Penjelasan Kode Python: Scraper Instagram Async dengan Playwright

Kode ini adalah **web scraper otomatis untuk Instagram** menggunakan pustaka **Playwright (asynchronous)**. Tujuannya:

1. Login ke akun Instagram.
2. Mengunjungi beberapa akun publik.
3. Mengambil caption dan timestamp dari postingan.
4. Menyimpan data dalam file JSON.

Bagian 1: Import Library

```
import asyncio
from playwright.async_api import async_playwright
import json
import re
import os
from datetime import date
```

- `asyncio`: Menjalankan fungsi async.
- `playwright.async_api`: API Playwright async.
- `json`: Simpan hasil ke JSON.
- `re`: Manipulasi teks (regex).
- `os`: Manipulasi folder & path.
- `date`: Ambil tanggal hari ini.

Bagian 2: Kredensial dan Parameter

```
IG_USERNAME = "IG_USERNAME"
IG_PASSWORD = "IG_PASSWORD"
```

Ganti dengan username & password asli.

```
TARGET_USERNAMES = [
    "bogor_update",
    "bogor24update",
    "depok24jam"
]
```

Daftar akun target.

```
MAX_POSTS = 5
OUTPUT_DIR = "json"
```

- `MAX_POSTS`: Maksimum jumlah postingan per akun.

- `OUTPUT_DIR`: Folder untuk file output.

Bagian 3: Login Instagram

```
async def login_instagram(page):
```

Melakukan login ke Instagram:

```
    await page.goto("https://www.instagram.com/accounts/login/", timeout=60000)
    await page.wait_for_selector('input[name="username"]', timeout=60000)
    await page.fill('input[name="username"]', IG_USERNAME)
    await page.fill('input[name="password"]', IG_PASSWORD)
    await page.click('button[type="submit"]')
    await page.wait_for_url("https://www.instagram.com", timeout=60000)
    await page.wait_for_timeout(1000)
```

Bagian 4: Scrape Akun Instagram

```
async def scrape_account(context, target_username):
```

1. Buka Profil

```
    page = await context.new_page()
    await page.goto(f"https://www.instagram.com/{target_username}/",
timeout=60000)
    await page.wait_for_selector('div.xg7h5cd.x1n2onr6', timeout=60000)
    await page.wait_for_timeout(3000)
```

2. Scroll untuk muat postingan

```
    for _ in range(3):
        await page.mouse.wheel(0, 3000)
        await page.wait_for_timeout(2000)
```

3. Ambil URL Postingan

```
    links = await page.locator('div.xg7h5cd.x1n2onr6 a').all()
    post_urls = []
    for link in links:
        href = await link.get_attribute("href")
        if href:
            post_urls.append(f'https://www.instagram.com/{href}')
    post_urls = list(dict.fromkeys(post_urls))[:MAX_POSTS]
```

4. Ambil Data Tiap Postingan

```
    posts = []

    for url in post_urls:
        await page.goto(url, timeout=60000)
        await page.wait_for_selector('div.xtOpsk2', timeout=60000)
        await page.wait_for_timeout(2000)

        caption_el = page.locator('div[role="button"] h1').first
        raw_caption = await caption_el.inner_text() if caption_el else ""
```

5. Bersihkan Caption

```
        caption = re.sub(r'[#@]\S+', '', raw_caption)
        caption = re.sub(r'[^\w\s,\.\-()/]', '', caption)
        caption = re.sub(r'\s+', ' ', caption).strip()
```

6. Ambil Timestamp

```
        time_el = page.locator('time').first
        timestamp = await time_el.get_attribute("datetime") if time_el else ""
```

7. Simpan Hasil

```
        posts.append({
            "platform": "instagram",
            "username": target_username,
            "text": caption,
            "timestamp": timestamp,
            "url": url,
        })
```

Bagian 5: Scraping Semua Akun

```
async def scrape_instagram():
```

1. Persiapan

```
    os.makedirs(OUTPUT_DIR, exist_ok=True)
    all_posts = []
```

2. Buka Browser

```
    async with async_playwright() as p:
        browser = await p.chromium.launch(headless=False)
        context = await browser.new_context()
```

3. Login Sekali

```
        page = await context.new_page()
        await login_instagram(page)
        await page.close()
```

4. Scrape Semua Akun

```
        for username in TARGET_USERNAMES:
            account_posts = await scrape_account(context, username)
            all_posts.extend(account_posts)
        await browser.close()
```

Simpan ke File JSON

```
    output_path = os.path.join(OUTPUT_DIR, f"{date.today()}_all_accounts.json")
    with open(output_path, "w", encoding="utf-8") as f:
        json.dump(all_posts, f, ensure_ascii=False, indent=2)
```

Eksekusi Program

```
if __name__ == "__main__":
    asyncio.run(scrape_instagram())
```

Memulai program scraping ketika file dijalankan. Command terminal untuk menjalankan program:

```
python [Nama_File].py
```

Instalasi Playwright (jika belum)

```
pip install playwright
python -m playwright install
```

atau

```
npm install -D playwright
npx playwright install
```

Contoh Output JSON

```
{
  "platform": "instagram",
  "username": "bogor_update",
  "text": "Contoh caption tanpa hashtag...",
  "timestamp": "2025-05-13T08:00:00.000Z",
  "url": "https://www.instagram.com/p/abcdefg/"
}
```

Catatan Tambahan

- Pastikan kredensial valid dan tidak terkunci.
- Instagram mungkin menampilkan CAPTCHA atau verifikasi login.
- Jangan scrape akun private tanpa follow dan persetujuan.
- Untuk scraping tambahan (media, komentar), kode perlu dikembangkan lebih lanjut.