

Web scraping with Python on Jordanian news websites to see how many articles have been published about Orange

Dana Ghazal

1. Web Scraping: Getting Data

Web scraping is an automatic method to obtain large amounts of data from websites. Web scraping requires two parts, namely the crawler and the scraper. The crawler is an artificial intelligence algorithm that browses the web to search for the particular data required by following the links across the internet. The scraper, on the other hand, is a specific tool created to extract data from the website.

The first step is to get data from news websites, so in this code, we will use the most popular Python libraries for web scraping: urllib.request, requests and BeautifulSoup.

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import requests
import mysql.connector

url = "http://www.almadenahnews.com/search?q=%D8%A3%D9%88%D8%B1%D9%86%D8%AC"
url1 =
requests.get("https://www.addustour.com/search.php?search=%D8%A3%D9%88%D8%B1%D9%86%D8%AC")
url2 =
"https://www.jordanzad.com/index.php?page=tag&hashtag=%D8%A3%D9%88%D8%B1%D9%86%D8%AC"
url3 = "http://www.sarahanews.net/?s=%D8%A3%D9%88%D8%B1%D9%86%D8%AC"

client = urlopen(url)
client1 = url1.content
client2 = urlopen(url2)
client3 = urlopen(url3)

html = client.read()
html1 = client2.read()
html2 = client3.read()

soup = BeautifulSoup(html, "html.parser")
soup1 = BeautifulSoup(client1, "lxml")
soup2 = BeautifulSoup(html1, "html.parser")
soup3 = BeautifulSoup(html2, "html.parser")

containers = soup.find_all("div", {"class": "search_cart"})
containers2 = soup2.find_all("li")
containers3 = soup3.find_all("div", {"class": "jeg_postblock_content"})
containers1 = soup1.find_all("li", {"class": "search"})
```

```

title1 = containers[0].a.text.replace('\n', '').strip()
title2 = containers[3].a.text.replace('\n', '').strip()

date1 = containers[2].text.replace('\n', '').strip()
date2 = containers[5].text.replace('\n', '').strip()

title4 = containers2[1].a.text.replace('\n', '').strip()
title5 = containers2[2].a.text.replace('\n', '').strip()
title6 = containers2[3].a.text.replace('\n', '').strip()
title7 = containers2[4].a.text.replace('\n', '').strip()

datei = soup2.find_all("span", {"class": "date"})
date4 = datei[0].text.strip()
date5 = datei[1].text.strip()
date6 = datei[2].text.strip()
date7 = datei[3].text.strip()

title8 = containers3[0].a.text.replace('\n', '').strip()
title9 = containers3[1].a.text.replace('\n', '').strip()
title10 = containers3[2].a.text.replace('\n', '').strip()
date8 = containers3[0].div.text.replace('\n', '').strip()
date9 = containers3[1].div.text.replace('\n', '').strip()
date10 = containers3[2].div.text.replace('\n', '').strip()

```

2. Saving our data

What do we have now? The second step we want to do now is to save that data so we don't have to make those requests again. We will store the scraped data in a MySQL database by this code.

```

"""**Saving data to Mysql Database**"""
mydb = mysql.connector.connect(
    host="localhost",
    user="root",
    password="123456",
    database="orange"
)

mycursor = mydb.cursor()

sql = "INSERT INTO news_orange (News_Title, Publish_Date, News_Website)
VALUES (%s,%s,%s)"
val1 = (title1, date1, "almadenah")
val2 = (title2, date2, "almadenah")
mycursor.execute(sql, val1)
mycursor.execute(sql, val2)

for i in containers1:
    titlee = i.findAll("h3")
    title3 = titlee[0].text.strip()

    datee = i.findAll("div", {"class": "date"})
    date3 = datee[0].text.strip()
    val3 = (title3, date3, "addustour")

```

```
mycursor.execute(sql, val3)

val4 = (title4, date4, "jordanzad")
val5 = (title5, date5, "jordanzad")
val6 = (title6, date6, "jordanzad")
val7 = (title7, date7, "jordanzad")
val8 = (title8, date8, "sarahaneews")
val9 = (title9, date9, "sarahaneews")
val10 = (title10, date10, "sarahaneews")
mycursor.execute(sql, val4)
mycursor.execute(sql, val5)
mycursor.execute(sql, val6)
mycursor.execute(sql, val7)
mycursor.execute(sql, val8)
mycursor.execute(sql, val9)
mycursor.execute(sql, val10)
mydb.commit()
```

Here, the figure shows how the data was stored in a table called news_orange in a database called orange.

The screenshot displays the MySQL Workbench interface. On the left, the 'SCHEMAS' pane shows the 'orange' database with a table named 'news_orange'. The table's columns are listed: News_ID (int, AI, PK), News_Title (varchar(300)), Publish_Date (varchar(200)), and News_Website (varchar(45)).

The main window shows the 'news_orange' table with a query result grid. The query is 'SELECT * FROM orange.news_orange;'. The result grid displays 10 rows of data, including News_ID, News_Title, Publish_Date, and News_Website.

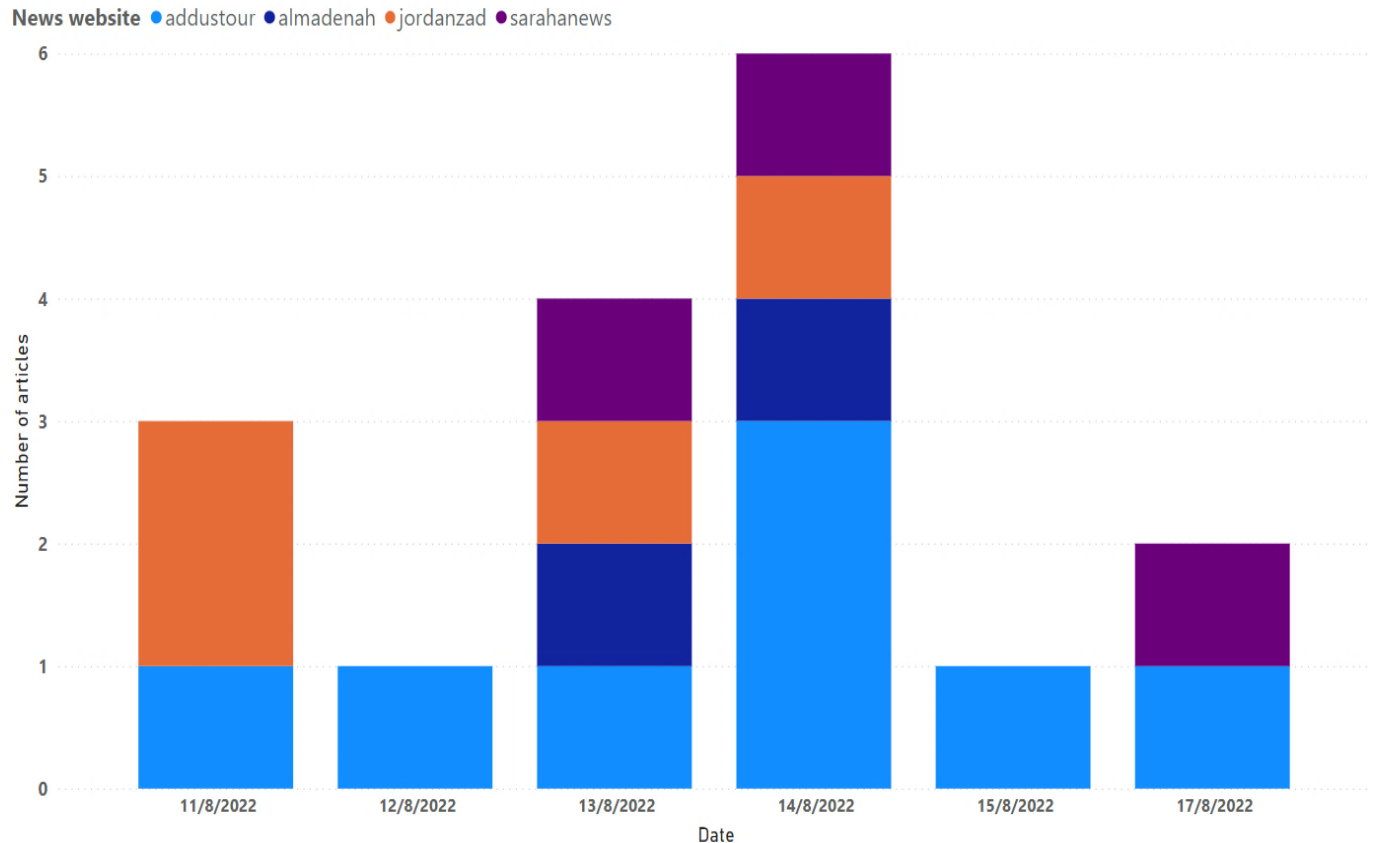
News_ID	News_Title	Publish_Date	News_Website
72	إطلاق مختبر أورنج للتصنيع الرقمي بمعهد تدريب ال...	14/8/2022	almadenah
73	أورنج والهاشمية يطلقان مختبرا للتصنيع الرقمي بد...	13/8/2022	almadenah
74	الأردن من أعلى نسب المتطوعين عربيا في التبرع ...	27/3/2022	addustour
75	«الصحبة»: 15 % من عبيات فحص كورونا تعود لـ«أورم»	8/1/2022	addustour
76	إطلاق مختبر أورنج للتصنيع الرقمي بمعهد تدريب ال...	14/8/2022	addustour
77	استكمالاً لمشروع «مساحة الابتكار» أورنج الأردن وا...	14/8/2022	addustour
78	تمديد «استدامة» لنهاية تشرين الأول	12/8/2022	addustour
79	قرارات مجلس الوزراء - تفاصيل	11/8/2022	addustour
80	مريضاً استفادوا من اليوم الطبي لجمعية البر وال...	6/7/2022	addustour
81	مريض استفادوا من اليوم الطبي المجاني لحم 892	5/7/2022	addustour
82	البر والإحسان تنظم يوما طبيا مجانيا في عجلون عدا	30/6/2022	addustour
83	الاتحاد الأوروبي وأورنج الأردن يفتتحان "قبرة أورنج ا...	27/6/2022	addustour
84	فتح باب التسجيل في الفوج الثاني من حاضنة أورنج ...	17/8/2022	addustour
85	«التدريب المهني» و«أورنج» لإطلاق مخ	15/8/2022	addustour
86	أورنج الأردن والاتحاد الأوروبي يعلنان عن برامج جديد	14/8/2022	addustour
87	أورنج الأردن والاتحاد الأوروبي يعلنان عن مختبر نصي	13/8/2022	addustour
88	... كالتفزيون : قبل ١٥ جرحى إصاباتهم خطيرة بإطلاق	17/5/2022	addustour
89	الفائزون بالسحب لجوائز تنقي مطعوم كورونا (أسماء)	21/1/2022	addustour
90	كل ما تريد معرفته عن زيدان	23/6/2022	addustour
91	جامعة الأميرة سمية تعقد مؤتمرها الأول لربط مشا	15/6/2022	addustour
92	افتتاح عمادة الابتكار ونقل التكنولوجيا والريادة ومركز	24/5/2022	addustour
93	فتح باب التسجيل في الفوج الثاني من حاضنة أورنج ...	17/8/2022	jordanzad
94	أورنج الأردن والاتحاد الأوروبي يعلنان عن برامج جديد	14/8/2022	jordanzad
95	أورنج الأردن والاتحاد الأوروبي يعلنان عن برامج جديد	13/8/2022	jordanzad

The right pane shows a message: 'Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.'

3. Visualizing the data

To create a stacked column chart showing the number of articles published per day about Orange on Jordanian news websites from 11/8/2022 to 17/8/2022, we will use the Power Bi tool.

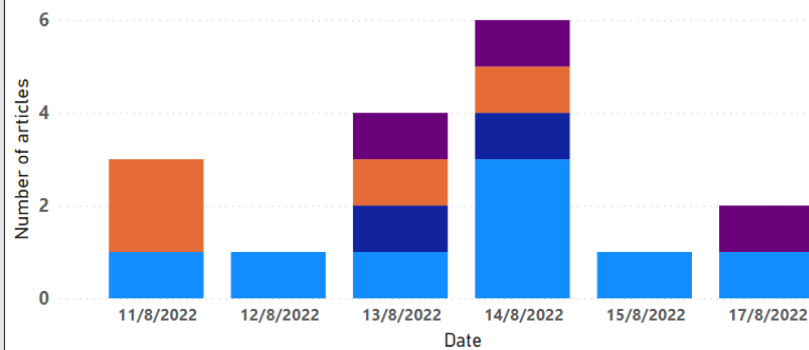
The number of articles published per day about Orange on Jordanian news websites



As shown in the figure, this graph shows that Addustour News received the largest number of articles, followed by Jordanzad News, Saraha News and Almadenah News, and that on August 14, 2022, news websites accounted for 17.65% of the number of articles, while on August 16, 2022, no article was published on news websites, and we also note that Addustour News and Jordanzad News tied with the highest average number of articles at 1.33.

The number of articles published per day about Orange on Jordanian news websites

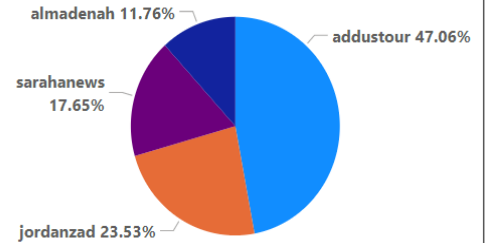
News website ● addustour ● almadenah ● jordanzad ● sarahanews



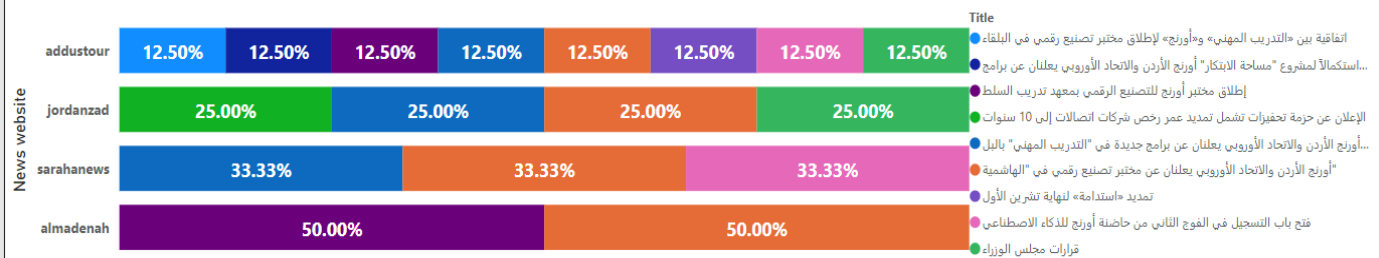
Total number of articles

17

Percent of articles by news website



Titles of articles published in every Jordanian news website



As shown in the dashboard, the total number of published articles is 17, Addustour News accounted for 47.06% of the articles. Also we note that the article titled “Orange Jordan and the European Union announce a digital manufacturing laboratory in Al Hashemite” is the most published by news websites.