 'nreplies',
    'tweet', 'username']]
53. print(Tweets_df)

```

Output

```
1622578947249811456 2023-02-06 18:22:21 +0530 <mynttra> @srikanth_vittam That doesn't sound like a happy situation, Vittam
! I've prioritized this concern & our case manager will contact you at the earliest. -KB https://t.co/WsLuXFysSb
1622578534584836096 2023-02-06 18:20:43 +0530 <mynttra> @Mumalrajawat63 I get that you're expecting an update, and it was
never our objective to keep you waiting, Mumal! Upon checking, I see that an email was sent to you under reference number
'IN23020316521082206453' on 06-Feb-2023 at 11:46 AM. Kindly check (cont) https://t.co/CbJqJhBR2o
1622578068916404224 2023-02-06 18:18:52 +0530 <mynttra> @bhavika_arora_ Bhavika! Upon checking, I see that one of our case
managers tried calling you today, 12:40 PM, but unfortunately, could not connect. A notification was sent to you via the
Myntra App with the reference number IN23020515312830651385. You (cont) https://t.co/TjL3hTmwJq
1622578026235195394 2023-02-06 18:18:42 +0530 <mynttra> @MuskanRahul6 Hi Upendra! I have replied to you here https://t.co
/qRsSJRKsim . Please check. - ST
1622577891925200896 2023-02-06 18:18:10 +0530 <mynttra> @MuskanRahul6 To hear this concerns us, Upendra! I regret the hass
le caused. Rest assured, we're closely following up with the logistics partners to have the product picked up from you so
on. One of our case managers will connect with you to share an update as early as possible. - ST
1622576957711056896 2023-02-06 18:14:27 +0530 <mynttra> @singhsal Hi Sakshi! As we can see that one of our case managers c
ontacted you on 01 Feb 2023, 01:15 PM and ask for images of the product with bit more information. Also, send an in-app w
ith reference id IN23013111020246736311. Kindly have a check and revert for further assistance. -AB
1622576901993955328 2023-02-06 18:14:14 +0530 <mynttra> @srikanth_vittam It was never our intention to cause any disruptio
n to our customers, Srikanth! Please be assured that I've highlighted your concern, and we are looking into it with the u
tmost priority. One of our case managers will connect with you at the earliest to provide an update. -MK
1622575586232377344 2023-02-06 18:09:00 +0530 <mynttra> @Sumitzrd It's never our intention to keep you waiting for the res
olution, Sumit! Please be assured that we are working towards a resolution on priority, and one of our case managers will
connect within the shortest time possible to address your concern with better precision. -MH
1622574933628051458 2023-02-06 18:06:24 +0530 <mynttra> @MuskanRahul6 Upendra! Rest assured, I've prioritized this concern
& one of our case managers will reach out to you at the earliest-- PK
1622574892456742912 2023-02-06 18:06:15 +0530 <mynttra> @mahima_nayyar Hi Mahima! As we can see that one of our case manag
ers contacted you on 06 Feb 2023, 02:19 PM and ask for necessary information. Also, send an in-app with reference id IN23
020500492948796241. Kindly have a check and revert. Awaiting your response. -AB
1622574431993479172 2023-02-06 18:04:25 +0530 <mynttra> @Rashi42691209 I hear you, Rashi! Upon checking, I see our case ma
nager has sent an In-App communication under the reference number IN23020312065441905756. You can track and view the same
```

Data columns (total 38 columns):			
#	Column	Non-Null Count	Dtype
0	id	20 non-null	object
1	conversation_id	20 non-null	object
2	created_at	20 non-null	float64
3	date	20 non-null	object
4	timezone	20 non-null	object
5	place	20 non-null	object
6	tweet	20 non-null	object
7	language	20 non-null	object
8	hashtags	20 non-null	object
9	cashtags	20 non-null	object
10	user_id	20 non-null	int64
11	user_id_str	20 non-null	object
12	username	20 non-null	object
13	name	20 non-null	object
14	day	20 non-null	int64
15	hour	20 non-null	object
16	link	20 non-null	object
17	urls	20 non-null	object
18	photos	20 non-null	object
19	video	20 non-null	int64
20	thumbnail	20 non-null	object
21	retweet	20 non-null	bool
22	nlikes	20 non-null	int64
23	nreplies	20 non-null	int64
24	nretweets	20 non-null	int64
25	quote_url	20 non-null	object


```
pwsh in Python
18 False False False False False ... False False False False False
19 False False False False False ... False False False False False

[20 rows x 38 columns]
   id      date      ...      tweet      username
0  1622581307883794433  2023-02-06 18:31:44 ... @TheBitterTruth To hear this concerns us. All... myntra
1  1622580814927269889  2023-02-06 18:29:47 ... @bhavika_arora_ I value the importance of your... myntra
2  1622580744190296064  2023-02-06 18:29:30 ... @amitmehta251091 It is not our intention to ke... myntra
3  1622580354019393536  2023-02-06 18:27:57 ... @Rishika07302711 Helping you out is our absolu... myntra
4  1622578947249811456  2023-02-06 18:22:21 ... @srikanth_vittam That doesn't sound like a hap... myntra
5  1622578534584836096  2023-02-06 18:20:43 ... @Mumalrajawat63 I get that you're expecting an... myntra
6  1622578068916404224  2023-02-06 18:18:52 ... @bhavika_arora_ Bhavika! Upon checking, I see ... myntra
7  1622578026235195394  2023-02-06 18:18:42 ... @MuskanRahul6 Hi Upendra! I have replied to yo... myntra
8  1622577891925200896  2023-02-06 18:18:10 ... @MuskanRahul6 To hear this concerns us, Upendr... myntra
9  1622576957711056896  2023-02-06 18:14:27 ... @singhsal Hi Sakshi! As we can see that one of... myntra
10 1622576901993955328  2023-02-06 18:14:14 ... @srikanth_vittam It was never our intention to ... myntra
11 1622575586232377344  2023-02-06 18:09:00 ... @Sumitzrd It's never our intention to keep you... myntra
12 1622574933628051458  2023-02-06 18:06:24 ... @MuskanRahul6 Upendra! Rest assured, I've prio... myntra
13 1622574892456742912  2023-02-06 18:06:15 ... @mahima_nayyar Hi Mahima! As we can see that o... myntra
14 1622574431993479172  2023-02-06 18:04:25 ... @Rashi42691209 I hear you, Rashi! Upon checkin... myntra
15 1622573953578590208  2023-02-06 18:02:31 ... @ChandrikaPathai I get that you are looking fo... myntra
16 1622573867817639936  2023-02-06 18:02:10 ... @angarikam To hear this concerns us, Angarika!... myntra
17 1622573569149661185  2023-02-06 18:00:59 ... @SujataMathur It was never our intention to pr... myntra
18 1622573147177517057  2023-02-06 17:59:18 ... @priyanka07551 I get that you are looking for ... myntra
19 1622572593562947591  2023-02-06 17:57:06 ... @angarikam Customer satisfaction is what we st... myntra

[20 rows x 7 columns]
C:\Desktop\Coding\Python 3.8.6 5.863s 18:32:05
```

Conclusion:

We have successfully cleaned Myntra's social media data from websites like Instagram, Facebook, and Twitter using various data preprocessing techniques.