

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ใช้ 100%

+ Gemini

+

+

โค้ด

ข้อความ

สมุดบันทึก

6630250435_HW2.csv

...

HW#2 _DataScrapping

นาย วรินทร์ สายปัญญา รหัสนิสิต 6630250435 sec. 870

คำตอบของคำถาม 2 ข้อ

1.) Website ที่เลือกคือ <https://www.rottentomatoes.com>

2.) เป็น AI model มีรูปแบบที่วิเคราะห์สรุปข้อความที่มาจากผู้ชมหนึ่งว่าหนังเรื่องนี้มีคุณภาพเป็นอย่างไรเหมาะกับผู้ชมแบบไหน โดยวิเคราะห์จากข้อความที่รีวิว และ คะแนนที่ให้

อธิบายการทำงานของโค้ด

0 วินาที

เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ใช้ 100%

+ Gemini

+

+

โค้ด

ข้อความ

สมุดบันทึก

6630250435_HW2.csv

...

อธิบายการทำงานของโค้ด

นำเข้า modules ของ libraries ที่ต้องใช้

[4] from pandas import DataFrame
from requests import get
from bs4 import BeautifulSoup

สร้างตัวแปรเก็บลิงค์ของ url ไว้จากนั้นทำการยิง http request โดยใช้ method get เพื่อเอาข้อมูลหน้าเว็บมา และ สร้าง object ชื่อ soup โดยส่ง argument เป็น ข้อมูลหน้าเว็บเข้าไป และ ทดสอบดึงข้อมูลออกมาแสดง

[9] url = 'https://www.rottentomatoes.com/m/venom_the_last_dance/reviews?type=verified_audience'
response = get(url)
soup = BeautifulSoup(response.content, "html.parser")

print("รีวิวของหนังเรื่อง " + soup.find("a", {'data-qa': 'sidebar-media-link'}).string.strip())

0 วินาที

เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ใช้ 100%

+ Gemini

+

+

โค้ด

ข้อความ

สมุดบันทึก

6630250435_HW2.csv

...

[9] url = 'https://www.rottentomatoes.com/m/venom_the_last_dance/reviews?type=verified_audience'
response = get(url)
soup = BeautifulSoup(response.content, "html.parser")

print("รีวิวของหนังเรื่อง " + soup.find("a", {'data-qa': 'sidebar-media-link'}).string.strip())

รีวิวของหนังเรื่อง Venom: The Last Dance

สร้าง function สำหรับดึงข้อมูล reviews

1. เขียน selector เพื่อเข้าถึง elements ที่ต้องการโดยต้องการ content ด้านใน element

2. สร้างตัวแปร reviews เป็น array ชนิด string แล้วใช้ method select ของ object soup ระบุ keyword argument เป็น selector เพื่อเข้าถึง elements จากนั้นจะได้ข้อมูลที่เป็น list ของ element จากนั้นทำการ loop element แต่ละตัวเพื่อดึงข้อความ content ออกมา และ ใช้ method strip เพื่อลบ whitespace ออกจากข้อความ

3. ต่อไป loop ข้อความใน list เพื่อลบตัวอักขระพิเศษ "\n" ออกจากข้อความรีวิวด้านในโดยใช้ method find เพื่อมองถึงตำแหน่ง index ของอักขระพิเศษ ถ้าเจอจะได้เลข index มา ถ้าไม่เจอคืนค่ากลับมา -1 แล้วใช้ if else เช็คถ้าเจออักขระพิเศษใช้ method replace แทนข้อความของอักขระพิเศษ เป็น ข้อความว่าง จากนั้น return list ชนิด string ออกจาก function

0 วินาที

เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ดีสก์

Gemini

สมุดบันทึก 6630250435_HW2.csv

[9] print("รีวิวของหนังเรื่อง " + soup.find("อ", {'data-qa': 'sidebar-media-link'}).string.strip())

รีวิวของหนังเรื่อง Venom: The Last Dance

สร้าง function สำหรับการดึงข้อมูล reviews

1. เขียน selector เพื่อเข้าถึง elements ที่ต้องการโดยต้องการ content ด้านใน element

2. สร้างตัวแปร reviews เป็น array ชนิด string แล้วใช้ method select ของ object soup ระบุ keyword argument เป็น selector เพื่อเข้าถึง elements จากนั้นจะได้ข้อมูลที่เป็น list ของ element มาจากนั้นทำการ loop element แต่ละตัวเพื่อดึงข้อความ content ออกมา และ ใช้ method strip เพื่อลบ whitespace ออกจากข้อความ

3. ต่อไป loop ข้อความใน list เพื่อลบตัวอักขระพิเศษ "ก" ออกจากข้อความรีวิวด้านในโดยใช้ method find เพื่อมองถึงตำแหน่งเลข index ของอักขระพิเศษ ถ้าเจอจะได้เลข index มา ถ้าไม่เจอคืนค่ากลับมา -1 แล้วใช้ if else เช็คว่าเจออักขระพิเศษใช้ method replace แทนข้อความของอักขระพิเศษ เป็นข้อความว่าง จากนั้น return list ชนิด string ออกมาจาก function

4. ทำการแสดงผลลัพธ์ข้อมูลของ function

[10] def get_all_reviews():
 selector = "drawer-more > p.audience-reviews__review"

0 วินาที เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ดีสก์

Gemini

สมุดบันทึก 6630250435_HW2.csv

[10] def get_all_reviews():
 selector = "drawer-more > p.audience-reviews__review"
 reviews = [review.string.strip() for review in soup.select(selector=selector)]

 target_char = "\n"
 for i in range(len(reviews)):
 idx = reviews[i].find(target_char)

 if idx == -1:
 pass
 else:
 reviews[i] = reviews[i].replace(target_char, "")

 return reviews

 print(get_all_reviews())

[Fun if you enjoy the dumb humor and buddy cop dynamic they have, but not for me. Adios spidermanless spiderman villain', 'The bar wasn't set very high for me so I thoroughly e

0 วินาที เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ดีสก์

Gemini

สมุดบันทึก 6630250435_HW2.csv

สร้าง function สำหรับดึงชื่อผู้รีวิว

1. เขียน selector เพื่อเข้าถึง elements

2. สร้างตัวแปร reviewers เก็บเป็น list ของ string ด้านในใช้ method select ส่ง argument เป็น selector จากนั้นจะได้ list ของ element มาทำการ loop แต่ละ element เพื่อเอาข้อความด้านใน element โดยเข้าถึง attribute string

3. return ตัวแปร reviewers

4. แสดงผลลัพธ์ข้อมูลของ function

def get_all_reviewers():
 selector = "div.audience-review-row > div.review-data > div.audience-reviews__name-wrap > span.audience-reviews__name"
 reviewers = [reviewer.string for reviewer in soup.select(selector=selector)]
 return reviewers

 print(get_all_reviewers())

['Paxton Wheat', 'Larry G', 'Anthony A', 'Jacob', 'william c', 'Jeffrey', 'Evelyn C', 'Cody', 'Tyreneis', 'K B', 'Eduardo Sanchez', 'Ivan', 'Jim', 'Kevin H', 'Maria uriarte', 'thomas s', 'frederic',

0 วินาที เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

RAM 100% ดิสก์ 100%

Gemini

สมุดบันทึก 6630250435_HW2.csv

สร้าง function สำหรับดึงวันที่รีวิว

1. เขียน selector เพื่อเข้าถึง elements

2. สร้างตัวแปร dates เก็บเป็น list ของ string ด้านในใช้ method select ส่ง argument เป็น selector จากนั้นจะได้ list ของ element มาทำการ loop แต่ละ element เพื่อเอาข้อความด้านใน element โดยเข้าถึง attribute string

3. return ตัวแปร dates

4. แสดงผลลัพธ์ข้อมูลของ function

[12]

def get_all_dates():
 selector = "div.audience-review-row > div.review-text-container > div.audience-review-meta > rating-stars-group + span.audience-reviews__verified + span.audience-reviews_
 dates = [date.string for date in soup.select(selector=selector)]
 return dates

 print(get_all_dates())

['Dec 18, 2024', 'Dec 18, 2024', 'Dec 17, 2024', 'Dec 17, 2024', 'Dec 16, 2024', 'Dec 14, 2024', 'Dec 13, 2024', 'Dec 13, 2024', 'Dec 12, 2024', 'Dec 10, 2024', 'Dec 10, 2024', 'Dec 9,

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

RAM 100% ดิสก์ 100%

Gemini

สมุดบันทึก 6630250435_HW2.csv

สร้าง function สำหรับดึงคะแนนรีวิว

1. เขียน selector เพื่อเข้าถึง elements

2. สร้างตัวแปร scores เก็บเป็น list ของ float ด้านในใช้ method select ส่ง argument เป็น selector จากนั้นจะได้ list ของ element มาทำการ loop แต่ละ element เพื่อเข้าถึง attribute score เราต้องการค่าของ attribute score โดยต้องระบุ key ให้ attrs ก็จะได้ค่าของ attribute score มาแล้วเราใช้ function float สำหรับแปลงข้อความที่ได้ให้เป็น เลขทศนิยม

3. return ตัวแปร scores

4. แสดงผลลัพธ์ข้อมูลของ function

[13]

def get_all_scores():
 selector = "div.audience-review-row > div.review-text-container > div.audience-review-meta > rating-stars-group"
 scores = [float(score.attrs["score"]) for score in soup.select(selector=selector)]
 return scores

 print(get_all_scores())

[1.0, 4.0, 3.0, 5.0, 5.0, 5.0, 5.0, 5.0, 3.0, 4.0, 3.5, 5.0, 4.0, 3.0, 5.0, 4.0, 5.0, 0.5, 5.0, 2.0]

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

RAM 100% ดิสก์ 100%

Gemini

สมุดบันทึก 6630250435_HW2.csv

1. สร้างตัวแปรชื่อ data_table มีโครงสร้างข้อมูลเป็น dict ใส่ชื่อ keys ให้เรียบร้อยแล้วกำหนด values คือการเรียกใช้ functions ที่เราสร้างขึ้นมา function จะ return list ของค่าแต่ละแถวตามลำดับ

2. สร้าง object dataframe ขึ้นมาส่ง argument เป็น dict ที่เราสร้างเมื่อที่

3. แสดงผลข้อมูลในตารางออกมา

[14]

data_table = {
 'reviewer': get_all_reviewers(),
 'review': get_all_reviews(),
 'score': get_all_scores(),
 'date': get_all_dates()
}

df = DataFrame(data_table)

print(df.info())
display(df)

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ดิสก์

Gemini

สมุดบันทึก 6630250435_HW2.csv

[14] ✓
รันใหม่

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 4 columns):
Column Non-Null Count Dtype

0 reviewer 20 non-null object
1 review 20 non-null object
2 score 20 non-null float64
3 date 20 non-null object
dtypes: float64(1), object(3)
memory usage: 768.0+ bytes
None

	reviewer	review	score	date
0	Paxton Wheat	Fun If you enjoy the dumb humor and buddy cop ...	1.0	Dec 18, 2024
1	Larry G	The bar wasn't set very high for me so I throu...	4.0	Dec 18, 2024
2	Anthony A	It was good. Not great. But worth watching	3.0	Dec 17, 2024
3	Jacob	Need to see Venom get a 4th movie with Spider-...	5.0	Dec 17, 2024
4	william c	I liked it all.The end was a little sad.	5.0	Dec 16, 2024

0 วินาที เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ดิสก์

Gemini

สมุดบันทึก 6630250435_HW2.csv

[14] ✓
รันใหม่

	reviewer	review	score	date
0	Paxton Wheat	Fun If you enjoy the dumb humor and buddy cop ...	1.0	Dec 18, 2024
1	Larry G	The bar wasn't set very high for me so I throu...	4.0	Dec 18, 2024
2	Anthony A	It was good. Not great. But worth watching	3.0	Dec 17, 2024
3	Jacob	Need to see Venom get a 4th movie with Spider-...	5.0	Dec 17, 2024
4	william c	I liked it all.The end was a little sad.	5.0	Dec 16, 2024
5	Jeffrey	Funny, this is the new type of super hero	5.0	Dec 14, 2024
6	Evelyn C	Loved it and seeing them all work together for...	5.0	Dec 13, 2024
7	Cody	This movie was a pretty good way to wrap up th...	5.0	Dec 13, 2024
8	Tyreneis	Venom the last dance is an ok movie I enjoyed...	3.0	Dec 12, 2024
9	K B	Waiting for Venom 4!	4.0	Dec 10, 2024
10	Eduardo Sanchez	Awesome absolutely loved it !	3.5	Dec 10, 2024
11	Ivan	Movie was cool had good action could have had	5.0	Dec 9, 2024

0 วินาที เสร็จสมบูรณ์เมื่อ 21:11

6630250435_HW2.ipynb

ไฟล์ แก้ไข มุมมอง แทรก รันใหม่ เครื่องมือ ความช่วยเหลือ บันทึกการเปลี่ยนแปลงทั้งหมดแล้ว

✓ RAM
ดิสก์

Gemini

สมุดบันทึก 6630250435_HW2.csv

[14] ✓
รันใหม่

11	Ivan	Movie was cool had good action could have had ...	5.0	Dec 9, 2024
12	Jim	Great follow-up to previous two movies with a ...	4.0	Dec 9, 2024
13	Kevin H	some parts are too long and would almost be be...	3.0	Dec 9, 2024
14	Maria uriarte	Absolutely love the movie I am willing to go s...	5.0	Dec 8, 2024
15	thomas s	Really cool use of Marvel visual effects. I t...	4.0	Dec 7, 2024
16	frederic	did really like it....feel sad at the end of t...	5.0	Dec 5, 2024
17	Leah	This movie was rushed and all over the place. ...	0.5	Dec 4, 2024
18	Kevin	It's just like every 2000s Marvel Flick. Good ...	5.0	Dec 4, 2024
19	Elexia j	The decliners in this auditorium are very suck...	2.0	Dec 4, 2024

ขั้นตอนถัดไป: สร้างโค้ดด้วย df ดูแผนภูมิที่แนะนำ New interactive sheet

สร้างไฟล์ csv โดยใช้ method to_csv กำหนดชื่อไฟล์ และระบุ keyword argument index เป็น False รันแล้วจะได้ไฟล์ .csv มา

0 วินาที เสร็จสมบูรณ์เมื่อ 21:11

