CMU-Africa

**Course: Programming for Data Analytics (04-638 A)**

**Title: Web Scraping and Sentiment Analysis**

**Submitted By:** Mustapha Olalekan Alaba and Alice Mugengano

**AndrewID:** malaba, amugenga

**Date:** November 1, 2024

Libraries used.

Objectives

To design and implement a web scraping project that extracts, processes, and performs sentiment analyses on articles about African countries from Wikipedia. The project aims to demonstrate proficiency in web scraping, data cleaning, natural language processing, visualisation and drawing insights from data.

# Sentiment analysis report on African countries article on Wikipedia

## Introduction

This report presents a comprehensive analysis of data collected from Wikipedia pages of African countries, offering insights into various aspects of their representation using sentiment analysis and selected themes. The study is based on a dataset that includes information about African nations, covering metrics such as word counts, frequent words, sentiment scores, and mentions of specific topics like war, poverty, and tourism.

## Steps

The following sections will delve into the specific findings, methodologies used, and findings from our analysis, providing an overview of the insights on African countries based on content from each country's Wikipedia pages.

## Web Scrapping and Data Preparation

Firstly, the list of African countries was used to dynamically visit Wikipedia.org using selenium to search for each country, selecting the first link in the result. Each resulting country web page was then saved as an HTML file. The webpages were further processed by removing all HTML tags, special characters, and extra spaces before saving cleaned aggregated text about each country into individual text documents. Finally, the aggregated text was preprocessed into a CSV file that contains each country article's number of paragraphs, frequent words, frequent word count, word count, polarity, subjectivity and other columns that focus on counts on thematic words such as war, poverty, corruption, insecurity, crime, art and tourism.

**Data Loading and Initial Exploration**

The extracted dataset from the Wikipedia pages was loaded from a CSV file containing information about African countries into a Pandas DataFrame. The data structure was previewed, and columns were checked for missing values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53 entries, 0 to 52
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country               53 non-null     object
 1   Number of Paragraphs  53 non-null     int64
 2   Frequent Words        53 non-null     object
 3   Frequent Word Count   53 non-null     object
 4   Word Count            53 non-null     int64
 5   Polarity              53 non-null     float64
 6   Subjectivity          53 non-null     float64
 7   war                   53 non-null     int64
 8   poverty               53 non-null     int64
 9   corruption            53 non-null     int64
 10  insecurity            53 non-null     int64
 11  crime                 53 non-null     int64
 12  art                   53 non-null     int64
 13  tourism               53 non-null     int64
dtypes: float64(2), int64(9), object(3)
memory usage: 5.9+ KB
```

There were no missing values, and all columns appeared to be in the correct data type and format.

**Descriptive Statistics**

The basic statistics for numerical columns, such as word count, polarity, and subjectivity, were calculated using pandas.

| | Number of Paragraphs | Word Count | Polarity | Subjectivity | war | poverty | corruption | insecurity | crime | art | tourism |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 53.000000 | 53.000000 | 53.000000 | 53.000000 | 53.000000 | 53.000000 | 53.000000 | 53.000000 | 53.000000 | 53.000000 | 53.000000 |
| mean | 117.452830 | 8938.830189 | 0.063132 | 0.325358 | 35.471698 | 3.981132 | 5.169811 | 0.452830 | 2.471698 | 123.811321 | 7.188679 |
| std | 38.181533 | 3290.066616 | 0.018841 | 0.051039 | 25.705884 | 3.968581 | 5.355572 | 1.600073 | 3.190152 | 38.624452 | 6.827705 |
| min | 1.000000 | 40.000000 | 0.000000 | 0.000000 | 2.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 2.000000 | 0.000000 |
| 25% | 88.000000 | 6298.000000 | 0.051000 | 0.315000 | 16.000000 | 1.000000 | 1.000000 | 0.000000 | 0.000000 | 98.000000 | 1.000000 |
| 50% | 116.000000 | 7939.000000 | 0.063000 | 0.333000 | 28.000000 | 2.000000 | 3.000000 | 0.000000 | 1.000000 | 123.000000 | 5.000000 |
| 75% | 148.000000 | 11755.000000 | 0.075000 | 0.350000 | 45.000000 | 6.000000 | 8.000000 | 0.000000 | 4.000000 | 146.000000 | 12.000000 |
| max | 221.000000 | 17904.000000 | 0.107000 | 0.370000 | 97.000000 | 15.000000 | 21.000000 | 11.000000 | 16.000000 | 237.000000 | 30.000000 |

From the statistics, the following insights can be deducted:

- **Average substantial content**: The African countries' wiki pages have an average paragraph count of 117 and an average word count of approximately 8,939, indicating substantial content length. Also, the word count with a standard deviation of 3,290 indicates a significant variation in text length.
- **Polarity and Subjectivity**: The mean polarity is 0.063, suggesting the overall sentiment is slightly positive with an average of 0.325 subjectivity; the texts tend to be more objective than subjective. A low standard deviation of 0.0188 for polarity suggests consistent sentiment across texts, while the standard deviation of 0.051 for subjectivity shows moderate variation in subjectivity levels.
- **Thematic summary**: On average, "war" appears frequently (mean = 35.47), while "poverty" is less common (mean = 3.98). Corruption and insecurity have moderate mentions, with means of 5.17 and 0.45, respectively.

It is also worth noting that the minimum word count is 40, signifying that there are some outliers; for the rest of the analysis, those outliers ( word count < 100 ) were removed from the dataset.

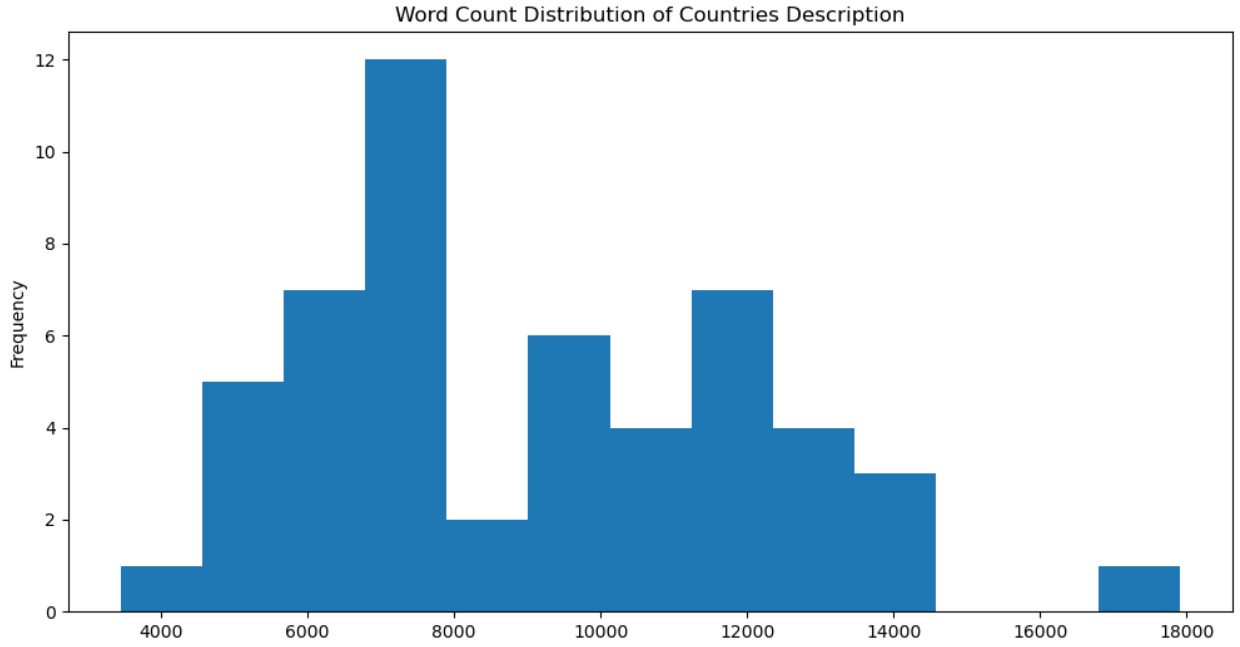
**Correlation Analysis**

A correlation analysis was done by finding the correlation between all the numerical columns in the dataset in tabular format before visualising it graphically for easy interpretation.


From the correlation matrix plot, the number of times “war” was mentioned moderately correlates (0.5) with the number of times “crime” was mentioned. Similarly, the mention of “war” has a higher moderate correlation (0.63) with the number of times “art” was mentioned in the countries' articles. A moderate correlation (0.63) was also noticed between the average polarity and subjectivity.

**Word Count Analysis**

The total word count for each country's wiki article was plotted with a histogram to visualise the words used for each page.

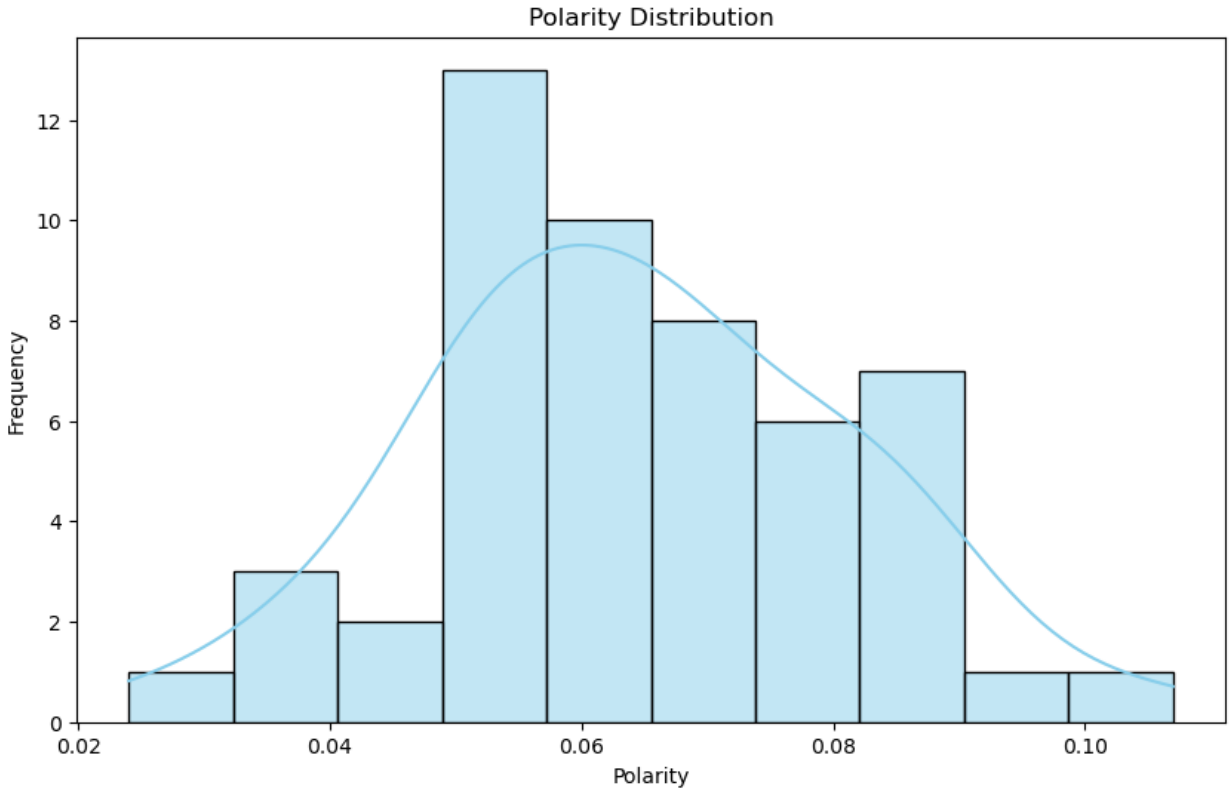


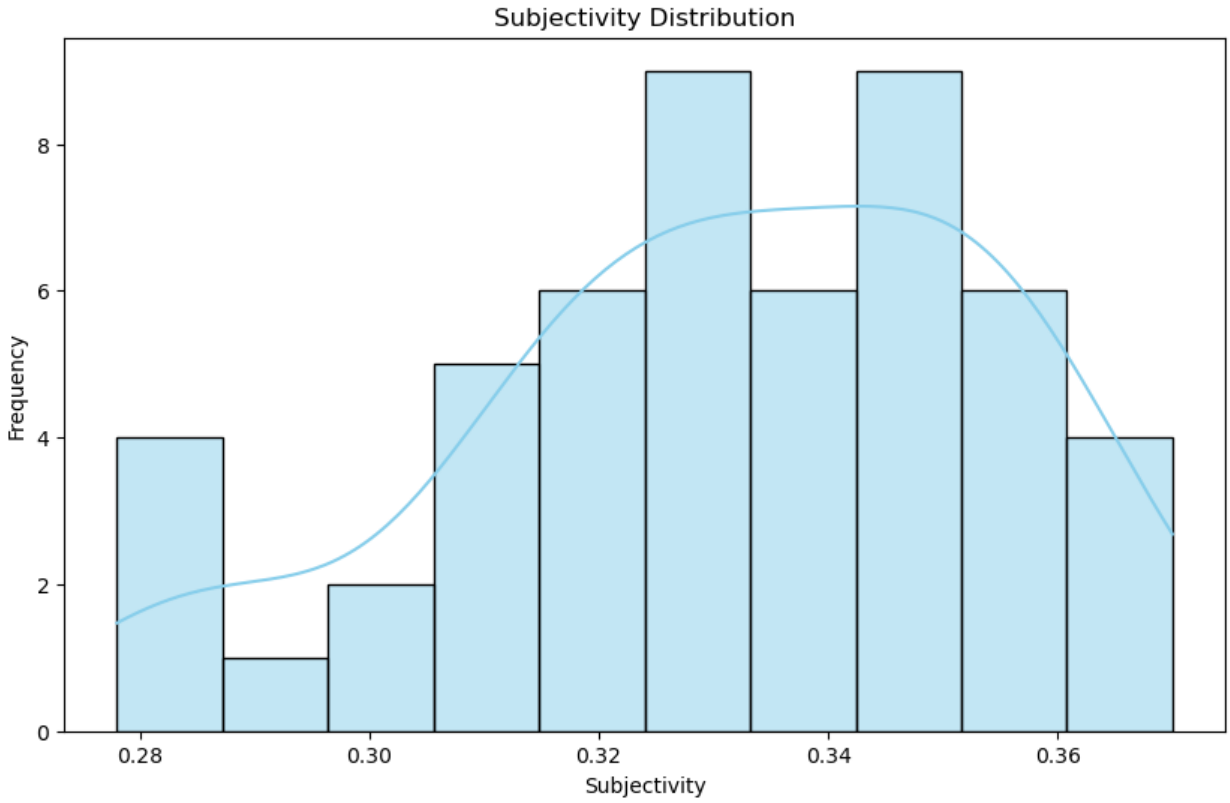

Most articles are between 6,500 and 7,800 words, with an outlier exceeding 17,000.

To further explore the most frequent words, each country's ten most frequent words were extracted and used to plot a WordCloud of the most frequent words across all countries' article

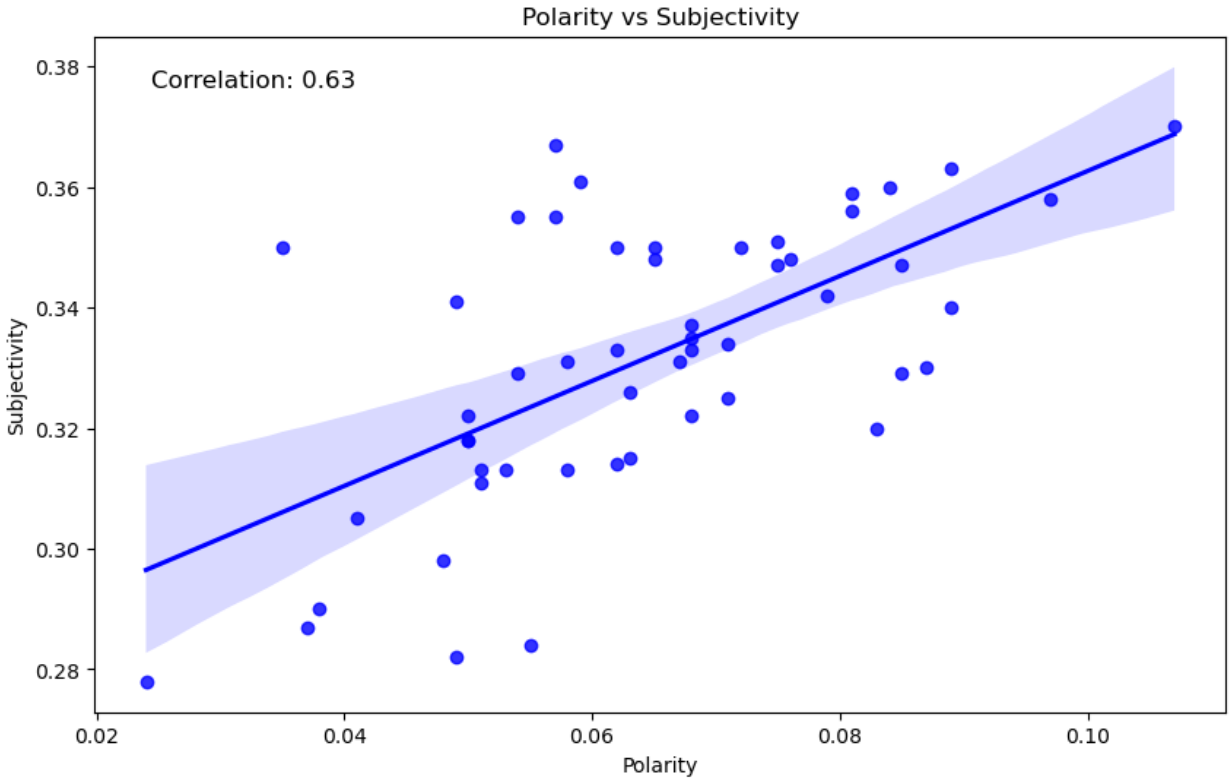
**Polarity and Subjectivity Analysis**

The average polarity and subjectivity for each country page were plotted on histogram to visualise the distribution; both indicate a normal distribution with subjectivity distribution being a bimodal distribution






The correlation between the average polarity and subjectivity for each country wiki page was also visualised using a scatter which further confirms the moderate correlation between the two variables.




**Thematic Analysis**

To analyse the prevalence of specific themes across various countries' Wikipedia pages, we evaluated word frequency analysis focusing on specific themes including war, poverty, corruption, and tourism. Subsequently, we identified and extracted the countries with the highest mentions for each of these selected themes.

Top 5 countries with the highest mention of war:

| Country | Mention |
|---|---|
| Sudan | 97 |
| Somalia | 90 |
| Libya | 86 |
| Ethiopia | 83 |
| Sierra Leone | 83 |

Top 5 countries with the highest mention of poverty:

| Country | Mention |
|---|---|
| Tanzania | 15 |
| Uganda | 15 |
| Chad | 14 |
| Nigeria | 13 |
| South Africa | 10 |

Top 5 countries with the highest mention of corruption:

| Country | Mention |
|---|---|
| Kenya | 21 |
| Congo (Kinshasa) | 18 |
| Liberia | 17 |
| Uganda | 17 |
| Cameroon | 16 |

Top 5 countries with the highest mention of insecurity:

| Country | Mention |
|---|---|
| Burkina Faso | 11 |
| Zimbabwe | 3 |
| Lesotho | 2 |
| Madagascar | 2 |
| Cameroon | 1 |

Top 5 countries with the highest mention of crime:

| Country | Mention |
|---|---|
| South Africa | 16 |
| Sudan | 10 |
| Congo (Kinshasa) | 8 |
| Liberia | 8 |
| Nigeria | 8 |

Top 5 countries with the highest mention of art:

| Country | Mention |
|---|---|
| South Africa | 237 |
| Tunisia | 188 |
| Sierra Leone | 187 |
| Tanzania | 182 |
| Sudan | 173 |

Top 5 countries with the highest mention of tourism:

| Country | Mention |
|---|---|
| Namibia | 30 |
| Egypt | 22 |
| Seychelles | 21 |
| Zimbabwe | 19 |
| Morocco | 16 |

As outlined in the table, most of the themes and the countries correlate with the real events in those countries.

**Conclusion**

This analysis provides insights into how African countries are represented on Wikipedia, highlighting patterns in sentiment, content length, and thematic focus. The findings can be useful for understanding biases, information gaps, and prevalent narratives in online encyclopedic content about African nations.