Configuration Manual

MSc Research Project
News Article Analysis for Indian Election 2024

Utkarsh Singh
Student ID: x21199922

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Utkarsh Singh

Student ID: X21199922

Programme: MSc Data Analytics **Year:** 2023

Module: ……

Lecturer: ……

Submission Due Date: ……

Project Title: News Article Analysis for Indian Election 2024

Word Count: …… **Page Count:** ……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Utkarsh Singh
Student ID: x21199922

1. Table of Contents

2. System Configuration

Operating System:

- Windows, macOS, or Linux. The code seems to be written with Windows paths in mind (e.g., C:\Users\User\Downloads\chromedriver_win32\chromedriver.exe), but it can be adapted for other operating systems with minor changes.

Processor:

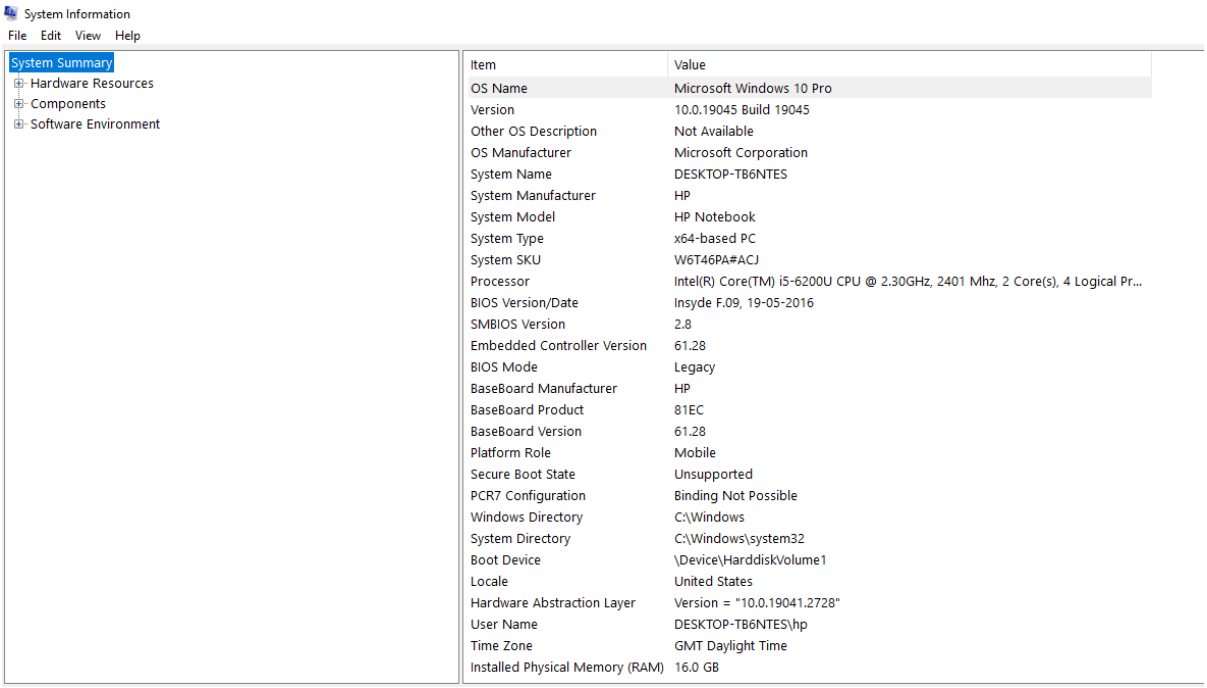
- Minimum: Dual-core CPU
- Recommended: Quad-core CPU or better

Memory (RAM):

- Minimum: 4GB
- Recommended: 8GB or more

Storage:

- Minimum: 10GB of free space (for software installations, temporary files, and data storage)
- Recommended: SSD with 20GB or more of free space for faster data processing.



Item	Value
OS Name	Microsoft Windows 10 Pro
Version	10.0.19045 Build 19045
Other OS Description	Not Available
OS Manufacturer	Microsoft Corporation
System Name	DESKTOP-TB6NTES
System Manufacturer	HP
System Model	HP Notebook
System Type	x64-based PC
System SKU	W6T46PA#ACJ
Processor	Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz, 2401 Mhz, 2 Core(s), 4 Logical Pr...
BIOS Version/Date	Insyde F.09, 19-05-2016
SMBIOS Version	2.8
Embedded Controller Version	61.28
BIOS Mode	Legacy
BaseBoard Manufacturer	HP
BaseBoard Product	81EC
BaseBoard Version	61.28
Platform Role	Mobile
Secure Boot State	Unsupported
PCR7 Configuration	Binding Not Possible
Windows Directory	C:\Windows
System Directory	C:\Windows\system32
Boot Device	\Device\HarddiskVolume1
Locale	United States
Hardware Abstraction Layer	Version = "10.0.19041.2728"
User Name	DESKTOP-TB6NTES\hp
Time Zone	GMT Daylight Time
Installed Physical Memory (RAM)	16.0 GB

3. Prerequisites:

- Python installed. Can be done using directions from below link https://youtu.be/Kn1HF3oD19c

- Libraries Required after the python installed and ran first code **bs4**, **selenium**, **newspaper**, **wordcloud**, **nltk**, **pandas**, **transformers**, **googletrans**, **langdetect**.

```
!pip install --upgrade tensorflow
import tensorflow as tf
from tensorflow.keras.models import Model
!pip install wordcloud
!pip install transformers
!pip install TensorFlow
!pip install --upgrade scipy --user
!pip install torch torchvision
```

```
from newspaper import Article
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from newspaper import Config
from wordcloud import WordCloud
from collections import Counter
import nltk
import pandas as pd
import transformers
from transformers import pipeline
nltk.download('stopwords')
nltk.download('punkt')
df_articles = pd.DataFrame()
sentiment_pipeline = pipeline("sentiment-analysis")

Title = []
SUMMARY = []
Keys = []
Sentiment = []

```

- Chrome browser installed.

Download the appropriate version of Chromedriver from the official site and place it in a known directory.
(https://sites.google.com/a/chromium.org/chromedriver/downloads)

4. Configuration

Update the path to the Chromedriver in the code

```
def is_social_media_or_wikipedia(url):
    return any(domain in url for domain in ['twitter', 'facebook', 'instagram', 'linkedin', 'pinterest', 'wikipedia'])

def get_google_search_links(keyword, n_pages):


    chrome_service = ChromeService(executable_path=r'C:\Users\User\Downloads\chromedriver_win32\chromedriver.exe')  # Replac
    chrome_options = webdriver.ChromeOptions()
    #chrome_options.add_argument('--headless')  # Run Chrome in headless mode to hide the browser window
    driver = webdriver.Chrome(service=chrome_service, options=chrome_options)

    try:

        # Scroll down the page to load more results

        query = keyword
        links = [] # Initiate empty list to capture final results
        # Specify number of pages on google search, each page contains 10 #links

        for page in range(1, n_pages):
            url = "http://www.google.com/search?q=" + query + "&start=" +      str((page - 1) * 10)
            driver.get(url)
            time.sleep(5)
            soup = BeautifulSoup(driver.page_source, 'html.parser')
            # soup = BeautifulSoup(r.text, 'html.parser')
```

5. Execution

1. For scripts with a main() function:

In the code, several sections have a main() function defined. This function serves as the entry point for the script.When you run the entire script, the code within the main() function will be executed.

This is because of the conditional statement if __name__ == "__main__":, which checks if the script is being run as the main program and not being imported elsewhere.

Upon execution, the main() function in the code often prompts the user for input.

```
    finally:
        driver.quit()
article_links = []
def main():
    keywords = input("Enter the keywords to search (comma-separated): ").split(',')
    n_pages = int(input("Enter the number of pages to scrape: "))

    for keyword in keywords:

        article_link = get_google_search_links(keyword, num_results)
        article_links.append(article_link)



    print(f"URL links from Google search results for '{keywords}':")
    for idx, link in enumerate(article_links, 1):
        print(f"{idx}. {link}")

if __name__ == "__main__":
    main()


#modi news article, bjp news article, indian national congress news article, rahul gandhi news article
```

6. Important Notes during execution

- The code contains multiple scripts that scrape Google search results, analyze articles, and visualize data.
- Ensure you have a stable internet connection when running the scripts.
- Web scraping scripts may break if the structure of the website changes. Regularly check and update the scraping logic if needed.
- Respect the **robots.txt** of websites and avoid making too many requests in a short period to prevent IP bans.
- Some scripts use the **googletrans** library for translation. This library uses the free Google Translate API, which has limits. If you encounter issues, consider using the paid API or another translation service.(My IP got banned for few instances when I was scrapping high amount 100 of articles at a time.)
- The sentiment analysis is done using the **transformers** library. Ensure you have the required models downloaded.

7. Visualization

The code contains scripts to generate word clouds and bar charts. Ensure you have the required libraries installed and run the appropriate functions to visualize the data.
Library → import matplotlib.pyplot as plt is used for bar charts

```
!pip install wordcloud
```

```
from newspaper import Article
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from newspaper import Config
from wordcloud import WordCloud
from collections import Counter
```

```
def create_wordcloud(text, keyword):
    wordcloud = WordCloud(width=800, height=400, background_color='white', colormap='viridis').generate(text)
    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.title(f"Word Cloud for Articles containing '{keyword}'")
    plt.show()
def main():
    global top_20_words

    keyword = input("Enter the keyword to search: ")
    num_results = int(input("Enter the number of article URLs to process: "))
```

8. Additional Configuration:

- If want to run Chrome in headless mode (without displaying the browser window), uncomment the line → chrome_options.add_argument('--headless')

- If you encounter pop-ups or other elements that interfere with scraping, you may need to update the handle_popups function or add additional logic to handle these elements.

References

Python Software Foundation (2019). 3.7.3 Documentation. [online] Python.org. Available at: https://docs.python.org/3/.

NLTK (2009). Natural Language Toolkit — NLTK 3.4.4 documentation. [online] Nltk.org. Available at: https://www.nltk.org/.

py-googletrans.readthedocs.io. (n.d.). *Googletrans: Free and Unlimited Google translate API for Python — Googletrans 3.0.0 documentation*. [online] Available at: https://py-googletrans.readthedocs.io/en/latest/.

huggingface.co. (n.d.). *Transformers — transformers 3.4.0 documentation*. [online] Available at: https://huggingface.co/transformers/.

sites.google.com. (n.d.). *ChromeDriver - WebDriver for Chrome*. [online] Available at: https://sites.google.com/a/chromium.org/chromedriver/

Richardson, L. (2019). *Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation*. [online] Crummy.com. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

Readthedocs.io. (2011). *Selenium with Python — Selenium Python Bindings 2 documentation*. [online] Available at: https://selenium-python.readthedocs.io/.