

Kaggle is not the Norm:

Using snsrape to collect data

Evan Simpson

p.evansimpson@gmail.com

[EducatorsRLearners](#)

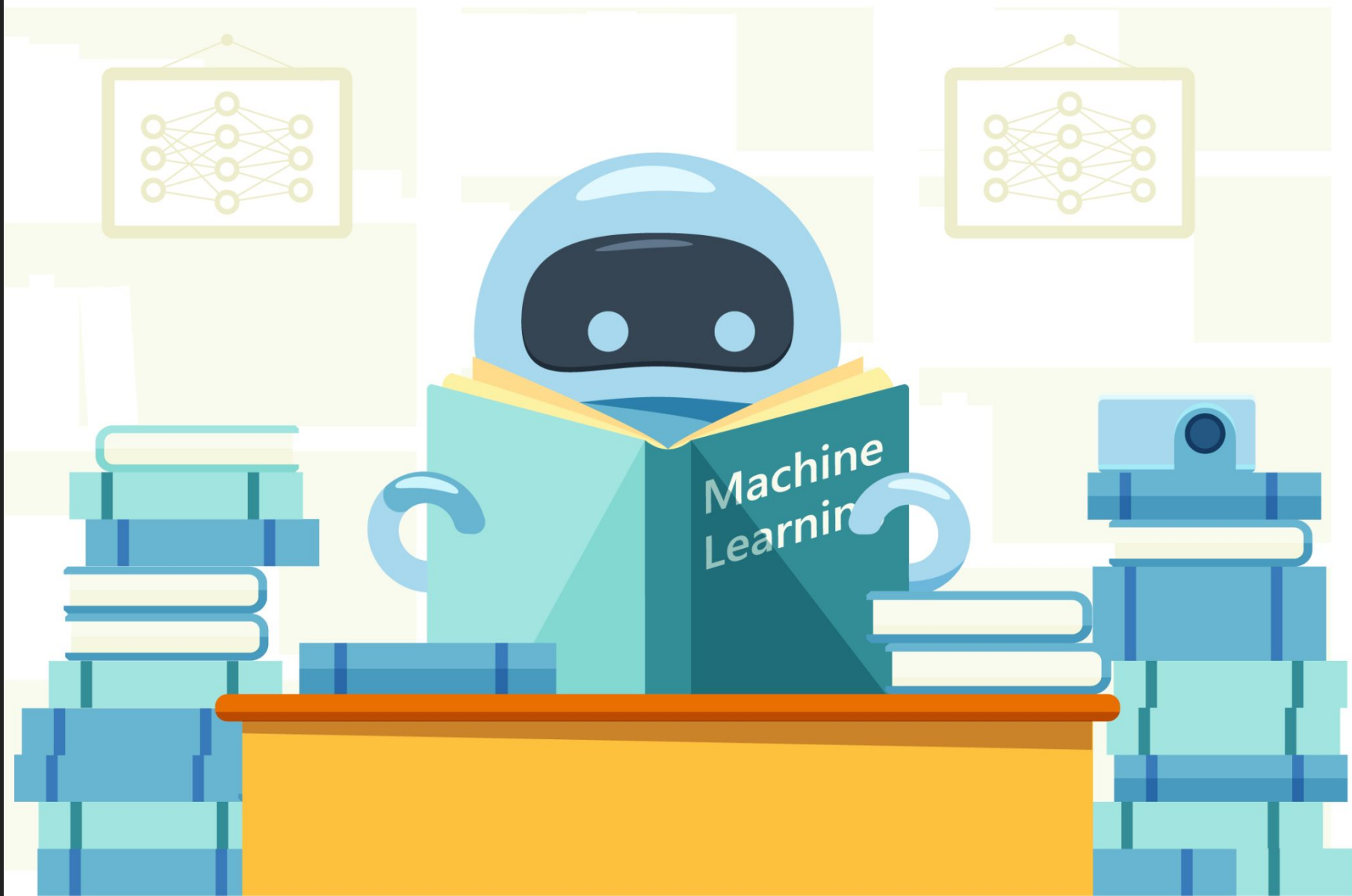
Agenda

- Introduction
- snsrape
- practice
- Future Directions

A stylized illustration of a person from the chest up, wearing a dark blue suit jacket, a white dress shirt, and a red necktie. The person's head is not visible. On the right side of the chest, there is a white rectangular name tag with a yellow clip at the top. The background is a solid yellow color.

HELLO,
My Name Is..





Put in Order

- A) Prepare the Data
- B) Frame the problem
- C) Inspect the Feature Importance
- D) Conduct Exploratory Data Analysis (EDA)
- E) Install and Load the Libraries
- F) Build and Evaluate the Model
- G) Collect/Load the Data

Sample Workflow

B) Frame the problem

E) Install and Load the Libraries

G) Collect/Load the Data

A) Prepare the Data

D) Conduct Exploratory Data Analysis (EDA)

F) Build and Evaluate the Model

C) Inspect the Feature Importance

Today's Goals

B) Frame the problem

E) Install and Load the Libraries

G) Collect/Load the Data

A) Prepare the Data

D) Conduct Exploratory Data Analysis (EDA)

The question isn't
'*We have this tech.
What can we use if
for?*' The question is
'*We have this
problem. What tech
can we use to solve
it?*'

- Alan Maley





snsrape

A scraper for Social Networking Services



twitter-hashtag

twitter-list-posts

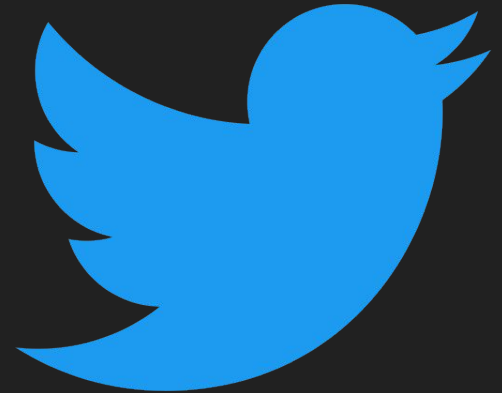
twitter-profile

twitter-search

twitter-trends

twitter-tweet

twitter-user





Get All Tweets by user @elonmusk

```
snsrape --jsonl twitter-user elonmusk > elonmusk.json
```

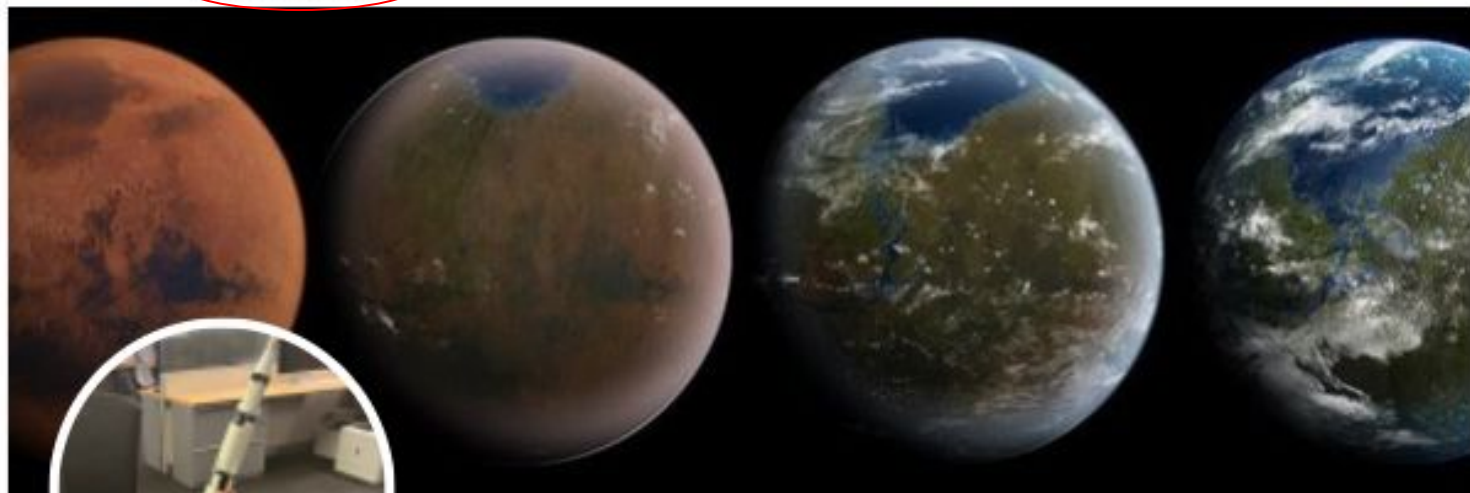
1 2 3

1. `--jsonl`: Return data in json format
2. `twitter-user`: What to search for
3. `>`: save the output to this file



Elon Musk

17.7K Tweets



Following

Elon Musk

@elonmusk

Joined June 2009

114 Following **92M** Followers

Get the Last 100 Tweets by user @elonmusk

```
snsrape --jsonl --max-results 100 twitter-user elonmusk > elonmusk.json
```

1 2 3 4

1. --jsonl: Return data in json format

2. --max-results: Return the last N number of tweets

3. twitter-user: What to search for

4. >: Save the output to this file



Get All Tweets with the hashtag MAGA

snsrape --jsonl twitter-hashtag MAGA > MAGA.json

1 2 3

1. --jsonl: Return data in json format
2. twitter-hashtag: What to search for
3. >: save the output to this file



Get All Tweets with the hashtag MAGA from January 6th, 2020

```
snsrape --jsonl --progress --since 2020-01-06 twitter-hashtag \  
               1                2  
"MAGA --until:2020-01-07" > MAGA_jan_06.json  
               3
```

1. --progress: Displays every 100 tweets scraped
2. --since : Start date and time
3. --until : End date and time

Get All Tweets from @narendramodi which contain “money”

```
snsrape --jsonl twitter-search “money from:narendramodi” >modi.json
```

1 2 3 4

1. `--jsonl`: Return data in json format
2. `twitter-search`: Type of search
3. `money`: Term to search for
4. `from:narendramodi`: User who sent the tweet

“How do I know what is possible?”

× **Advanced search** Search

Words

All of these words

Example: what's happening · contains both “what's” and “happening”

This exact phrase

Example: happy hour · contains the exact phrase “happy hour”

Any of these words

Example: cats dogs · contains either “cats” or “dogs” (or both)

None of these words

Example: cats dogs · does not contain “cats” and does not contain “dogs”

These hashtags

What is returned?

```
['_type', 'url', 'date', 'content',  
'renderedContent', 'id', 'user',  
'replyCount', 'retweetCount',  
'likeCount', 'quoteCount', 'conversationId',  
'lang', 'source', 'sourceUrl', 'sourceLabel',  
'outlinks', 'tcooutlinks', 'media',  
'retweetedTweet',  
'quotedTweet', 'inReplyToTweetId',  
'inReplyToUser', 'mentionedUsers',  
'coordinates', 'place', 'hashtags',  
'cashtags']
```


date: date and time (GMT) tweet was sent

content: Content of the tweet

replyCount: number of replies











retweetCount: number of retweets

likeCount: number of likes

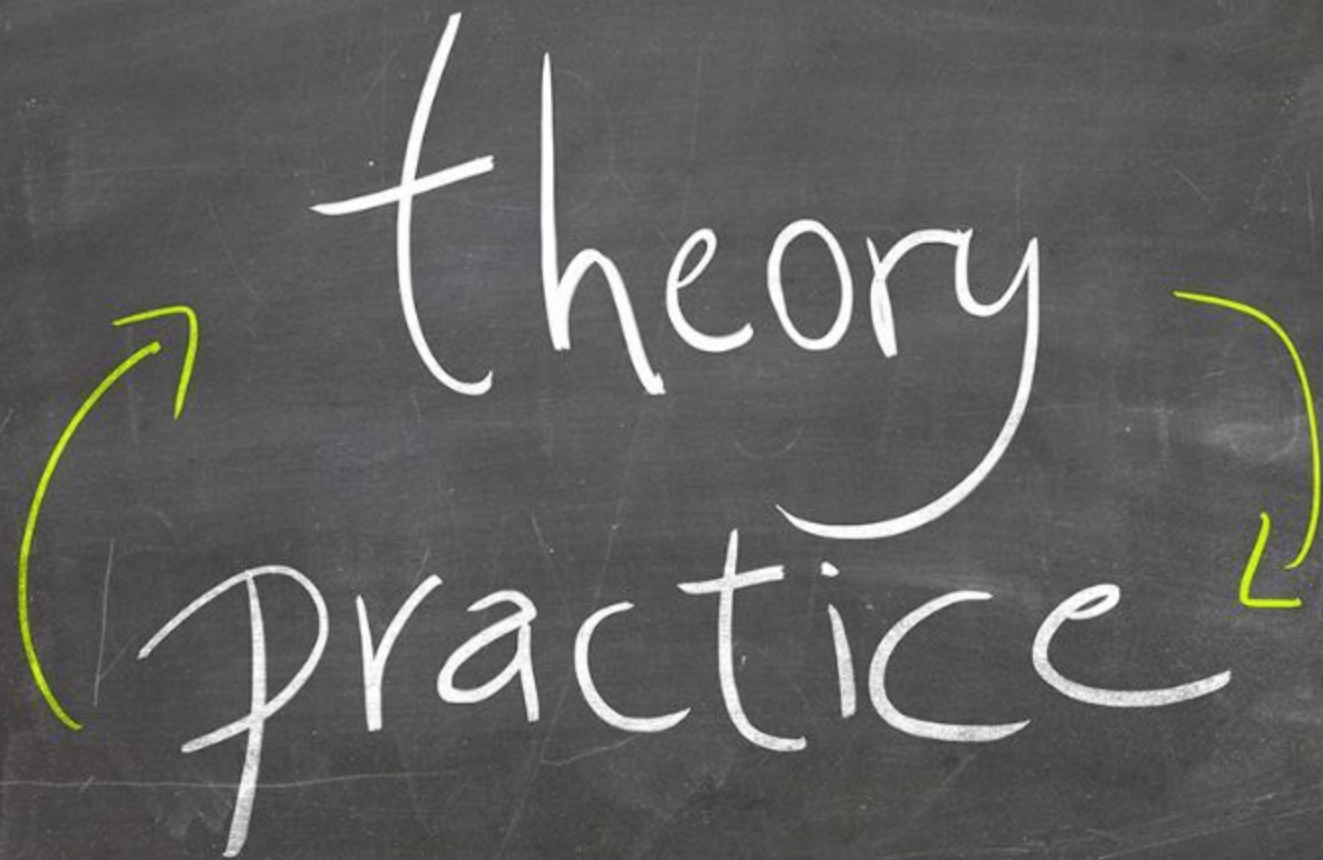
hashtags: hashtags used

user: information about the user

mentionedUsers: user information for every
user mentioned in tweet

```
'_type': 'snsrape.modules.twitter.User',
'username': 'Ow_My_Back_',
'id': 2341257349,
'displayname': 'Chris Loewen   guy living with chronic pain from a  
degenerative lumbar spine and arthritis.  Coffee, Cannabis and Cats  NO DM's   guy living with chronic pain from a  
degenerative lumbar spine and arthritis.  Coffee, Cannabis and Cats  NO DM's  location': 'Victoria BC or Mars ',
'protected': False,
'linkUrl': None,
'linkTcourl': None,
'profileImageUrl':
https://pbs.twimg.com/profile_images/1500243345829892096/tCIg8tDZ_normal.jpg',
'profileBannerUrl': 'https://pbs.twimg.com/profile_banners/2341257349/1650606681',
'label': None,
'url': 'https://twitter.com/Ow_My_Back_'
```

theory
practice



```
graph TD; theory --> practice; practice --> theory;
```

Next Steps



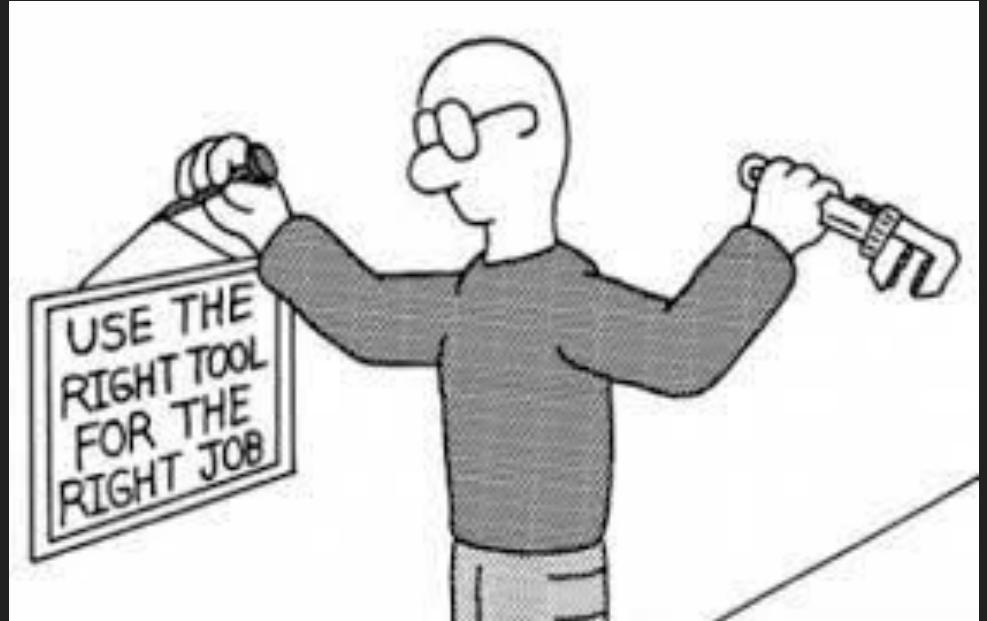
Where to learn more about snsrape

- [Official Github Repo](#)
- [Projects using snsrape](#)
- [Medium Reading List](#)

Summary



**DAS
LEBEN
IST
KEIN
PONYHOF**



[illegible]