Special topics I (data analysis)

(الحج و العمرة)

**Students name ID**

Muhammad Yousef 439005652

Yazeed Alhomidhi 439004374

Muayyad Al-Subhi(Leader) 439005487

**Supervised by**

Dr. Mohammed Al Ghamdi

**Task 1**

**Introduction:**

Our mission in this assignment is to collect 100,000 tweets or more from twitter about specific topic ( الحج و العمرة ), then after we collect all data that we want we will put it in excel file.

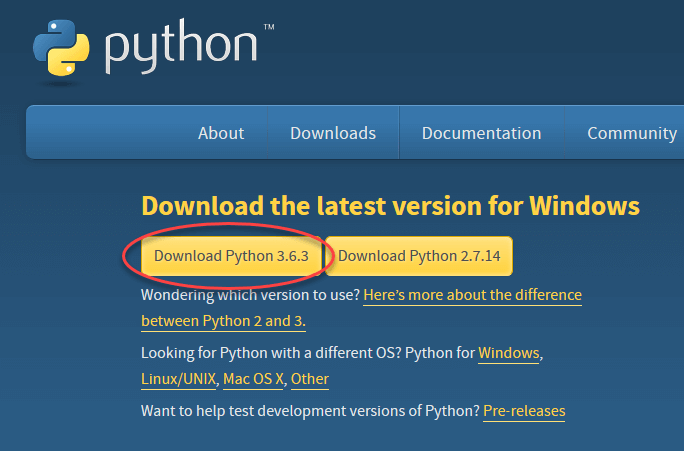
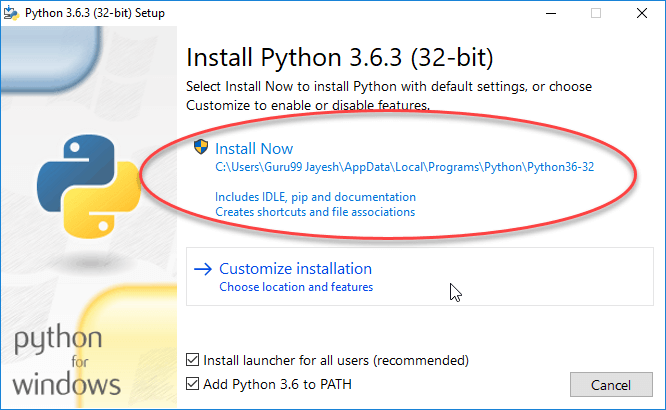
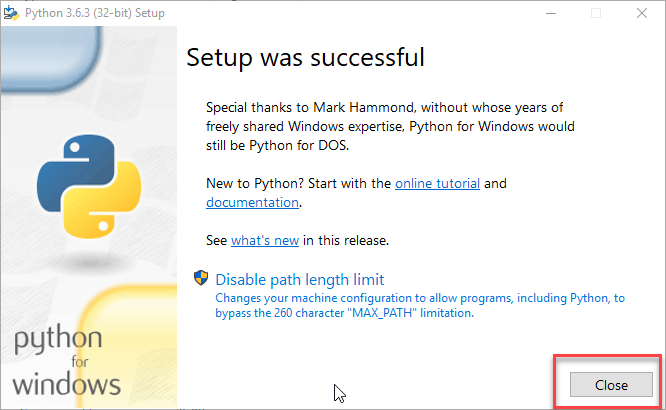
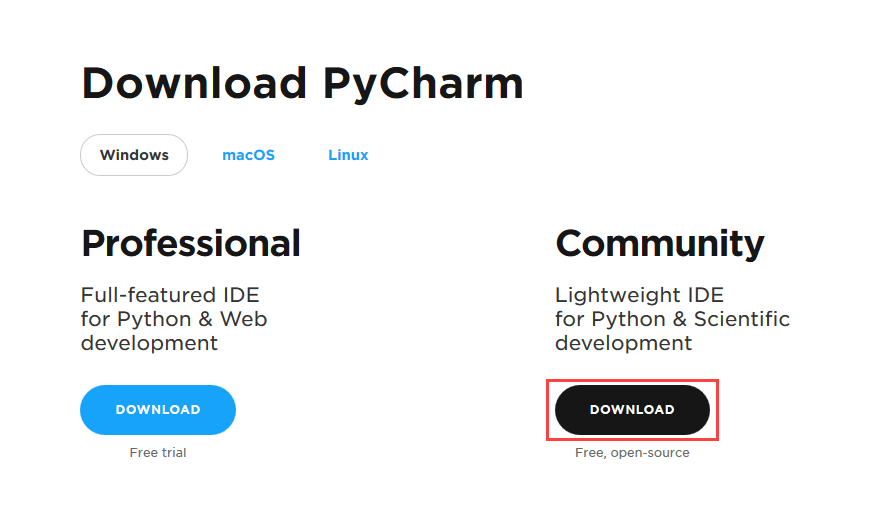
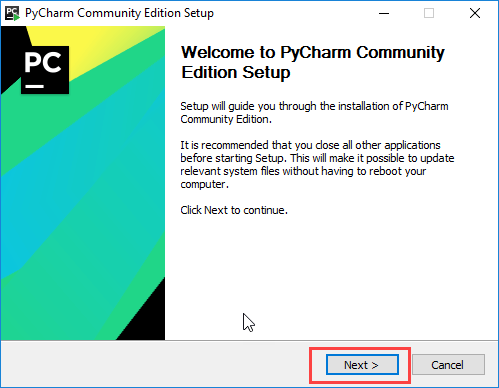
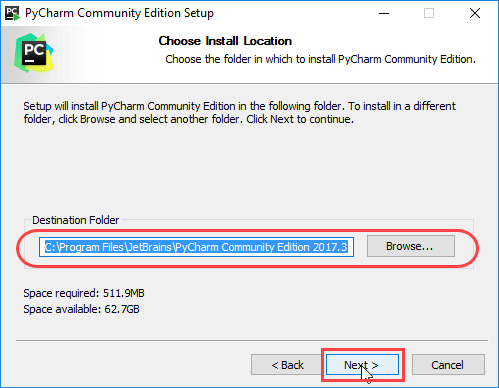
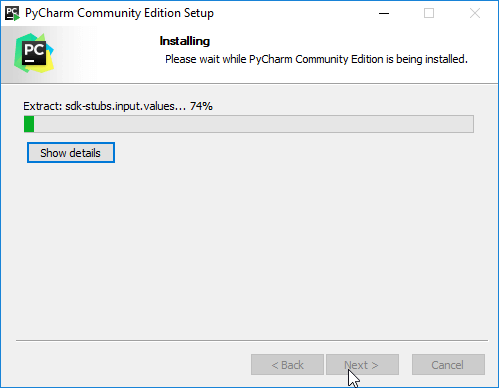
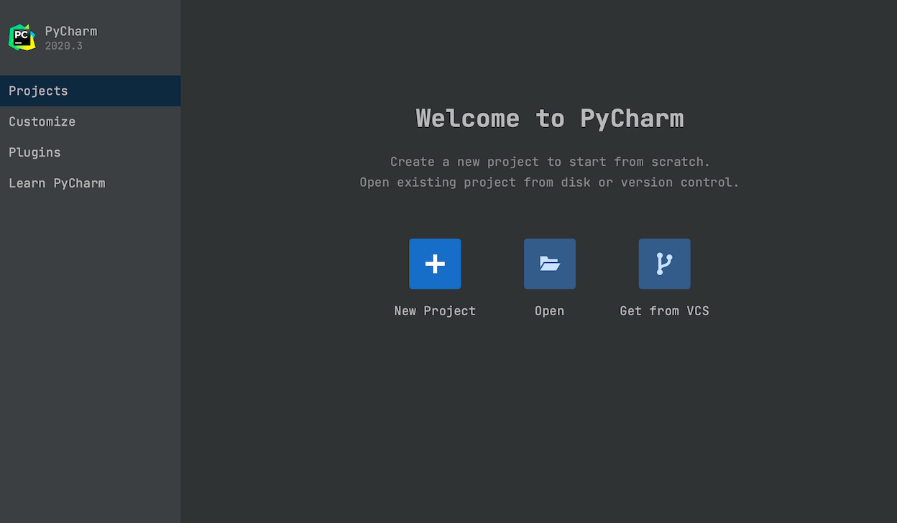
**Requirements:**

1. Collect 100,000 tweet or more.
2. Convert it into excel file.
3. Write a report for the assignment.

**Steps:**

1. **Download PyCharm to deal with Python language.**

**PyCharm** is a cross-platform editor developed by JetBrains. Pycharm provides all the tools you need for productive Python development.

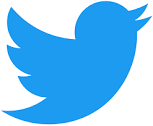
1. download and install Python <https://www.python.org/downloads/> .
2. We have selected Python version 3.6.3 
3. run the .exe file to install Python. Now click on Install Now 
4. Setup was successful. Now click on “Close” 
5. Install Pycharm form <https://www.jetbrains.com/pycharm/download/>
6. Click the “DOWNLOAD” link under the Community Section. 
7. run the exe for install PyCharm. The setup wizard should have started. Click “Next”. 
8. Change the installation path if required. Click “Next”. 
9. Wait for the installation to finish. 
10. “Run PyCharm Community Edition” box first and click “Finish”. 
11. **set the device environment for IDE use.** 
12. **Import library SNSCRAPE and PANDAS for collection of the tweets.**

* **pandas** is an open source data analysis library built on top of the Python programming language.
* **snscrape** is a scraper for social networking services (SNS). It scrapes things like user profiles, hashtags, or searches and returns the discovered items, e.g. the relevant posts. Source code [<https://github.com/JustAnotherArchivist/snscrape>]

1. Installation snscrape (pip3 install snscrape)
2. Installation pandas (import pandas as pd)
3. Select how the will be type file, we selected Excel CSV because the library supports CSV.
4. search in hashtags about Hajj and Umrah seasons and set the time period
5. run code from 2012 to 2022.
6. After collecting tweets inside the type file CSV, we will convert it to Excel XLSX, because CSV does not read the Arabic language.

**Software environment**:

* **Twitter**

is a microblogging, social networking service owned by American company Twitter, Inc., on which users post and interact with messages known as "tweets".

* **Python for collect data**

Icon

Description automatically generatedis a high-level, general-purpose programming language. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

* **Microsoft Excel for store data**

 is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA).

* **Icon

  Description automatically generatedWord for writing**

is a word processing software developed by Microsoft, it was first released on October 25, 1983, under the name Multi-Tool Word for Xenix systems.

* **Twint twitter intelligence tool**

Logo, company name

Description automatically generated“Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API.

Twint utilizes Twitter's search operators to let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like e-mail and phone numbers.

Twint also makes special queries to Twitter allowing you to also scrape a Twitter user's followers, Tweets a user has liked, and who they follow without any authentication, API, Selenium, or browser emulation.” [GitHub twint]

**Advantages:**

* Easy
* Fast
* No need to have account on twitter.
* Unlimited (+3600 tweets)

**Disadvantage:**

* Take a long time

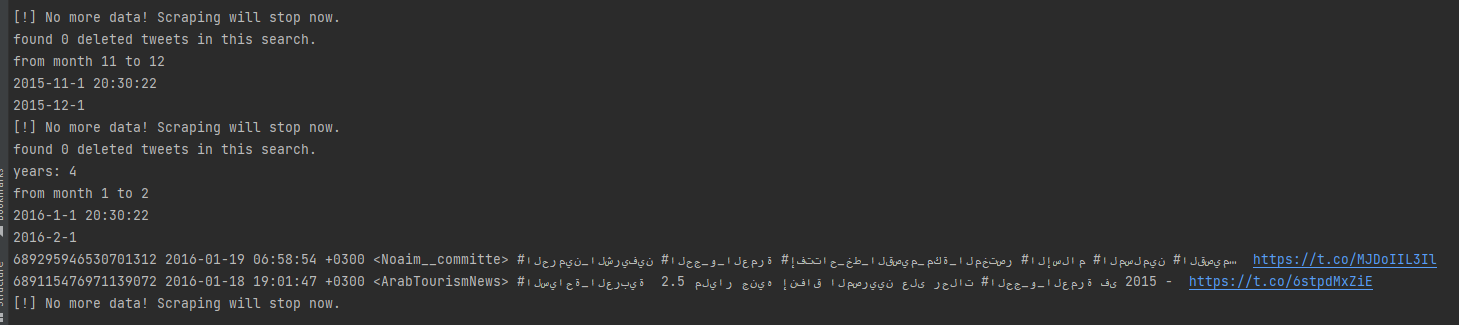
**Screenshots from code and results:**

* IDE and environment are ready:

Text

Description automatically generated with low confidence

* When running the code:



* File type csv (Not supported Arabic):

Graphical user interface, text

Description automatically generated

* After convert to xlsx:

Graphical user interface, application

Description automatically generated

**At the end we collect 1016822 tweets**

**Reference and Resources:**

* <https://www.youtube.com/watch?v=_SqgSh3aR1g&ab_channel=AhmedBesbes>
* Python Tutorial <https://www.w3schools.com/python/default.asp>
* TWINT - Twitter Intelligence Tool <https://github.com/twintproject/twint>
* <https://en.wikipedia.org/wiki/Microsoft_Word>

**Task 2**

**Introduction:**

Our mission in this assignment is filtering and cleaning the data (tweets) that we collect and organized it into a perfect shape.

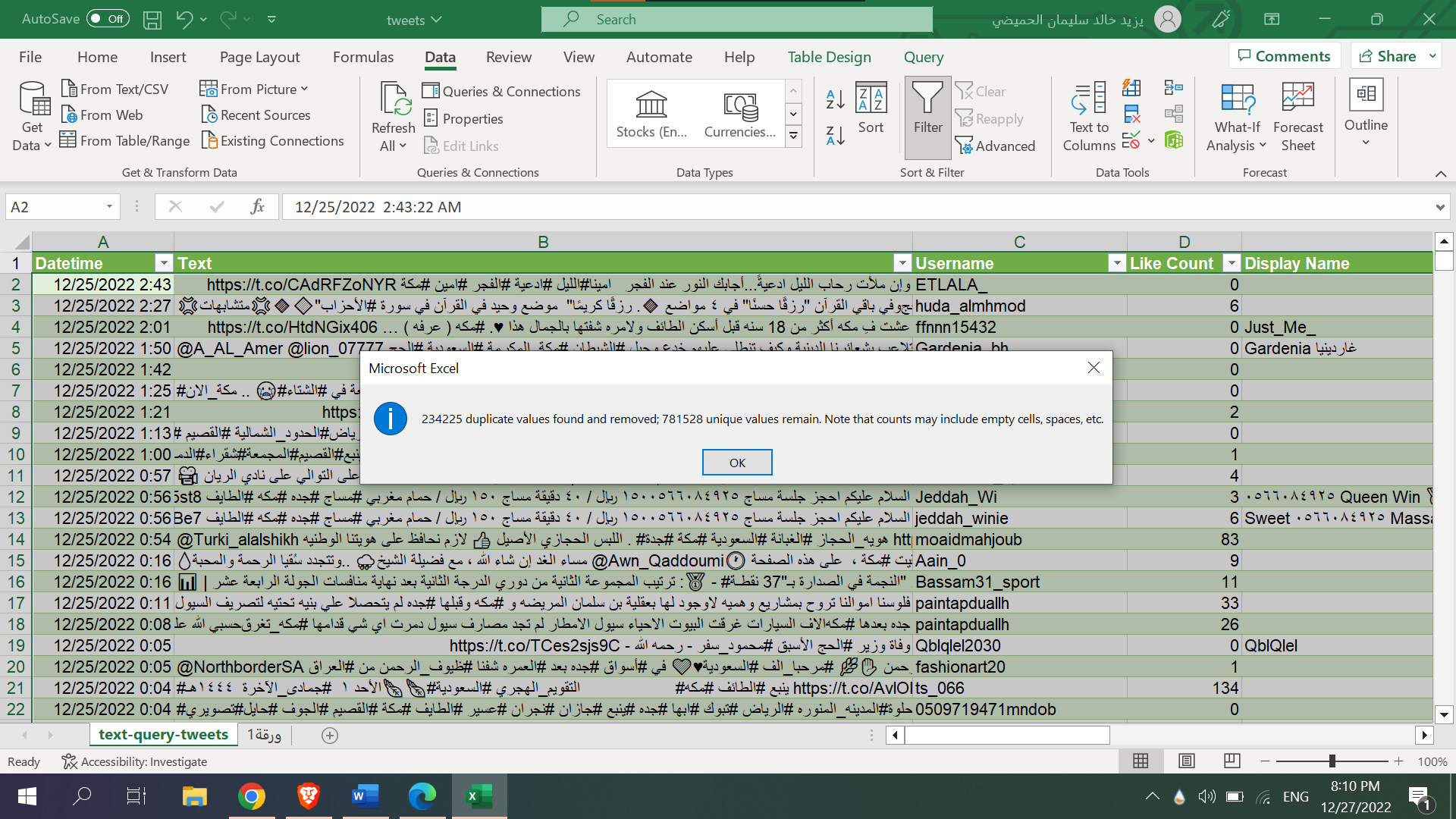
**Requirements**

* Clean data (tweets)
* Filtering data (tweets)
* Write a report for the assignment.

**Steps**

1. **Remove duplicate**

* Remove duplicate **before** removing special characters



* Remove duplicate **after** removing special characters

Graphical user interface, application, table, Excel

Description automatically generated

1. **Remove special characteristics from tweets**

* The function

Graphical user interface, application

Description automatically generated

* Before

Graphical user interface, application, Excel

Description automatically generated

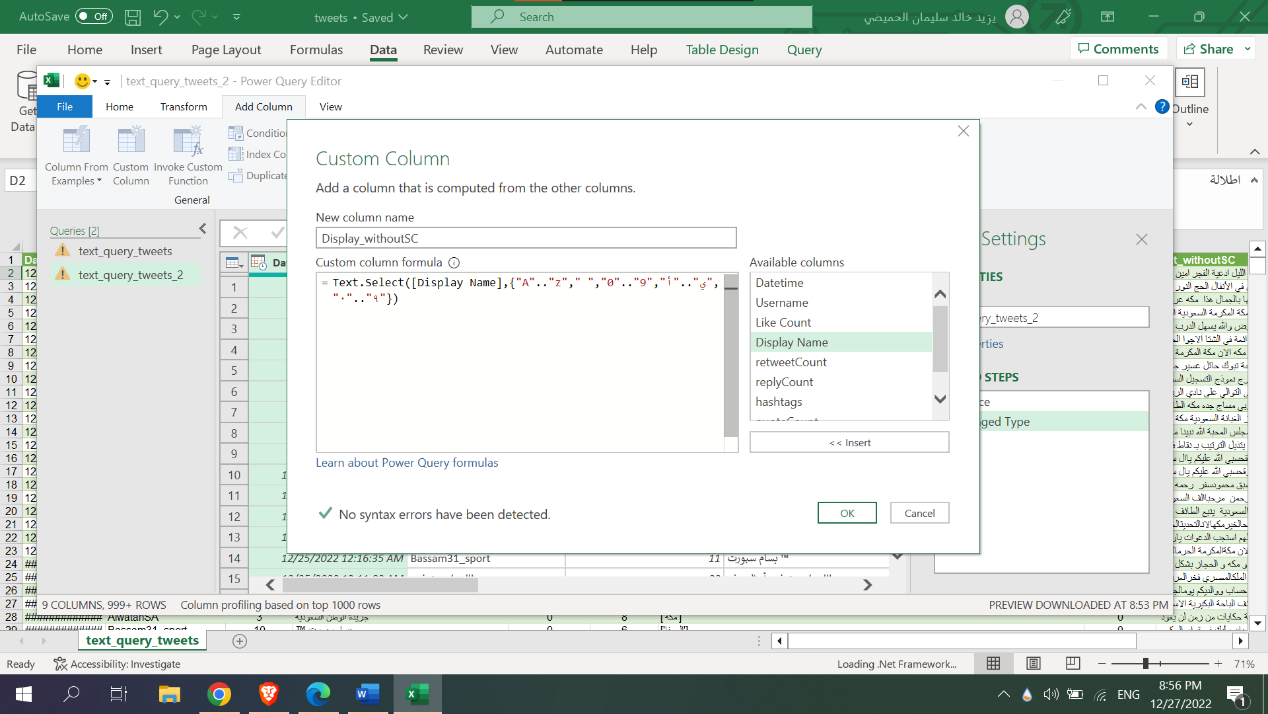
* After

Graphical user interface, application, table, Excel

Description automatically generated

1. **Remove special characteristics from display name**

* The function



* Before

Graphical user interface, application, table

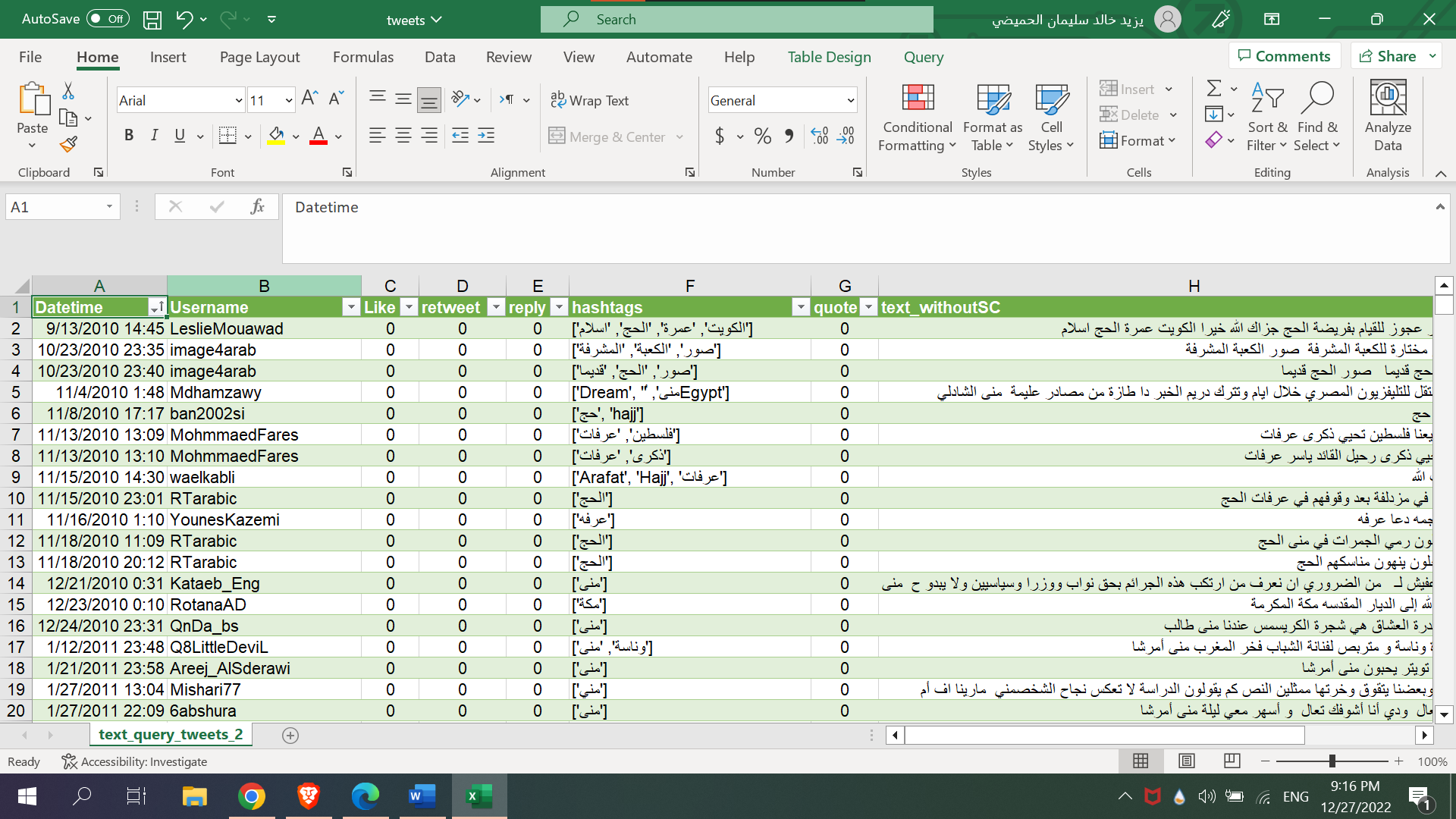
Description automatically generated

* After

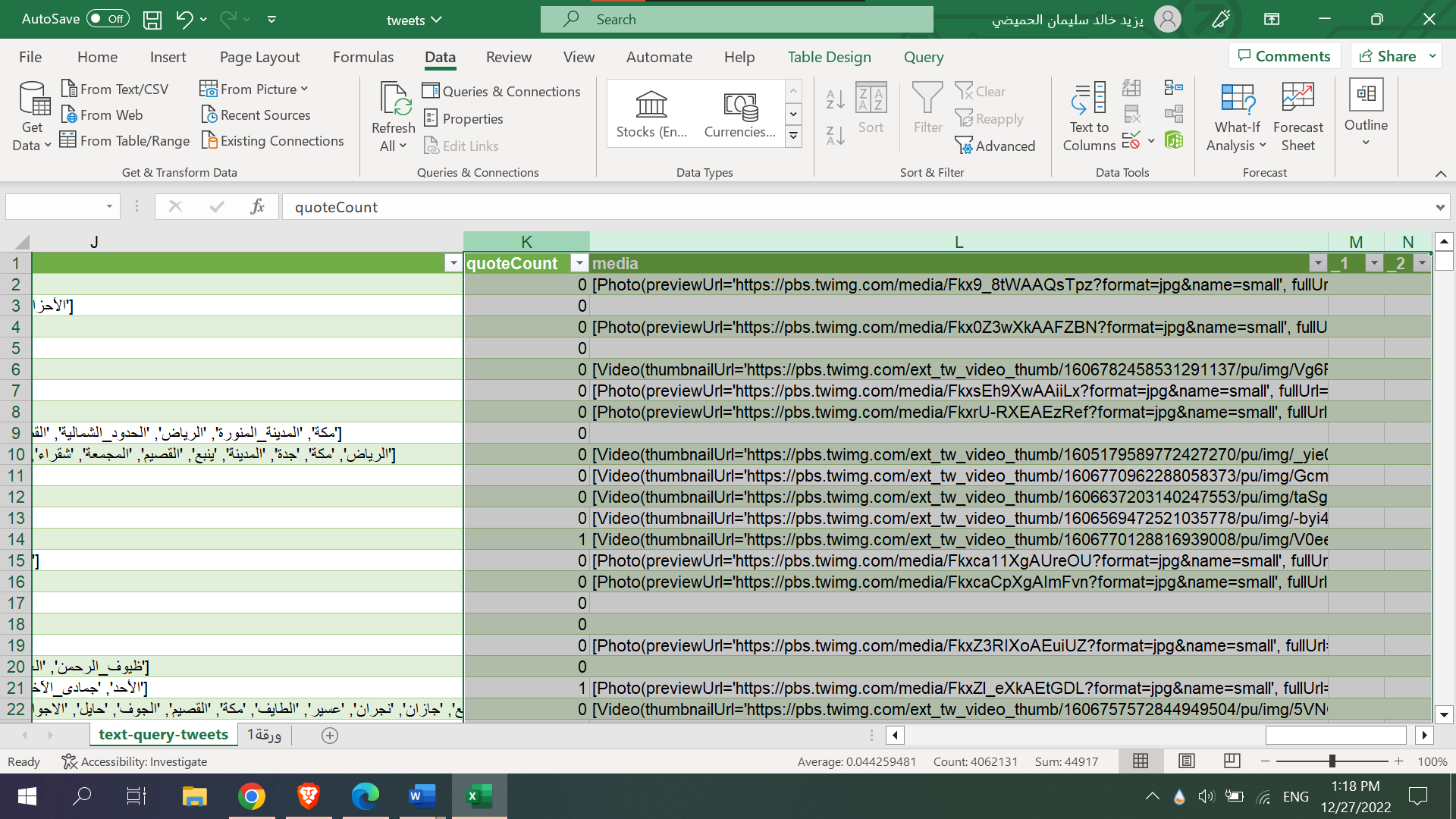
Graphical user interface, application

Description automatically generated

1. **organized data by the date from oldest to newest**



1. **Remove any columns that are unnecessary.**



**Reference and Resources:**

* <https://www.youtube.com/watch?v=e0TfIbZXPeA>
* <https://www.youtube.com/watch?v=_jmiEGZ6PIY>
* <https://www.youtube.com/watch?v=8POlb547SDQ&t=47s>

**Task 3**

**Introduction:**

In this task, it is required to convert the tweets into Arabic stemmers and format them into a column in Excel

**Inference:**

After conducting research, reading and comparisons between other libraries that perform the Arabic Stemmers process for tweets, we found the **ISRIStemmer** library the most efficient in terms of time and speed compared to other libraries such as **Tashaphyne** and **Snowball Stemmer** and in terms of the ease of converting tweets and decoding them word by word and rooting them, so we decided to use the **ISRIStemmer** library and continue in The logging of texts and tweets took about 8 hours due to the density and size of the large data.

**Tashaphyne** is a light stemmer and segmentor for Arabic. It provides all conceivable segmentations and mostly enables light stemming (removing prefixes and suffixes). In order to create all segmentations, it uses a modified finite state automaton.

**Snowball Stemmer:** This stemming method, commonly referred to as the Porter2 stemming algorithm, is an improved version of the Porter Stemmer since some of its shortcomings have been addressed.

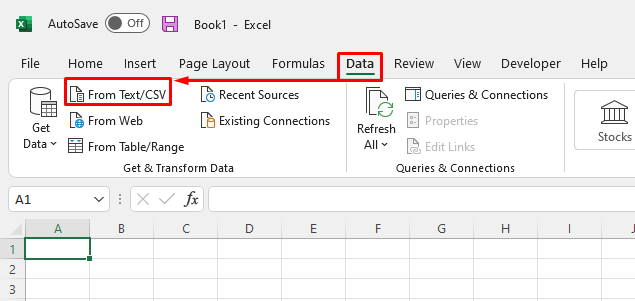
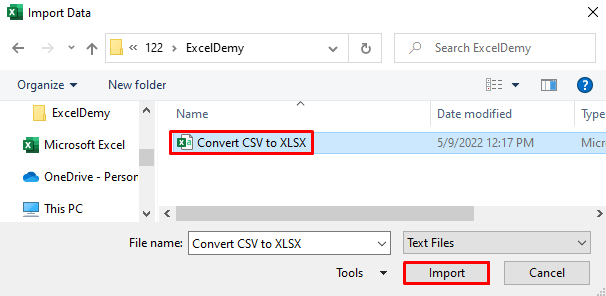
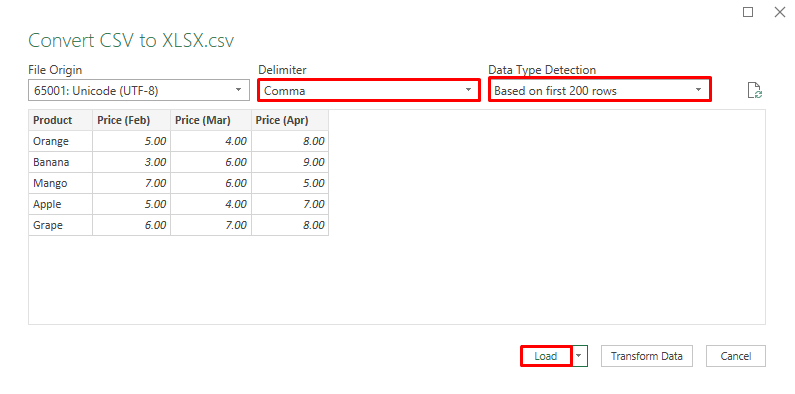
**Requirements:**

* **Installing NLTK:** requires Python versions 3.7, 3.8, 3.9 or 3.10
* **Setting up a Python Environment:**
  + Install NLTK: run pip install --user -U nltk
  + Install Numpy (optional): run pip install --user -U numpy
  + Test installation: run python then type import nltk
* **Import Arabic-Stemmers:** Arabic-Stemmers is a proxy that allows access to market arabic stemmers.
  + **Install** : from nltk.stem.isri import ISRIStemmer
  + **Import pandas:** is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way towards this goal.
  + **Install**: import pandas as pd

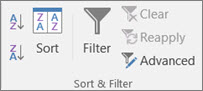
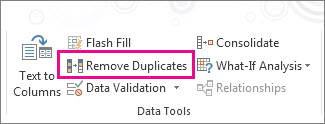
**Code:**

After collecting the data in a csv file, we will convert it to xlsx, to show the data in a consistent manner.

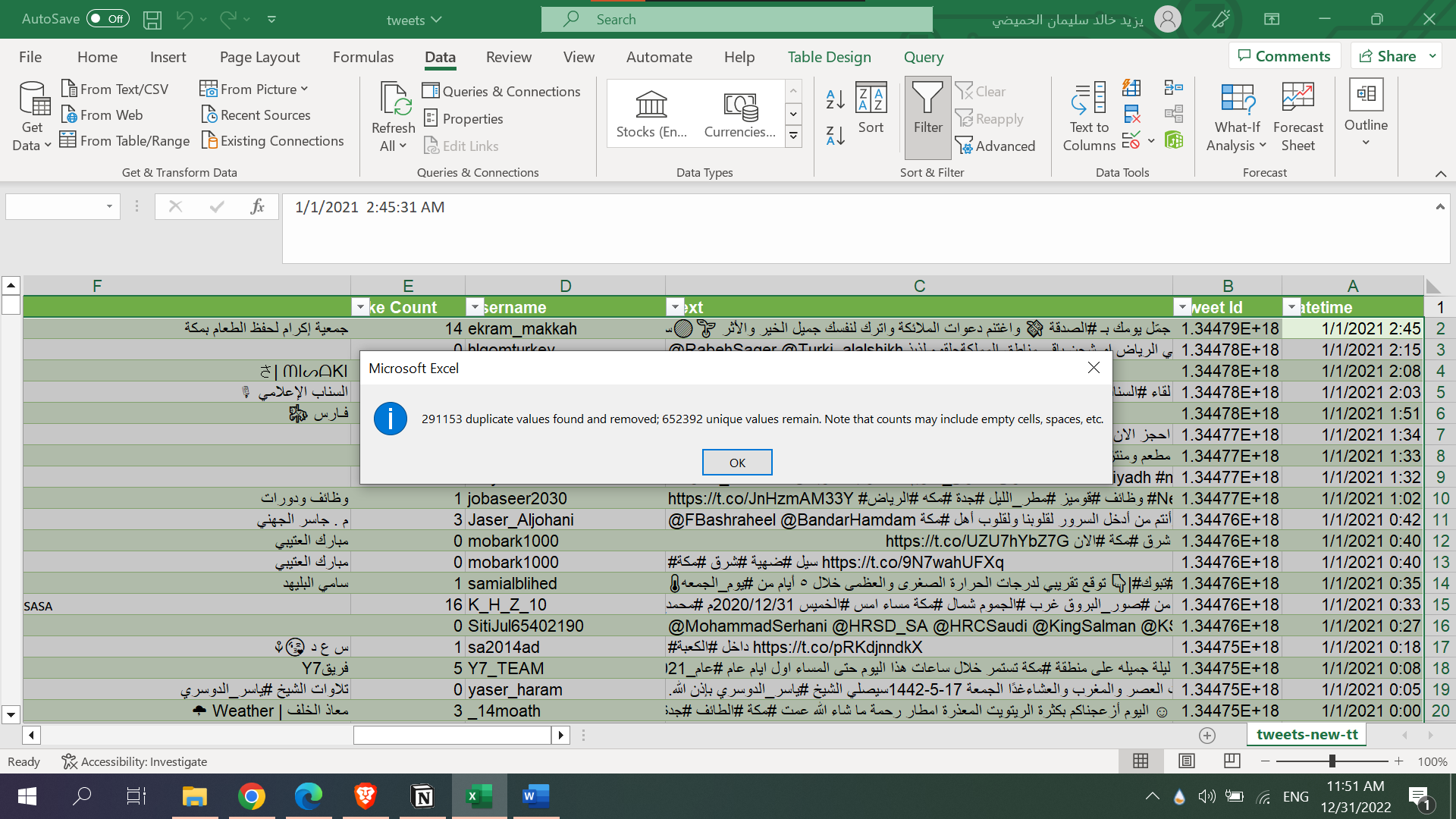
**Steps:**

* Open the Excel app.
* Data ➤ From Text/CSV. 
* after appearing the **Import**, select your **CSV file.** 
* Make sure select **comma** and **Based on the first 200 rows** 
* click **Load**.

**Remove duplicate:**

* To filter for unique values, click Data > Sort & Filter > Advanced.  
  
* To remove duplicate values, click Data > Data Tools > Remove Duplicates. ****

**Result:**

1- 

2-

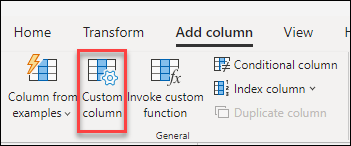
Graphical user interface

Description automatically generated

**Function for DispalyName**

* **Create a custom column**

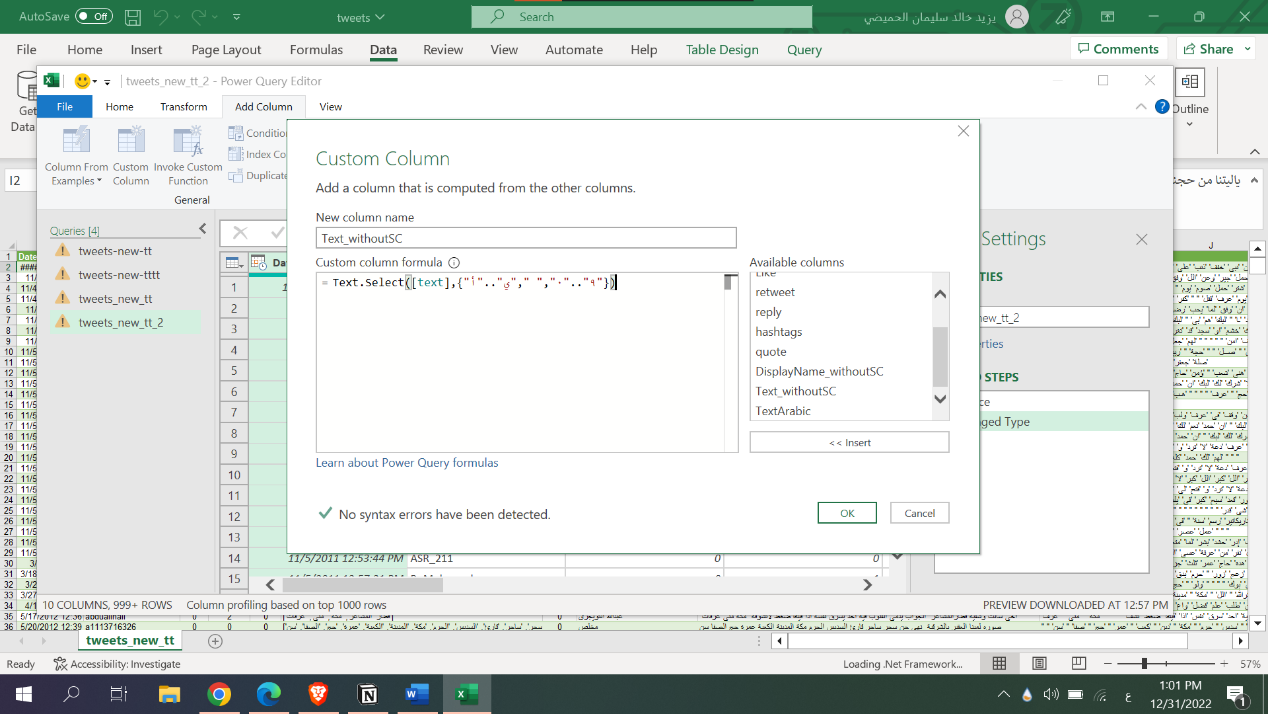
On the Add column tab, select Custom column.

****

**You specify the formula to construct your column in this dialog box.**



Function for Text

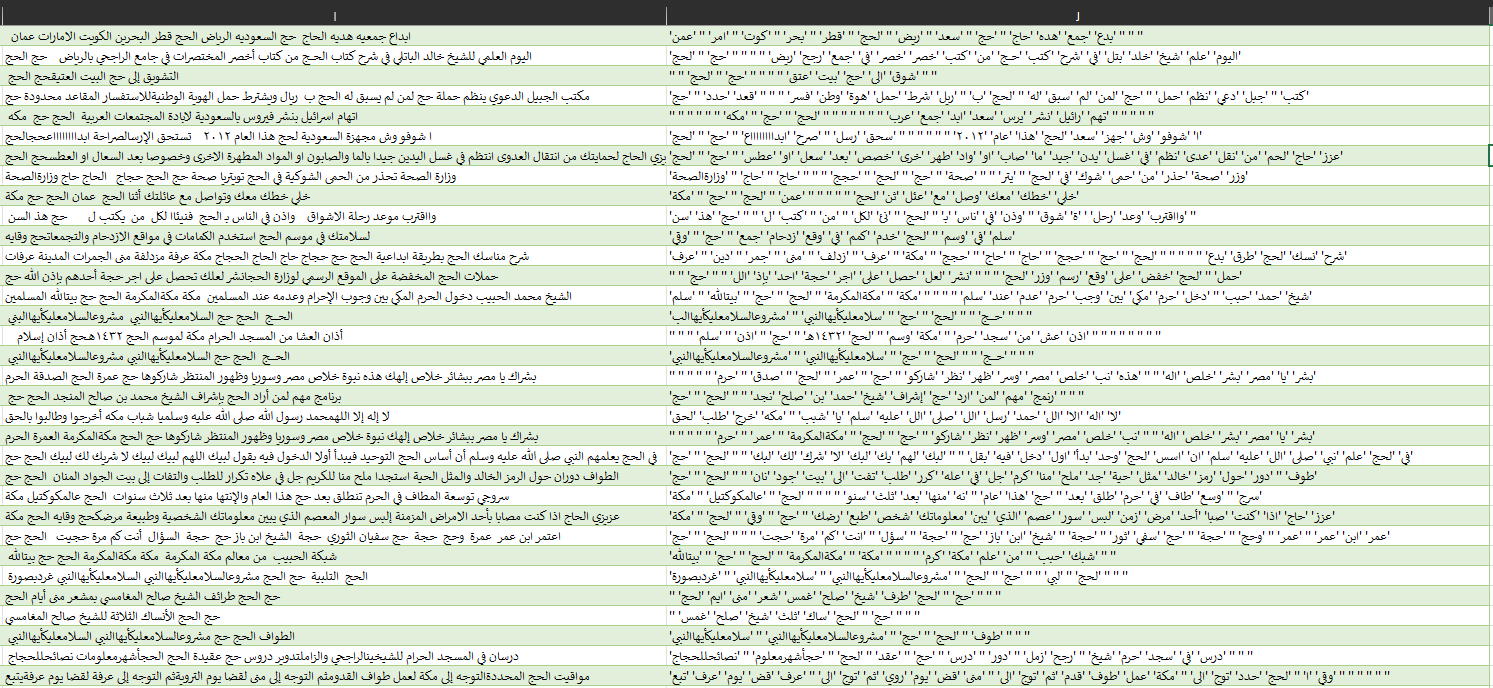


Function for TextArabic

Graphical user interface

Description automatically generated

**The Final Result:**

After data collection, sorting, organizing and getting rid of duplicate data our Stemmers Arabic column is ready for use and analysis

**Reference:**

* <https://learn.microsoft.com/en-us/power-query/add-custom-column>
* <https://www.exceldemy.com/convert-csv-to-xlsx/>
* <https://github.com/MaafiHanene/Arabic-Stemmers>
* <https://github.com/pandas-dev/pandas>
* <https://www.nltk.org/install.html>