Tugas Praktikum-1 Natural Language Processing (NLP)

Nama : Gavrilla Claudia

NIM : 21110004

```
# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" |

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
```

```
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:12 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [1,337 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,266 kB]
Get:14 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy/main amd64 Packages [21.7 kB]
Get:15 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Get:16 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [1,156 kB]
Get:17 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [1,037 kB]
Get:18 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [1,000 kB]
Fetched 7,941 kB in 3s (2,550 kB/s)
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
ca-certificates is already the newest version (20230311ubuntu0.22.04.1).
curl is already the newest version (7.81.0-1ubuntu1.13).
gnupg is already the newest version (2.2.27-3ubuntu2.1).
gnupg set to manually installed.
0 upgraded, 0 newly installed, 0 to remove and 30 not upgraded.
deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_20.x nodistro main
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Get:2 https://deb.nodesource.com/node_20.x nodistro InRelease [12.1 kB]
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Hit:4 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:6 https://deb.nodesource.com/node_20.x nodistro/main amd64 Packages [2,902 B]
Hit:7 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:8 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:9 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:12 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Fetched 15.0 kB in 1s (10.4 kB/s)
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  nodejs
0 upgraded, 1 newly installed, 0 to remove and 30 not upgraded.
Need to get 31.1 MB of archives.
After this operation, 195 MB of additional disk space will be used.
Get:1 https://deb.nodesource.com/node_20.x nodistro/main amd64 nodejs amd64 20.8.0-1nodesource1 [31.1 MB]
Fetched 31.1 MB in 0s (71.2 MB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/Debconf/F
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package nodejs.
(Reading database ... 120895 files and directories currently installed.)
Preparing to unpack .../nodejs_20.8.0-1nodesource1_amd64.deb ...
Unpacking nodejs (20.8.0-1nodesource1) ...
Setting up nodejs (20.8.0-1nodesource1) ...
Processing triggers for man-db (2.10.2-1) ...
v20.8.0
```

```
# Crawl Data

filename = 'Jungkook.csv'
search_keyword = 'Jungkook'
limit = 1000

!npx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit}
```

```
Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 598
Scrolling more...

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 613
Scrolling more...

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 626

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 640

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 658

--Taking a break, waiting for 10 seconds...
Scrolling more...

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 673
Scrolling more...

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 686

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 701
Scrolling more...

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/Jungkook.csv
Total tweets saved: 717

Got some tweets, saving to file...
Error parsing response json: {"_type":"Response","_guid":"response@f3b017142fd796d7070e15c507848745"}
Most likely, you have already exceeded the Twitter rate limit. Read more on https://twitter.com/elonmusk/status/1675187969420828
Scrolling more...
Scrolling more...
Scrolling more...
Scrolling more...
Already got 717 tweets, done scrolling...
npm notice
npm notice New minor version of npm available! 10.1.0 -> 10.2.0
npm notice Changelog: https://github.com/npm/cli/releases/tag/v10.2.0
npm notice Run npm install -g npm@10.2.0 to update!
npm notice
```

```
import pandas as pd

# Specify the path to your CSV file
file_path = f"tweets-data/{filename}"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path, delimiter=";")

# Display the DataFrame
display(df)
```

	created_at	id_str	full_text	quote_count	reply_count	retweet_count	favorite_count	lang	use
0	Wed Oct 04 03:23:42 +0000 2023	1709408926565826963	#JungKook_GOLDEN 👇👇	0	0	0	0	qme	16743113987
1	Wed Oct 04 03:23:42 +0000 2023	1709408926435889254	This announcement doesn't say anything about a...	0	0	0	0	en	3
2	Wed Oct 04 03:23:40 +0000 2023	1709408921524281824	Inspired by the golden moments of #JungKook, h...	0	0	0	0	en	11463178923
3	Wed Oct 04 03:23:40 +0000 2023	1709408921104908629	Inspired by the golden moments of #JungKook, h...	0	0	0	0	en	15876838426
4	Wed Oct 04 03:23:40 +0000 2023	1709408920123683144	방탄소년단 방탄 버터 가디건 포카 양도 bts butter cardigan mini...	0	0	0	0	ko	13105759952
...	...	...	...	...	...	...	...	...	
712	Wed Oct 04 03:15:40 +0000 2023	1709406906966577280	Noturnas CHARTMYS no SPF x3 e na AM rodando, ...	0	0	2	3	pt	15853355580
713	Wed Oct 04 03:15:40 +0000 2023	1709406905230373215	Jungkook GOLEDN album unopen sealed sell 정국 골...	0	0	0	0	ko	10513826080
714	Wed Oct 04 03:15:39 +0000 2023	1709406903237743058	LISTOO TAEHYUNG LE HACE EL WATER CHALLENGE A J...	0	0	0	0	en	14929863376
	Wed Oct 04								5
									2

717 rows × 12 columns

```
# Cek jumlah data yang didapatkan

num_tweets = len(df)
print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")
```

```
Jumlah tweet dalam dataframe adalah 717.
```

```
df1 = pd.DataFrame(df['full_text'])
```

```
df1
```

	full_text
0	#JungKook_GOLDEN 👇👇
1	This announcement doesn't say anything about a...
2	Inspired by the golden moments of #JungKook, h...
3	Inspired by the golden moments of #JungKook, h...
4	방탄소년단 방탄 버터 가디건 포카 양도 bts butter cardigan mini...
...	...
712	Noturnas CHARTMYS no SPF x3 e na AM rodando, ...
713	Jungkook GOLEDN album unopen sealed sell 정국 골...
714	LISTOO TAEHYUNG LE HACE EL WATER CHALLENGE A J...
715	jungkook bobo keknya yh
716	Is There Such A Thing As Original Asian Sound?...

717 rows × 1 columns

```
df.to_csv('Jungkook.csv')
```