Task 8HD : Data Cleansing and Text Analysis Challenge

Truong Khang Thinh Nguyen

2024-05-22

Table of contents

Truong Khang Thinh Nguyen - 223446545

Email : s223446545@gmail.com

SIT220 - Undergraduate

This report endeavors to delve deeper into the History Stack Exchange site, aiming to extract valuable patterns, trends, and insights about its users through the utilization of regular expressions. While traditional exploratory data analysis (EDA) provides an overview of the data, the use of regular expressions enables us to unexplored details and relationships within the text data. For instance, by employing regular expressions, we can extract specific information such as locations mentioned in the posts, allowing us to visualize user activity geographically through mapping. Beyond merely skimming the surface of the data, this approach empowers us to uncover hidden gems of knowledge and gain a more comprehensive understanding of user interactions and interests on the History Stack Exchange platform.

Moreover, this report goes beyond basic exploratory analysis by delving into the relationship between sentiment expressed in comments and various factors such as post length, sentiment of the post title and body, and other contextual attributes. By leveraging regular expressions to extract sentiment-related information from the text data, we can discern patterns and correlations that may otherwise remain unnoticed. For example, we may discover that longer posts tend to elicit more diverse sentiments in the comments, or that the sentiment expressed in the post title strongly influences the sentiment of subsequent comments. This deeper analysis not only enriches our understanding of user behavior and engagement on the platform but also provides valuable insights for content creators, moderators, and platform administrators to enhance user experience and foster community engagement.

```
# Packages for csv and XML files manipulation
import xml.etree.ElementTree as ET
import csv

# Data Manipulation packages
import pandas as pd
import numpy as np
import re

# Packages for plotting
import folium
from folium.plugins import MarkerCluster
import seaborn as sns
import matplotlib.pyplot as plt
from wordcloud import WordCloud

# Packages for building a regresison model
from sklearn.preprocessing import StandardScaler
import statsmodels.api as sm
from statsmodels.graphics.gofplots import qqplot

# Import neccessary packages to analyse the sentiments and the polarity
 ↪  scores of the text
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
from scipy.special import softmax
```

```
# Create a function to parse the XML file and then create a corresponded csv
 ↪  file
def xml_to_csv(input_file, output_file):
    # Parse the XML file
    tree = ET.parse(input_file)
    root = tree.getroot()

    # Open the CSV file for writing
    with open(output_file, 'w', newline='', encoding='utf-8') as csvfile:
        # Create a CSV writer object
        csvwriter = csv.writer(csvfile)

        # Extract all column names from the first element (assuming all
         ↪  elements have the same structure)
        first_element = root[0]
        column_names = first_element.attrib.keys()
        csvwriter.writerow(column_names)  # Write the header

        # Write data rows
        for element in root:
            row = [element.attrib.get(col, '') for col in column_names]
            csvwriter.writerow(row)

# Use the defined function to pass the XML file and the wanted csv file name
xml_to_csv('Badges.xml', 'Badges.csv')
xml_to_csv('Comments.xml', 'Comments.csv')
xml_to_csv('PostHistory.xml', 'PostHistory.csv')
xml_to_csv('PostLinks.xml', 'PostLinks.csv')
xml_to_csv('Posts.xml', 'Posts.csv')
xml_to_csv('Tags.xml', 'Tags.csv')
xml_to_csv('Users.xml', 'Users.csv')
xml_to_csv('Votes.xml', 'Votes.csv')
```

```
# Load datasets from the created csv files
Badges = pd.read_csv("Badges.csv")
Cmt = pd.read_csv("Comments.csv")
PostHistory = pd.read_csv("PostHistory.csv")
PostLink = pd.read_csv("PostLinks.csv")
Post = pd.read_csv("Posts.csv")
Tags = pd.read_csv("Tags.csv")
Users = pd.read_csv("Users.csv")
```

```
Vote = pd.read_csv("Votes.csv")
```

To kick off our analysis, we first need to perform an Exploratory Data Analysis (EDA) to get a broad understanding of the dataset's characteristics.

Initially, we should generate copies of each dataset we've just created. This way, any modifications made to the duplicates won't impact the original datasets.

```
badges_df = Badges
cmt_df = Cmt
posthis_df = PostHistory
post_df = Post
tag_df = Tags
user_df = Users
vote_df = Vote
```

As we are currently in Australia so I'll compile historical events by summarizing the top 10 posts with the highest views from Stack Exchange data.

```
# Function to remove tags from the text
def remove_html_tags(text):
    # Define the regular expression pattern to match HTML tags
    html_pattern = re.compile(r'<.*?>')

    # Use re.sub() to replace HTML tags with an empty string
    clean_text = re.sub(html_pattern, '', text)

    return clean_text
```

```
# Function to extract the capitalized words from titles ==> Get Keywords
def extract_capitalized_words(title):
    # Use regex to find capitalized words
    capitalized_words = re.findall(r'\b[A-Z][a-z]*\b', title)
    return ' '.join(capitalized_words)
```

Initially, we'll eliminate rows that don't have any tags.

```
post_df.dropna(subset = "Tags",inplace = True)
```

Then filter out rows that just have the “australian”

```
# Filter rows where the "tags" column contains "Australia"
aus_df = post_df[post_df["Tags"].str.contains(r"\baustralia\b")]
print(aus_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 59 entries, 65 to 43635
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Id                59 non-null     int64
 1   PostTypeId        59 non-null     int64
 2   CreationDate      59 non-null     object
 3   Score             59 non-null     int64
 4   ViewCount         59 non-null     float64
 5   Body              59 non-null     object
 6   OwnerUserId       51 non-null     float64
 7   LastEditorUserId  41 non-null     float64
 8   LastEditDate      43 non-null     object
 9   LastActivityDate  59 non-null     object
 10  Title             59 non-null     object
 11  Tags              59 non-null     object
 12  AnswerCount       59 non-null     float64
 13  CommentCount      59 non-null     int64
 14  ContentLicense    59 non-null     object
dtypes: float64(4), int64(4), object(7)
memory usage: 7.4+ KB
None
```

```
# Let's have a look at our dataframe that just has Aus tag
print(aus_df)
```

```
          Id  PostTypeId             CreationDate  Score  ViewCount  \
65        73           1  2011-10-11T22:18:54.920      9      633.0
76        85           1  2011-10-11T23:24:12.227     14      903.0
252      273           1  2011-10-13T21:55:31.070      4     2548.0
257      279           1  2011-10-13T22:29:48.187     13      491.0
```

304	328	1	2011-10-14T21:42:48.133	5	1063.0
1011	1120	1	2012-01-05T21:32:19.320	2	629.0
1458	1652	1	2012-03-26T00:00:18.023	5	154.0
1575	1785	1	2012-04-17T17:14:53.843	22	109061.0
1844	2073	1	2012-05-16T10:44:12.327	5	360.0
2305	2575	1	2012-07-12T17:19:08.040	26	32613.0
2349	2629	1	2012-07-18T16:45:42.340	19	2991.0
3000	4381	1	2012-10-22T11:40:59.687	25	11028.0
6536	10649	1	2013-10-28T09:24:27.717	9	1109.0
6545	10664	1	2013-10-29T17:32:06.570	2	227.0
7205	11526	1	2014-01-27T21:25:21.073	10	2765.0
7415	11784	1	2014-02-22T05:22:52.353	5	5402.0
7475	11855	1	2014-02-26T22:03:31.113	6	408.0
7577	11981	1	2014-03-08T03:44:39.397	0	3795.0
7912	12392	1	2014-04-09T09:18:03.893	3	422.0
8006	12496	1	2014-04-15T01:06:41.733	6	473.0
9807	16773	1	2014-10-23T22:42:44.587	15	3716.0
10141	17224	1	2014-11-22T01:51:28.670	7	328.0
10932	19256	1	2015-02-07T11:43:17.103	25	35350.0
12113	22801	1	2015-05-18T01:17:48.323	10	3518.0
14553	26967	1	2015-12-27T08:49:34.237	2	432.0
15714	28456	1	2016-04-25T10:53:38.920	45	25830.0
15865	28650	1	2016-05-07T05:24:54.550	3	620.0
15893	28685	1	2016-05-09T21:32:06.710	3	892.0
16526	30525	1	2016-06-25T04:19:08.227	0	96.0
16793	30894	1	2016-07-15T10:53:52.443	3	232.0
19672	35586	1	2017-02-18T16:47:09.013	11	7004.0
19747	35676	1	2017-02-24T18:53:56.500	7	2861.0
21199	38495	1	2017-06-09T05:48:00.517	4	371.0
21315	38644	1	2017-06-21T01:40:13.490	4	307.0
21608	38997	1	2017-07-14T02:53:06.253	6	356.0
22566	40140	1	2017-09-10T19:11:42.920	6	245.0
22924	40588	1	2017-09-29T19:10:48.053	7	285.0
23388	41148	1	2017-10-21T00:56:03.233	4	785.0
23586	41408	1	2017-11-02T00:22:35.180	0	194.0
26255	45869	1	2018-04-28T04:55:44.380	4	313.0
27358	47256	1	2018-07-21T07:26:17.063	1	149.0
28289	48430	1	2018-09-29T22:07:38.357	1	236.0
28595	48829	1	2018-10-20T04:36:59.253	3	407.0
29916	50501	1	2019-01-07T09:54:32.467	1	225.0
30880	51839	1	2019-03-28T03:55:49.050	2	224.0
31646	52822	1	2019-05-26T02:14:19.963	4	706.0
31918	53158	1	2019-06-12T01:25:30.720	1	1295.0

32301	53641	1	2019-07-13T08:53:49.177	0	131.0
32304	53644	1	2019-07-13T10:06:35.343	5	289.0
32357	53714	1	2019-07-17T03:55:58.390	2	164.0
33259	54976	1	2019-10-11T14:23:10.420	46	15301.0
36035	59826	1	2020-06-15T14:24:25.217	3	335.0
36158	59995	1	2020-06-27T12:14:10.787	1	131.0
36387	60312	1	2020-07-16T05:41:32.667	2	154.0
37295	61742	1	2020-11-03T04:32:03.310	16	2294.0
38728	63856	1	2021-05-08T17:40:40.427	4	781.0
39585	66028	1	2021-09-10T19:55:42.740	18	1179.0
41500	69756	1	2022-09-04T16:01:26.637	2	230.0
43635	72881	1	2023-11-29T13:05:48.433	3	304.0

	Body	OwnerUserId	\
65	Why were many Aboriginal names kept for tow...	73.0	
76	In my hometown (Eastwood NSW, Australia) we...	73.0	
252	This wikipedia article on <a href="http://e...	73.0	
257	There are two that I am aware of: \n\n<u...	73.0	
304	Frank Welsh's mammoth history of Australia,...	106.0	
1011	A simple question that is not so simple to ...	NaN	
1458	The <a href="http://en.wikipedia.org/wiki/P...	49.0	
1575	I asked this question in various places on ...	NaN	
1844	Anyone who listens to Question Time in Aust...	49.0	
2305	My Australian friend says yes. I'm not so s...	282.0	
2349	I recently read that the Australian Aborigi...	1110.0	
3000	Australia hosted aboriginal populations sin...	1436.0	
6536	The Wikipedia article about the <a href="ht...	2830.0	
6545	The English were hostile to most of the nat...	2684.0	
7205	In 1939, Australia had not ratified the 193...	876.0	
7415	The common explanation for the usage of Aus...	3682.0	
7475	<a href="http://en.wikipedia.org/wiki/Marga...	3850.0	
7577	In Australia's History, the Europeans tried...	NaN	
7912	After reviewing a source about American-Aus...	3810.0	
8006	The <a href="http://www8.austlii.edu.au/cgi...	49.0	
9807	The gold rush in Australia saw many <a href...	49.0	
10141	I am trying to identify the ship in this pi...	8521.0	
10932	The aborigines are believed to have migrate...	8451.0	
12113	While there are many articles discussing th...	4545.0	
14553	Being European, I know history of World War...	15803.0	
15714	Leaving aside the naval battles of the Paci...	10028.0	
15865	I have been wondering for a while and have ...	17481.0	
15893	In the spring of 1942, there was the danger...	120.0	
16526	Under the Australian Constitution, who will...	NaN	

```
16793  <p>Australian Federal Politics has been tumult...        454.0
19672  <p>Is there any evidence of ancient Chinese ex...      23474.0
19747  <p>Asians were trading with each other. Indian...      23798.0
21199  <p>According to H.H. Bancroft's "History of Ca...      18968.0
21315  <p>The Wikipedia article about <a href="https:...      25303.0
21608  <p>I have found the below picture of my ancest...      25859.0
22566  <p>Some time ago, I read a story on the intern...      26877.0
22924  <p><a href="https://en.wikipedia.org/wiki/Euro...        251.0
23388  <p>I am restoring a Chevy truck from 1942. The...      27531.0
23586  <p>Every year on the first tuesday of novememb...      26410.0
26255  <p>When the first penal colony was founded in ...      31522.0
27358  <p>I found this attached to a rock, on a beach...      32721.0
28289  <p>I was told by a guide that when European se...      31876.0
28595  <p>"White Australians" in the 19th and early 2...        876.0
29916  <p>How motivated were the Australian, Canadian...       4853.0
30880  <p>Can someone identify this badge please. </p...      37258.0
31646  <p>Australia was a penal colony, so since ther...      31722.0
31918  <p>I'm reading a book about the first world wa...       1455.0
32301  <p>On 1939-09-03, the UK declared war on Germa...          NaN
32304  <p>I'm looking at a dataset claims that Austra...          NaN
32357  <p>I once read that an aboriginal man from Aus...        994.0
33259  <p>From what I've gathered at a glance, Aborig...      40601.0
36035  <p>When <a href="https://en.wikipedia.org/wiki...          NaN
36158  <ol>\n<li><p>What was life like for Australian...          NaN
36387  <p>During the Australian constitutional crisis...      34547.0
37295  <p>Did Aboriginal Australians know slings?</p>...      25755.0
38728  <p>What kind of religions did ancient aborigin...      48231.0
39585  <p>It has been rather well-established that (s...      52220.0
41500  <p>I'm Australian. From a few years after our ...      58939.0
43635  <p>This text was written in a <a href="https:/...      63339.0

       LastEditorUserId             LastEditDate         LastActivityDate  \
65                 16.0  2011-10-13T20:52:02.737  2011-12-06T01:29:53.427
76                 85.0  2012-11-20T16:18:53.453  2023-08-24T20:53:24.863
252                16.0  2011-10-13T22:24:54.863  2018-07-26T05:10:33.223
257                16.0  2011-10-14T16:45:39.860  2017-12-04T22:56:17.030
304               106.0  2013-07-05T22:01:55.030  2019-11-20T21:05:12.523
1011                NaN                      NaN  2012-02-02T20:00:54.850
1458              102.0  2012-03-26T13:18:53.657  2012-05-14T03:49:00.580
1575               -1.0  2015-07-06T09:35:26.337  2018-05-28T10:09:10.750
1844                NaN                      NaN  2012-08-30T02:22:41.437
2305             3810.0  2014-03-15T11:38:12.450  2017-02-09T15:05:49.873
2349            26454.0  2018-03-22T09:31:40.750  2018-03-22T09:31:40.750
```

3000	1110.0	2012-11-03T19:02:36.883	2021-01-28T09:14:11.107
6536	2830.0	2013-10-28T10:58:30.053	2013-10-28T10:58:30.053
6545	48903.0	2021-03-10T15:59:05.737	2021-03-10T15:59:05.737
7205	26454.0	2018-03-15T22:42:21.003	2018-05-10T05:01:26.330
7415	3682.0	2014-03-06T14:55:04.387	2014-03-06T15:09:47.163
7475	NaN	NaN	2014-04-14T06:07:40.660
7577	NaN	NaN	2014-03-09T22:05:16.410
7912	NaN	NaN	2014-04-09T11:24:48.943
8006	24858.0	2019-04-28T13:50:15.033	2019-04-29T15:06:58.863
9807	NaN	NaN	2018-11-20T04:39:13.560
10141	3871.0	2014-11-22T13:37:49.240	2014-12-31T02:18:40.430
10932	8451.0	2016-03-12T14:37:21.250	2018-03-28T00:22:32.030
12113	4545.0	2015-05-19T15:14:05.960	2024-03-24T04:57:42.677
14553	NaN	NaN	2015-12-27T08:49:34.237
15714	4390.0	2016-04-26T21:25:31.293	2024-02-03T05:01:41.707
15865	11674.0	2016-05-09T09:36:10.043	2016-05-09T09:36:10.043
15893	NaN	NaN	2016-05-10T04:34:19.660
16526	NaN	NaN	2016-06-29T05:00:56.890
16793	NaN	NaN	2016-07-15T14:32:53.720
19672	56412.0	2022-10-06T16:37:28.960	2023-11-24T00:21:38.513
19747	4545.0	2017-02-24T20:41:34.753	2020-06-03T08:34:38.970
21199	18968.0	2018-04-04T18:36:07.950	2018-07-13T14:34:07.117
21315	NaN	NaN	2017-10-18T14:54:47.150
21608	38309.0	2019-07-23T09:02:42.400	2019-07-23T09:02:42.400
22566	1401.0	2017-09-11T18:31:41.427	2022-10-06T04:38:47.617
22924	771.0	2020-06-28T03:53:24.150	2020-06-28T03:53:24.150
23388	24858.0	2017-10-21T20:07:45.093	2017-10-22T12:35:41.440
23586	NaN	NaN	2017-11-02T01:05:49.883
26255	31522.0	2018-04-28T05:09:15.670	2018-06-03T04:22:09.833
27358	1401.0	2018-07-21T12:12:37.357	2018-07-21T12:12:37.357
28289	31876.0	2018-09-30T01:01:55.567	2018-09-30T04:10:06.583
28595	33089.0	2018-10-20T13:22:04.513	2018-10-21T11:52:11.533
29916	24858.0	2019-01-07T09:57:29.667	2019-01-07T09:57:29.667
30880	24858.0	2019-03-28T04:13:11.453	2019-03-28T04:13:11.453
31646	NaN	NaN	2019-05-26T14:22:12.520
31918	NaN	NaN	2021-03-30T19:17:28.517
32301	NaN	2019-07-13T09:01:48.593	2019-07-13T09:25:02.067
32304	NaN	2019-07-17T00:53:13.087	2019-07-17T16:06:17.837
32357	NaN	NaN	2019-07-19T15:43:29.320
33259	40601.0	2019-10-14T06:39:49.547	2024-03-22T11:08:29.800
36035	40598.0	2020-06-17T12:54:23.303	2020-06-17T12:54:23.303
36158	1401.0	2020-06-27T16:24:58.063	2020-06-27T18:05:11.620
36387	6489.0	2020-07-16T05:54:36.993	2020-07-16T09:50:52.817

```
37295            771.0  2020-11-03T05:01:29.677  2020-11-03T21:58:51.403
38728          48231.0  2021-05-09T09:29:31.663  2021-05-09T22:06:33.823
39585          32206.0  2021-09-12T09:36:32.677  2021-09-16T10:33:46.697
41500              NaN                      NaN  2022-09-07T16:31:01.733
43635           1401.0  2023-11-29T13:23:37.997  2023-11-29T13:23:37.997

                                                  Title  \
65       Australian towns / cities with Aboriginal names
76     Was the Granny Smith Apple the first green ski...
252    Why did Arthur Philip decide to move to Port J...
257    How many recorded incidents are there of attac...
304    How was Australia able to start to demobilize ...
1011   How did the early settlers of Australia settle...
1458   What caused the decline in support for the Pro...
1575   Why did Canada, Australia and New Zealand sepa...
1844   Why did Dorothy Dixers become prevalent in que...
2305   Can the Queen of England fire the prime minist...
2349   Why did civilisation/city states never take ro...
3000   Why did Austronesian/Polynesian people not col...
6536   Did France actually intend to colonise West Au...
6545   Were there any major battles between Australia...
7205   When did Australia declare war on Germany in WWII
7415           America vs. Australia as a "Penal Colony"
7475    Was the penal system used to colonise Australia?
7577   Did any other countries try to breed a race ou...
7912   Was a direct telegraph line between America an...
8006   Execution at the scene of the crime under Aust...
9807   Which fruits and vegetables did Chinese migran...
10141               Identifying a ship in Sydney Harbour
10932  Did the aborigines of Australia and the Maoris...
12113       Why weren't Australian Aborigines enslaved?
14553           Was Australia ever attacked during WW2?
15714  Why was the Japanese Army's fatalities inflict...
15865                Australian Immigration 1945-present
15893  Was there a viable alternate supply route to A...
16526                           Australian Head of State
16793  This is the second time Australia has had 6 PM...
19672  Is there evidence of ancient Chinese explorati...
19747              Why didn't Asians discover Australia?
21199     Convict stowaways crossing the Pacific in 1796
21315  Why were penal colony members from Australia m...
21608  What possible uniform with decoration is my an...
22566  Who was the 18th century French eccentric who ...
```

```
22924  Which personal weapons, if any, were carried b...
23388  When were 1942 lend-lease trucks sold to the p...
23586  Why do people in Melbourne Victoria Australia ...
26255  When did keeping convicts in Australia become ...
27358                Is this a coin, token or medallion?
28289  Did Indigenous Australians burn land to get Eu...
28595  When did the accusations of convict stain agai...
29916  How motivated were the Australian, Canadian an...
30880                 Help to identify WW1 or WWII badge
31646  How did early colonial Australia deal with its...
31918  In World War 1, why were the Australian and Ca...
32301  Did any UK declaration of war extend automatic...
32304  Why might one think that Australia gained inde...
32357  Has anyone written Australian history from the...
33259  Why didn't Aboriginal Australians invent agric...
36035  Why didn't Indonesians/Indian Ocean traders br...
36158  What was life like for Australian convicts in ...
36387    Why didn't Sir John Kerr just call an election?
37295             Did Aboriginal Australians know slings?
38728  What kind of religions did ancient aboriginals...
39585  Why wasn't awareness of Australia more widespr...
41500  Why Oppress a Negligible Population (Stolen Ge...
43635  Factcheck: Did Australia request 200.000 17 ye...

                                                    Tags  AnswerCount  \
65                                |australia|aboriginals|          1.0
76                                      |australia|food|          2.0
252                  |australia|18th-century|settlement|          1.0
257                               |war|australia|defense|          4.0
304                   |world-war-two|military|australia|          2.0
1011                              |australia|aboriginals|          1.0
1458                        |government|australia|election|          1.0
1575  |british-empire|united-kingdom|ireland|austral...          9.0
1844                         |australia|political-history|          1.0
2305                                  |australia|monarchy|          4.0
2349  |native-americans|australia|civilizations|nort...          1.0
3000            |colonization|australia|south-east-asia|          8.0
6536              |colonization|australia|french-empire|          1.0
6545         |colonization|australia|battle|aboriginals|          0.0
7205  |world-war-two|germany|australia|diplomatic-hi...          2.0
7415                     |united-states|australia|colony|          1.0
7475                                          |australia|          3.0
7577                              |australia|aboriginals|          1.0
```

```
7912                          |united-states|australia|          1.0
8006                        |law|crime|racism|australia|          1.0
9807                        |food|agriculture|australia|          1.0
10141                  |20th-century|australia|nautical|          1.0
10932                |australia|aboriginals|new-zealand|          4.0
12113                               |slavery|australia|          5.0
14553                         |world-war-two|australia|          0.0
15714      |united-states|world-war-two|japan|australia|          7.0
15865               |20th-century|australia|immigration|          1.0
15893                         |world-war-two|australia|          1.0
16526                       |britain|monarchy|australia|          1.0
16793                           |20th-century|australia|          1.0
19672                            |australia|exploration|          1.0
19747                   |asia|australia|south-east-asia|          2.0
21199  |18th-century|california|spanish-empire|age-of...          2.0
21315                   |united-states|australia|colony|          2.0
21608  |20th-century|19th-century|identification|unif...          1.0
22566                     |18th-century|australia|dutch|          1.0
22924       |weapons|17th-century|australia|exploration|          2.0
23388                         |world-war-two|australia|          1.0
23586                                |australia|sports|          1.0
26255                                       |australia|          1.0
27358                                 |naval|australia|          0.0
28289  |18th-century|colonization|australia|settlemen...          1.0
28595                    |british-empire|australia|race|          1.0
29916  |world-war-two|world-war-one|canada|australia|...          0.0
30880           |world-war-two|identification|australia|          1.0
31646                  |british-empire|australia|gender|          1.0
31918                   |world-war-one|canada|australia|          4.0
32301  |war|law|united-kingdom|international-relation...          1.0
32304  |war|law|united-kingdom|international-relation...          1.0
32357  |political-history|cultural-history|social-his...          1.0
33259                                       |australia|          8.0
36035                       |trade|australia|indonesia|          2.0
36158  |19th-century|slavery|transportation|australia...          1.0
36387   |political-history|monarchy|australia|democracy|          1.0
37295                    |weapons|australia|aboriginals|          1.0
38728                   |religion|australia|aboriginals|          1.0
39585   |china|australia|indonesia|precolumbian-contact|          5.0
41500                                 |racism|australia|          1.0
43635              |world-war-two|immigration|australia|          0.0

       CommentCount ContentLicense
```

65	0	CC BY-SA 3.0
76	11	CC BY-SA 3.0
252	1	CC BY-SA 3.0
257	2	CC BY-SA 3.0
304	1	CC BY-SA 3.0
1011	3	CC BY-SA 3.0
1458	0	CC BY-SA 3.0
1575	6	CC BY-SA 3.0
1844	4	CC BY-SA 3.0
2305	3	CC BY-SA 3.0
2349	8	CC BY-SA 3.0
3000	11	CC BY-SA 3.0
6536	2	CC BY-SA 3.0
6545	8	CC BY-SA 4.0
7205	0	CC BY-SA 3.0
7415	1	CC BY-SA 3.0
7475	2	CC BY-SA 3.0
7577	13	CC BY-SA 3.0
7912	2	CC BY-SA 3.0
8006	3	CC BY-SA 4.0
9807	4	CC BY-SA 3.0
10141	1	CC BY-SA 3.0
10932	4	CC BY-SA 3.0
12113	9	CC BY-SA 3.0
14553	7	CC BY-SA 3.0
15714	12	CC BY-SA 3.0
15865	0	CC BY-SA 3.0
15893	5	CC BY-SA 3.0
16526	4	CC BY-SA 3.0
16793	7	CC BY-SA 3.0
19672	3	CC BY-SA 4.0
19747	2	CC BY-SA 3.0
21199	7	CC BY-SA 3.0
21315	0	CC BY-SA 3.0
21608	4	CC BY-SA 4.0
22566	5	CC BY-SA 3.0
22924	5	CC BY-SA 4.0
23388	2	CC BY-SA 3.0
23586	3	CC BY-SA 3.0
26255	5	CC BY-SA 3.0
27358	8	CC BY-SA 4.0
28289	2	CC BY-SA 4.0
28595	5	CC BY-SA 4.0

```
29916             6   CC BY-SA 4.0
30880             0   CC BY-SA 4.0
31646             4   CC BY-SA 4.0
31918            14   CC BY-SA 4.0
32301             0   CC BY-SA 4.0
32304             7   CC BY-SA 4.0
32357             7   CC BY-SA 4.0
33259             6   CC BY-SA 4.0
36035            10   CC BY-SA 4.0
36158             2   CC BY-SA 4.0
36387             2   CC BY-SA 4.0
37295             4   CC BY-SA 4.0
38728             5   CC BY-SA 4.0
39585             1   CC BY-SA 4.0
41500             3   CC BY-SA 4.0
43635            24   CC BY-SA 4.0
```

To effectively identify historical events within each post, we'll analyze both the title and the body. Currently, the bodies of each post are in HTML format, so we'll need to reformat them for better readability.

```
aus_df["Clean Body"] = aus_df["Body"].apply(remove_html_tags)
print(aus_df[["Clean Body"]].head())
```

```
                                            Clean Body
65   Why were many Aboriginal names kept for town a...
76   In my hometown (Eastwood NSW, Australia) we an...
252  This wikipedia article on Botany Bay suggests ...
257  There are two that I am aware of:\n\n\nJapanes...
304  Frank Welsh's mammoth history of Australia, Gr...

C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\3680200629.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

Next, we'll extract keywords from the titles to identify significant events. Typically, keywords related to major events will be capitalized.

```
aus_df["Key Words"] = aus_df["Title"].apply(extract_capitalized_words)
print(aus_df[["Title","Key Words"]].head())
```

```
                                                Title  \
65     Australian towns / cities with Aboriginal names
76   Was the Granny Smith Apple the first green ski...
252  Why did Arthur Philip decide to move to Port J...
257  How many recorded incidents are there of attac...
304  How was Australia able to start to demobilize ...

                                  Key Words
65                      Australian Aboriginal
76                     Was Granny Smith Apple
252  Why Arthur Philip Port Jackson Botany Bay
257                            How Australian
304                             How Australia

C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\3774467798.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

Following that, we'll select the top 10 posts with the highest views which may contain significant events in Australian history.

```
aus_df.sort_values(by = "ViewCount", ascending = False,inplace = True)
top10_aus = aus_df.head(10)
print(top10_aus)
```

```
          Id  PostTypeId            CreationDate  Score  ViewCount  \
1575    1785           1  2012-04-17T17:14:53.843     22   109061.0
10932  19256           1  2015-02-07T11:43:17.103     25    35350.0
2305    2575           1  2012-07-12T17:19:08.040     26    32613.0
15714  28456           1  2016-04-25T10:53:38.920     45    25830.0
33259  54976           1  2019-10-11T14:23:10.420     46    15301.0
3000    4381           1  2012-10-22T11:40:59.687     25    11028.0
19672  35586           1  2017-02-18T16:47:09.013     11     7004.0
```

```
7415   11784           1  2014-02-22T05:22:52.353      5     5402.0
7577   11981           1  2014-03-08T03:44:39.397      0     3795.0
9807   16773           1  2014-10-23T22:42:44.587     15     3716.0

                                                    Body  OwnerUserId  \
1575   <p>I asked this question in various places on ...          NaN
10932  <p>The aborigines are believed to have migrate...       8451.0
2305   <p>My Australian friend says yes. I'm not so s...        282.0
15714  <p>Leaving aside the naval battles of the Paci...      10028.0
33259  <p>From what I've gathered at a glance, Aborig...      40601.0
3000   <p>Australia hosted aboriginal populations sin...       1436.0
19672  <p>Is there any evidence of ancient Chinese ex...      23474.0
7415   <p>The common explanation for the usage of Aus...       3682.0
7577   <p>In Australia's History, the Europeans tried...          NaN
9807   <p>The gold rush in Australia saw many <a href...         49.0

       LastEditorUserId             LastEditDate         LastActivityDate  \
1575               -1.0  2015-07-06T09:35:26.337  2018-05-28T10:09:10.750
10932            8451.0  2016-03-12T14:37:21.250  2018-03-28T00:22:32.030
2305             3810.0  2014-03-15T11:38:12.450  2017-02-09T15:05:49.873
15714            4390.0  2016-04-26T21:25:31.293  2024-02-03T05:01:41.707
33259           40601.0  2019-10-14T06:39:49.547  2024-03-22T11:08:29.800
3000             1110.0  2012-11-03T19:02:36.883  2021-01-28T09:14:11.107
19672           56412.0  2022-10-06T16:37:28.960  2023-11-24T00:21:38.513
7415             3682.0  2014-03-06T14:55:04.387  2014-03-06T15:09:47.163
7577                NaN                      NaN  2014-03-09T22:05:16.410
9807                NaN                      NaN  2018-11-20T04:39:13.560

                                                   Title  \
1575   Why did Canada, Australia and New Zealand sepa...
10932  Did the aborigines of Australia and the Maoris...
2305   Can the Queen of England fire the prime minist...
15714  Why was the Japanese Army's fatalities inflict...
33259  Why didn't Aboriginal Australians invent agric...
3000   Why did Austronesian/Polynesian people not col...
19672  Is there evidence of ancient Chinese explorati...
7415           America vs. Australia as a "Penal Colony"
7577   Did any other countries try to breed a race ou...
9807   Which fruits and vegetables did Chinese migran...

                                                    Tags  AnswerCount  \
1575   |british-empire|united-kingdom|ireland|austral...          9.0
10932                |australia|aboriginals|new-zealand|          4.0
```

```
2305                                |australia|monarchy|          4.0
15714      |united-states|world-war-two|japan|australia|          7.0
33259                                        |australia|          8.0
3000            |colonization|australia|south-east-asia|          8.0
19672                           |australia|exploration|          1.0
7415                    |united-states|australia|colony|          1.0
7577                            |australia|aboriginals|          1.0
9807                         |food|agriculture|australia|          1.0

       CommentCount ContentLicense  \
1575              6   CC BY-SA 3.0
10932             4   CC BY-SA 3.0
2305              3   CC BY-SA 3.0
15714            12   CC BY-SA 3.0
33259             6   CC BY-SA 4.0
3000             11   CC BY-SA 3.0
19672             3   CC BY-SA 4.0
7415              1   CC BY-SA 3.0
7577             13   CC BY-SA 3.0
9807              4   CC BY-SA 3.0

                                                  Clean Body  \
1575   I asked this question in various places on the...
10932  The aborigines are believed to have migrated f...
2305   My Australian friend says yes. I'm not so sure...
15714  Leaving aside the naval battles of the Pacific...
33259  From what I've gathered at a glance, Aborigine...
3000   Australia hosted aboriginal populations since ...
19672  Is there any evidence of ancient Chinese explo...
7415   The common explanation for the usage of Austra...
7577   In Australia's History, the Europeans tried to...
9807   The gold rush in Australia saw many Chinese mi...

                                       Key Words
1575              Why Canada Australia New Zealand
10932  Did Australia Maoris New Zealand Europeans
2305                   Can Queen England Australia
15714                    Why Japanese Army Pacific
33259                  Why Aboriginal Australians
3000         Why Austronesian Polynesian Australia
19672                         Is Chinese Australia
7415                America Australia Penal Colony
7577                                Did Australia
```

```
9807                        Which Chinese Australia

C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\3766916687.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

```python
# Print out the rows that contain the title, body, and keyword of that post
for row_idx, (title, body, keyword) in enumerate(zip(top10_aus["Title"],
 ↪  top10_aus["Clean Body"], top10_aus["Key Words"]), start=1):
    print(f"ROW {row_idx} :")
    print(f"Title: {title}")
    print(body)
    print(f"****KEYWORD: {keyword}")
    print("==" * 30)
    print()
```

```
ROW 1 :
Title: Why did Canada, Australia and New Zealand separate from the UK?
I asked this question in various places on the Web. I haven't received a clear answer.

The USA separated from the UK because the people felt that they were not British, and they we

Weren't Canadians/Australians/New Zealanders also descended from the British?

Or was it that UK was unable to administer them for financial reasons? I will not accept this

Why are Canada/Australia/New Zealand not administered from the UK?

****KEYWORD: Why Canada Australia New Zealand
============================================================

ROW 2 :
Title: Did the aborigines of Australia and the Maoris in New Zealand know about each other's
The aborigines are believed to have migrated from India in prehistoric times. The Maoris are

It's one and a half thousand miles from NZ to OZ, about the same as the distance between Brit

****KEYWORD: Did Australia Maoris New Zealand Europeans
```

==========

ROW 3 :
Title: Can the Queen of England fire the prime minister of Australia?
My Australian friend says yes. I'm not so sure, considering the monarchy collects no taxes fi

****KEYWORD: Can Queen England Australia
==========

ROW 4 :
Title: Why was the Japanese Army's fatalities inflicted:suffered ratio so low in the Pacific
Leaving aside the naval battles of the Pacific, why did the Japanese army do so poorly agains

While the Battle of Kaiapit is a great example of this, where Australia lost 14 men to over 2

What is the explanation for this? It was not as if the Imperial Army had not had ample traini

****KEYWORD: Why Japanese Army Pacific
==========

ROW 5 :
Title: Why didn't Aboriginal Australians invent agriculture?
From what I've gathered at a glance, Aborigine Australians have about 50000 years of history

While Jared Diamond argues that agriculture was not as easily possible, a cursory Wikipedia s

I'm not much of a geography expert, but if I were to apply the concept of a "hot, dry" climat

Were there any other sources of troubles for Aborigines that prevented their discovering othe

****KEYWORD: Why Aboriginal Australians
==========

ROW 6 :
Title: Why did Austronesian/Polynesian people not colonize Australia?
Australia hosted aboriginal populations since prehistory. However technologically advanced ci

Why was it not colonized by those people? Is there any evidence of interaction/invasions?

****KEYWORD: Why Austronesian Polynesian Australia
==========

ROW 7 :

Title: Is there evidence of ancient Chinese exploration of Australia?
Is there any evidence of ancient Chinese exploration of Australia? Specifically, up through t

****KEYWORD: Is Chinese Australia
==========

ROW 8 :
Title: America vs. Australia as a "Penal Colony"
The common explanation for the usage of Australia as a penal colony is that the American Revo
From Wikipedia, "On Convicts in Australia":

> Alternatives to the American colonies were investigated...

But I'm wondering why then are the methods so different? If they just switched over, the mode

The usage of America as a penal colony is explained as a transfer of prisoners sold as indent

****KEYWORD: America Australia Penal Colony
==========

ROW 9 :
Title: Did any other countries try to breed a race out like what happened in Australia?
In Australia's History, the Europeans tried to breed out Aboriginals colour by pairing a Abor

That Aboriginals were a dying race
That Australia was a 'white mans' country
If they did this than if they had a baby it would be a quarter white and so on until the blac

My question is did any other countries do this and was this an effect of Transnational Histor

Answers are appreciated.

****KEYWORD: Did Australia
==========

ROW 10 :
Title: Which fruits and vegetables did Chinese migrants introduce to Australia during the gol
The gold rush in Australia saw many Chinese migrate to the country, with the Chinese populati

```
Which fruits and vegetables were introduced to Australia by Chinese migrants during the gold

****KEYWORD: Which Chinese Australia
============================================================
```

Now that we've identified the major historical events based on the posts with the highest view counts, let's create a map and mark the locations to aid in our interpretation. Separation of Australia from the United Kingdom in Westminster.

```
# Define additional information for Westminster, London popup
westminster_popup_html = """
<h4>Westminster - Separation of Australia from the United Kingdom in
 ↪  Westminster</h4>

<section>
    <h5>Introduction</h5>
    <p>The separation of Australia from the United Kingdom in Westminster
 ↪  refers to the process through which Australia gained legislative
 ↪  independence from the UK Parliament.</p>
    <p>This process culminated in the passage of the Statute of Westminster
 ↪  Adoption Act 1942 by the Australian Parliament.</p>
    <p>The act effectively recognized Australia's legislative autonomy,
 ↪  allowing it to pass laws without requiring approval from the British
 ↪  Parliament.</p>
</section>

<section>
    <h5>Background</h5>
    <p>Before this separation, Australia, along with other British Dominions,
 ↪  operated under the legal framework of British parliamentary
 ↪  sovereignty.</p>
    <p>However, as Australia grew and developed as a nation, there was
 ↪  increasing sentiment for greater autonomy and self-governance.</p>
    <p>The Statute of Westminster 1931 laid the groundwork for this process
 ↪  by granting legislative independence to the self-governing Dominions of
 ↪  the British Empire, including Australia.</p>
    <p>However, Australia initially hesitated to adopt the statute due to
 ↪  concerns about its potential impact on the federation and constitutional
 ↪  arrangements.</p>
</section>

<section>
```

```
    <h5>Legislative Independence</h5>
    <p>It wasn't until the outbreak of World War II and the fall of Singapore
 ↪  in 1942 that the Australian government felt the urgency to assert its
 ↪  legislative independence.</p>
    <p>In response to these events, the Australian Parliament passed the
 ↪  Statute of Westminster Adoption Act 1942, formally adopting the Statute
 ↪  of Westminster and severing legislative ties with the UK Parliament.</p>
    <p>This act marked a significant milestone in Australia's journey towards
 ↪  full sovereignty and independence.</p>
    <p>It represented a symbolic and practical step towards asserting
 ↪  Australia's status as a fully independent nation within the
 ↪  Commonwealth.</p>
</section>
"""
```

Torres Strait Islands - interactions between the Aboriginal peoples of Australia and the Māori of New Zealand.

```
torrens_strait_popup_html = """
<h4>Torres Strait Islands</h4>
<section>
    <h5>Introduction</h5>
    <p>The Torres Strait Islands are a group of at least 274 small islands in
 ↪  the Torres Strait waterway between Australia and Papua New Guinea.</p>
    <p>The islands are home to the Indigenous Torres Strait Islander people
 ↪  and have a rich cultural heritage.</p>
</section>

<section>
    <h5>Interactions with Aboriginal Australians and the Māori</h5>
    <p>While there isn't extensive direct evidence of regular contact between
 ↪  Aboriginal Australians and the Māori prior to European arrival, it's
 ↪  theorized that some limited interaction may have occurred through the
 ↪  Torres Strait Islands. Both Aboriginal Australians and the Māori are
 ↪  thought to have had maritime capabilities, with the Māori famously known
 ↪  for their impressive voyaging canoes.</p>
    <p>The Torres Strait Islands were inhabited by Indigenous Torres Strait
 ↪  Islander peoples who had established trade networks and cultural
 ↪  connections with neighboring regions, including Papua New Guinea and
 ↪  other parts of the Pacific. It's possible that these islanders served as
 ↪  intermediaries or facilitators of contact between the Aboriginal peoples
 ↪  of Australia and the Māori of New Zealand.</p>
```

```
    <p>However, the exact nature and extent of these interactions remain
 ↪  subjects of ongoing research and debate among historians,
 ↪  anthropologists, and archaeologists. While evidence suggests some level
 ↪  of contact and exchange, further studies are needed to fully understand
 ↪  the dynamics of pre-European contact between these Indigenous groups
 ↪  through the Torres Strait Islands.</p>
</section>
"""
```

Canberra, the capital of Australia, is where decisions are made by the Prime Minister.

```
# Define the popup HTML content for Canberra
canberra_popup_html = """
<h4>Canberra, Australia</h4>
<section>
    <h5>Constitutional Monarchy</h5>
    <p>Australia operates as a constitutional monarchy with a parliamentary
 ↪  system of government. The Queen of England is the head of state, but her
 ↪  powers are largely ceremonial and symbolic.</p>
</section>
<section>
    <h5>Appointment of Prime Minister</h5>
    <p>The Prime Minister of Australia is appointed by the Governor-General,
 ↪  who is the Queen's representative in Australia. However, this appointment
 ↪  is based on the political realities of the Australian Parliament,
 ↪  specifically the party or coalition of parties that commands the majority
 ↪  in the House of Representatives.</p>
</section>
<section>
    <h5>Removal of Prime Minister</h5>
    <p>In practice, the Prime Minister serves at the pleasure of the
 ↪  Australian Parliament and can be removed through political processes such
 ↪  as a vote of no confidence or a leadership spill within the Prime
 ↪  Minister's party.</p>
</section>
"""
```

Guadalcanal holds importance for Australia due to its role as a significant battleground during World War II, where both Allied and Japanese forces were engaged in a major campaign.

```
japan_popup_html = """
<h4>Guadalcanal</h4>
<section>
    <h5>Background</h5>
    <p>Guadalcanal, located in the Solomon Islands, holds historical
 ↪  significance for Australia as it was the site of a major campaign during
 ↪  World War II involving Allied and Japanese forces.</p>
</section>
<section>
    <h5>Battle</h5>
    <p>The battle for Guadalcanal, which raged from August 1942 to February
 ↪  1943, saw heavy casualties for both sides. The Japanese suffered
 ↪  significant losses while attempting to defend the island against Allied
 ↪  invasion.</p>
</section>
<section>
    <h5>Significance</h5>
    <p>The campaign for Guadalcanal marked a pivotal moment in the Pacific
 ↪  theater of World War II. It represented the first major offensive
 ↪  launched by Allied forces against Japanese-held territory.</p>
    <p>The eventual Allied victory at Guadalcanal provided Australia with a
 ↪  strategic foothold in the Solomon Islands, strengthening its position in
 ↪  the region and contributing to the overall Allied effort against Japanese
 ↪  expansion in the Pacific.</p>
</section>
"""
```

Uluru, also known as Ayers Rock, which is a culturally significant landmark in central Australia and holds significance to the Aboriginal people. This location serves as a symbolic representation of Aboriginal culture and heritage.

```
uluru_popup_html = """
<h4>Uluru - Factors Influencing the Absence of Agriculture Among Aboriginal
 ↪  Australians</h4>
<section>
    <h5>Background</h5>
    <p>Despite their lengthy history of approximately 50,000 years on the
 ↪  continent, Aboriginal Australians did not develop agriculture to the
 ↪  extent seen in other parts of the world.</p>
    <p>While some argue that the temperate and subtropical climates of
 ↪  Australia could have supported agricultural practices, the development of
 ↪  advanced technology and agricultural systems among Aboriginal
 ↪  civilizations remained limited compared to counterparts in the Old
 ↪  World.</p>
```

```
</section>
<section>
    <h5>Factors</h5>
    <p>Factors such as isolation, limited resources, and cultural practices
 ↪  may have hindered the development and adoption of agricultural techniques
 ↪  among Aboriginal Australians.</p>
    <p>While geographical factors may have played a role, other challenges
 ↪  likely contributed to the absence of agricultural practices.</p>
</section>
<section>
    <h5>Conclusion</h5>
    <p>The absence of agriculture among Aboriginal Australians likely
 ↪  resulted from a combination of geographical, cultural, and historical
 ↪  factors that influenced their societies' development and technological
 ↪  advancements.</p>
</section>
"""
```

Absence of Austronesian/Polynesian Colonization.

```
polynesian_popup_html = """
<h4>Torres Strait Islands - Factors Influencing the Absence of
 ↪  Austronesian/Polynesian Colonization of Australia</h4>
<section>
    <p>Despite the presence of Aboriginal populations in Australia since
 ↪  prehistory, neighboring regions such as Indonesia, Polynesia, and New
 ↪  Guinea were home to technologically advanced civilizations. However, the
 ↪  Austronesian and Polynesian peoples did not colonize Australia, raising
 ↪  questions about the reasons behind this absence of colonization and
 ↪  whether there is evidence of interaction or invasion.</p>
</section>
<section>
    <h5> Geographic Barriers</h5>
    <p>The vast expanses of open sea, such as the Torres Strait, separating
 ↪  Australia from neighboring regions could have posed significant
 ↪  challenges to maritime navigation and exploration.</p>
</section>
<section>
    <h5> Cultural Priorities</h5>
    <p>Austronesian and Polynesian societies may have had different cultural
 ↪  priorities or objectives that did not include expansion into Australia.
 ↪  These societies might have focused on other areas for settlement,
 ↪  resource exploitation, or cultural development.</p>
```

```
</section>
<section>
    <h5> Adaptation to Local Environments</h5>
    <p>The ecosystems and environments of Australia differ significantly from
 ↪  those of neighboring regions. Austronesian and Polynesian peoples may
 ↪  have been better adapted to the maritime and island environments of their
 ↪  homelands, making the Australian continent less attractive for
 ↪  colonization.</p>
</section>
<section>
    <h5> Resource Availability</h5>
    <p>Australia's arid and harsh interior may have presented challenges in
 ↪  terms of resource availability for potential colonizers. The lack of
 ↪  reliable freshwater sources and fertile land suitable for agriculture
 ↪  could have deterred Austronesian and Polynesian peoples from establishing
 ↪  permanent settlements in Australia.</p>
</section>
<section>
    <h5> Limited Evidence of Interaction</h5>
    <p>While there is some evidence of contact and interaction between
 ↪  Aboriginal Australians and Austronesian/Polynesian peoples, such as the
 ↪  presence of traded goods and linguistic connections, there is limited
 ↪  evidence of large-scale colonization or invasion.</p>
</section>
<p>Overall, the absence of Austronesian/Polynesian colonization of Australia
 ↪  likely resulted from a combination of geographical, cultural, and
 ↪  environmental factors that influenced the priorities and capabilities of
 ↪  these societies.</p>
"""
```

Sydney - Ancient Chinese Exploration of Australia.

```
sydney_popup_html = """
<h4>Sydney - Evidence of Ancient Chinese Exploration of Australia</h4>

<section>
    <h5>Introduction</h5>
    <p>Exploring the possibility of ancient Chinese exploration of Australia
 ↪  before extensive contact with Europeans during the Middle Ages raises
 ↪  intriguing questions about maritime history and cross-cultural
 ↪  interactions.</p>
</section>
```

```
<section>
    <h5>Background</h5>
    <p>While concrete evidence of direct Chinese exploration remains elusive,
 ↪  several theories and pieces of evidence suggest the potential for early
 ↪  Chinese contact with the Australian continent.</p>
</section>

<section>
    <h5>Evidence</h5>
    <p>One notable piece of evidence is the presence of ancient Chinese maps,
 ↪  such as the Kangnido and Zheng He's maps, which depict lands far beyond
 ↪  China's borders, including regions that could correspond to parts of
 ↪  Australia.</p>
    <p>Additionally, archaeological findings, such as Chinese artifacts
 ↪  discovered along the northern coast of Australia, fuel speculation about
 ↪  possible contact between Chinese voyagers and Indigenous Australian
 ↪  populations.</p>
    <p>Historical records and accounts, such as those from Chinese voyages
 ↪  led by Zheng He in the early 15th century, offer further insights into
 ↪  the potential for Chinese maritime expeditions to have reached distant
 ↪  shores, including parts of Australia.</p>
</section>

<section>
    <h5>Conclusion</h5>
    <p>Overall, while conclusive evidence of ancient Chinese exploration of
 ↪  Australia remains elusive, various theories, artifacts, maps, and
 ↪  historical accounts suggest the intriguing possibility of early
 ↪  cross-cultural encounters between Chinese sailors and Indigenous
 ↪  Australians.</p>
    <p>Further archaeological research, interdisciplinary studies, and
 ↪  exploration of historical sources may shed more light on this fascinating
 ↪  topic.</p>
</section>
"""
```

America vs. Australia as a “Penal Colony”.

```
america_vs_australia_html = """
<h4>Sydney - America vs. Australia as a "Penal Colony"</h4>
```

```
<section>
    <h5>Introduction</h5>
    <p>The common explanation for the usage of Australia as a penal colony is
 ↪  that the American Revolution made the usage of America as one no longer
 ↪  tenable.</p>
</section>

<section>
    <h5>Background</h5>
    <p>Before the American Revolution, Britain sent many convicts to the
 ↪  American colonies, often as indentured servants. These indentured
 ↪  servants were sold to work for a specified number of years.</p>
    <p>After the American colonies gained independence, Britain had to find
 ↪  an alternative location for its convicts, leading to the establishment of
 ↪  Australia as a penal colony.</p>
</section>

<section>
    <h5>Differences in Methods</h5>
    <p>The American model involved selling convicts as indentured servants
 ↪  who were eventually integrated into the population. In contrast, the
 ↪  Australian model involved the establishment of penal colonies where
 ↪  convicts were isolated and worked on public projects or in designated
 ↪  settlements.</p>
</section>

<section>
    <h5>Development of the Australian Model</h5>
    <p>The Australian convict transportation model involved strict
 ↪  supervision, infrastructure establishment, and the gradual transition of
 ↪  convicts into free settlers once their sentences were served. This system
 ↪  significantly influenced the development of Australian society.</p>
</section>
"""
```

Did Any Other Countries Attempt to Breed Out a Race Like in Australia ?

```
breeding_out_popout_html = """
<h4>Canberra - Did Any Other Countries Attempt to Breed Out a Race Like in
 ↪  Australia?</h4>
```

```
<section>
    <p>In Australia's history, the policy of "breeding out the colour"
 ↪  involved pairing Aboriginal women with white men. This practice was part
 ↪  of a broader effort to assimilate Aboriginal people and erase their
 ↪  cultural and racial identity. The policy was driven by several
 ↪  beliefs:</p>
</section>

<section>
    <h5>Beliefs Behind the Policy</h5>
    <ul>
        <li>Aboriginals were a dying race: Europeans believed that Aboriginal
 ↪  people would eventually die out and saw interbreeding as a way to manage
 ↪  this perceived inevitability.</li>
        <li>Australia was a 'white man's' country: The policy was part of a
 ↪  broader effort to maintain Australia as a predominantly white
 ↪  nation.</li>
        <li>Erasing the black colour: By encouraging mixed-race marriages and
 ↪  subsequent generations, the aim was to gradually eliminate the physical
 ↪  characteristics associated with Aboriginal people.</li>
    </ul>
    <p>This policy was implemented during the era known as the Stolen
 ↪  Generation, roughly from 1909 to 1969. Many Aboriginal children were
 ↪  forcibly removed from their families and placed in institutions or with
 ↪  white families to be assimilated into white society.</p>
</section>

<section>
    <h5>Comparison with Other Countries</h5>
    <p>The concept of "breeding out" a race has been seen in other contexts
 ↪  as well, though the specifics and motivations may differ:</p>
    <ul>
        <li><strong>United States:</strong> Various assimilation policies
 ↪  aimed at Native American populations, including forced removals and
 ↪  boarding schools.</li>
        <li><strong>Canada:</strong> Similar to the US, with residential
 ↪  schools and policies to assimilate Indigenous populations.</li>
        <li><strong>Nazi Germany:</strong> Eugenics policies aimed at
 ↪  creating a pure Aryan race through selective breeding and elimination of
 ↪  those considered racially inferior.</li>
    </ul>
    <p>These policies reflect broader trends in transnational history where
 ↪  colonial and imperial powers implemented strategies to assimilate or
 ↪  eliminate indigenous populations.</p>
```

```
</section>
"""
```

Fruits and Vegetables Introduced by Chinese Migrants During the Australian Gold Rush.

```
chinese_migrants_popout_html = """
<h4>Ballarat - Fruits and Vegetables Introduced by Chinese Migrants During
 ↪  the Australian Gold Rush</h4>

<section>
    <p>During the Australian gold rush of the mid-19th century, many Chinese
 ↪  migrants came to Australia, reaching a population of around 40,000 by the
 ↪  1860s. These migrants brought with them various vegetable seeds and
 ↪  horticultural knowledge, significantly impacting Australia's agricultural
 ↪  landscape.</p>
</section>

<section>
    <h5>Vegetables Introduced</h5>
    <ul>
        <li>Bok Choy: A type of Chinese cabbage that became popular in
 ↪  Australian cuisine.</li>
        <li>Chinese Cabbage: Different from Western cabbages, these varieties
 ↪  added diversity to the available produce.</li>
        <li>Snow Peas: Known for their sweet taste and crunchy texture, snow
 ↪  peas became a staple in Australian markets.</li>
        <li>Chinese Broccoli (Gai Lan): This leafy green vegetable added to
 ↪  the variety of greens available in Australia.</li>
        <li>Garlic Chives: With their distinct flavor, these chives were used
 ↪  in various dishes.</li>
    </ul>
</section>

<section>
    <h5>Fruits Introduced</h5>
    <ul>
        <li>Chinese Gooseberries (Kiwifruit): Although more associated with
 ↪  New Zealand, the kiwifruit was introduced by Chinese migrants.</li>
        <li>Persimmons: Known for their sweet and unique taste, persimmons
 ↪  were another contribution.</li>
    </ul>
```

```
</section>

<section>
    <h5>Commercial Growth and Success</h5>
    <p>Many of these fruits and vegetables were successfully grown
 ↪  commercially by Chinese farmers. They dominated the fruit and vegetable
 ↪  markets by the late 19th century, even though the produce was not
 ↪  necessarily of Chinese origin.</p>
    <p>The Chinese growers were particularly successful in urban markets and
 ↪  mining towns where the demand for fresh produce was high. However, in
 ↪  cities and towns with lower concentrations of Chinese people, the
 ↪  acceptance of these new vegetables varied. Over time, many of these
 ↪  products became integrated into the broader Australian diet.</p>
</section>

<section>
    <h5>Impact on Agriculture</h5>
    <p>The Chinese migrants introduced innovative farming techniques and crop
 ↪  diversity, which enriched Australian agriculture. These introductions
 ↪  facilitated cultural exchange and dietary diversification in Australia,
 ↪  influencing Australian cuisine for generations.</p>
</section>
"""
```

```
# Create a map centered at Canberra
m = folium.Map(location=[-35.3075, 149.1244], zoom_start=10)

# Create a MarkerCluster to prevent overlapping markers
marker_cluster = MarkerCluster().add_to(m)


#  Westminster, London and Torrens Strait Islands
folium.Marker(
    location=[51.5007, -0.1246],
    popup=folium.Popup(westminster_popup_html, max_width=300,
 ↪  max_height=200), # Enable scrolling
    icon=folium.Icon(color="green")
).add_to(m)

# Torrens Strait
folium.Marker(
```

```
    location=[-10.0, 142.0],
    popup=folium.Popup(torrens_strait_popup_html, max_width=300,
↪   max_height=200),
    icon=folium.Icon(color="blue")
).add_to(marker_cluster)

# Canberra
folium.Marker(
    location=[-35.3075, 149.1244],
    popup=folium.Popup(canberra_popup_html, max_width=300, max_height=200),
     icon=folium.Icon(color="purple")
).add_to(marker_cluster)

# Guadalcanal
folium.Marker(
    location=[-9.1944, 159.9521],
    popup=folium.Popup(japan_popup_html, max_width=300, max_height=200),
     icon=folium.Icon(color="blue")
).add_to(m)

# Add a marker for Uluru with the popup containing the information
folium.Marker(
    location=[-25.3444, 131.0369],
    popup=folium.Popup(uluru_popup_html, max_width=300, max_height=200),
    icon=folium.Icon(color="orange")
).add_to(m)

# Torres Strait Islands and Austronesian/Polynesian Colonization
folium.Marker(
    location=[-10.0, 142.0],
    popup=folium.Popup(polynesian_popup_html, max_width=300, max_height=200),
↪
    icon=folium.Icon(color="red")
).add_to(marker_cluster)

# Sydney
folium.Marker(
    location=[-33.8688, 151.2093],
    popup=folium.Popup(sydney_popup_html, max_width=300, max_height=200),
    icon = folium.Icon(color = "green")
).add_to(marker_cluster)
```

```
folium.Marker(
    location=[-33.8688, 151.2093],
    popup=folium.Popup(america_vs_australia_html, max_width=600,
 ↪  max_height=400),
    icon=folium.Icon(color="blue")
).add_to(marker_cluster)

# Canberra
folium.Marker(
    location=[-35.3075, 149.1244],
    popup=folium.Popup(breeding_out_popout_html, max_width=600,
 ↪  max_height=400),
    icon=folium.Icon(color="red")
).add_to(marker_cluster)

# r Ballarat
folium.Marker(
    location=[-37.5622, 143.8503],
    popup=folium.Popup(chinese_migrants_popout_html, max_width=600,
 ↪  max_height=400),
    icon=folium.Icon(color='red')
).add_to(m)
# Display the map
m.save("map.html")
```

Click here to view the map view map

From the information gathered from our map and the posts, it is evident that the most viewed posts primarily focus on Aboriginal people, Chinese immigrants, issues related to World War I and II, and matters of independence.

These topics highlight key aspects of Australia's complex history, such as the rich cultural heritage and the challenges faced by Aboriginal communities, the significant contributions and experiences of Chinese immigrants during pivotal periods like the gold rush, the profound impact of global conflicts on Australia's national identity and development, and the country's journey towards legislative and political independence.

Now in the next step, we will examine the number of URLs mentioned in each post and its associated comments. Our goal is to determine which domain is referenced the most frequently and to compare its occurrence rate with that of other URLs.

```
# Function to extract and count domain URLs in a text
def extract_and_count_urls(text):
    url_pattern = re.compile(r'https?://([^/ ]+)')
    urls = re.findall(url_pattern, text)
    url_counts = {}
    for url in urls:
        if url in url_counts:
            url_counts[url] += 1
        else:
            url_counts[url] = 1
    return url_counts

# Return True if dictionary is not empty, False otherwise
def filter_empty_dicts(d):
    return bool(d)
```

```
# Create a Url dataframe that just has the body and the title of the post
url_df = post_df[["Body","Title"]]
```

```
print(url_df.head())
```

```
                                                 Body  \
0   <p>What factors related to the Eastern Crisis ...
1   <p>What language(s) were considered the primar...
2   <p>When in history do we first have record of ...
3   <p>The title above is fairly broad, so more sp...
10  <p>Political discussions leading up to the Bay...

                                                Title
0   What factors related to the Eastern Crisis con...
1   What language(s) were spoken within the Holy R...
2                         When did steel first appear?
3   What events led to the fall of the Zhou Dynast...
10  Who, if any, in JFK's inner circle argued agai...
```

```
# Again we will remove tags and clean the text
url_df["Clean Body"] = url_df["Body"].apply(remove_html_tags)
print(url_df.head())
```

```
                                                 Body  \
0   <p>What factors related to the Eastern Crisis ...
1   <p>What language(s) were considered the primar...
2   <p>When in history do we first have record of ...
3   <p>The title above is fairly broad, so more sp...
10  <p>Political discussions leading up to the Bay...

                                                Title  \
0   What factors related to the Eastern Crisis con...
1   What language(s) were spoken within the Holy R...
2                        When did steel first appear?
3   What events led to the fall of the Zhou Dynast...
10  Who, if any, in JFK's inner circle argued agai...

                                           Clean Body
0   What factors related to the Eastern Crisis con...
1   What language(s) were considered the primary l...
2   When in history do we first have record of ste...
3   The title above is fairly broad, so more speci...
10  Political discussions leading up to the Bay of...

C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\2880346372.py:2: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guid
```

```
# Apply the function to the cleanned body column and create a new column
 ↪  represents the urls counts
url_df["Url Counts Body"] = url_df["Clean
 ↪  Body"].apply(extract_and_count_urls)
print(url_df)
```

```
                                                 Body  \
```

```
0      <p>What factors related to the Eastern Crisis ...
1      <p>What language(s) were considered the primar...
2      <p>When in history do we first have record of ...
3      <p>The title above is fairly broad, so more sp...
10     <p>Political discussions leading up to the Bay...
...                                                  ...
44149  <p>I assume they would not call it the &amp;quot;B...
44151  <p>Was Spartacus related to Sparta?\n<a href="...
44153  <p>This would be a better fit for trains.stack...
44155  <p>I'm thinking about years later operation MK...
44161  <p>I recently came across this article - <a hr...

                                                  Title  \
0      What factors related to the Eastern Crisis con...
1      What language(s) were spoken within the Holy R...
2                              When did steel first appear?
3      What events led to the fall of the Zhou Dynast...
10     Who, if any, in JFK's inner circle argued agai...
...                                                  ...
44149  What would term would a lollard use to refer t...
44151  Was Spartacus related to Sparta? Was he a Spar...
44153              How does one "ride on the brake beam"?
44155  What is the best example of a "conspiracy theo...
44161  How much wealth did the US "earn" from Philipp...

                                              Clean Body Url Counts Body
0      What factors related to the Eastern Crisis con...                {}
1      What language(s) were considered the primary l...                {}
2      When in history do we first have record of ste...                {}
3      The title above is fairly broad, so more speci...                {}
10     Political discussions leading up to the Bay of...                {}
...                                                  ...               ...
44149  I assume they would not call it the &amp;quot;Bibl...             {}
44151  Was Spartacus related to Sparta?\nSpartacus li...                {}
44153  This would be a better fit for trains.stackexc...                {}
44155  I'm thinking about years later operation MKUlt...                {}
44161  I recently came across this article - 'There's...                {}

[14864 rows x 4 columns]

C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\3884573359.py:2: SettingWithCopyWarning:
```

```
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

```
# Convert the 'url_counts' column to a DataFrame with separate rows for each
 ↪  URL and its count
url_counts_body = pd.DataFrame(url_df["Url Counts Body"].tolist(),
 ↪  index=url_df.index).stack().reset_index()
url_counts_body.columns = ["index", "domain", "count_body"]
url_counts_body = url_counts_body.drop(columns = ["index"])

print(url_counts_body)
```

```
                                          domain  count_body
0                                 en.wikipedia.org         1.0
1                              niv.scripturetext.com         1.0
2                                 www.salon.com         1.0
3                                www.youtube.com         1.0
4                                 en.wikipedia.org         1.0
..                                             ...         ...
591                                    kith.com         1.0
592                                     doi.org         2.0
593                                   www.unm.edu         1.0
594                                i.stack.imgur.com         1.0
595  mcoecbamcoepwprd01.blob.core.usgovcloudapi.net         1.0

[596 rows x 2 columns]
```

Currently, we have the count of URLs within the body of each post, but this count is limited to individual posts. Since a URL can be referenced multiple times across different posts, we need to calculate the total number of times each URL appears in the entire dataset.

```
url_counts_body = url_counts_body.groupby("domain").size().to_frame()
print(url_counts_body)
```

```
                              0
domain
20agetravel.blogspot.pt       1
```

```
abbot.us                          1
acoup.blog                        1
afvdb.50megs.com                  1
albeityacademic.blogspot.co.uk    1
...                              ..
www2.idehist.uu.se                1
www7.uc.cl                        1
yogimami.com                      1
youtu.be                          7
zoeperissos.blogspot.de           1

[348 rows x 1 columns]
```

```
# Rename the column
url_counts_body.rename(columns={0: "Times appear"}, inplace=True)

# Sort the times appear to find the most referenced URL
url_counts_body.sort_values(by="Times appear", ascending=False, inplace=True)
url_counts_body = url_counts_body.reset_index()
print(url_counts_body)
```

```
                         domain  Times appear
0                en.wikipedia.org           137
1                 www.youtube.com            17
2               i.stack.imgur.com            12
3              en.m.wikipedia.org            11
4                         youtu.be             7
..                             ...           ...
343           military.wikia.org             1
344  military-history.fandom.com             1
345                   medium.com             1
346    media-cdn.tripadvisor.com             1
347      zoeperissos.blogspot.de             1

[348 rows x 2 columns]
```

Interestingly, Wikipedia is the most frequently referenced page, with its number of references far surpassing other domains. This suggests that users most commonly rely on Wikipedia to gather information and details for the content of their posts.

The next step is to examine the headers of the posts to check for any URLs included in each one.

```
# Apply the function to the Title column and create a new column represents
 ↪  the urls counts of that title
url_df["Url Counts Title"] = url_df["Title"].apply(extract_and_count_urls)
print(url_df.head())
```

```
                                                 Body  \
0   <p>What factors related to the Eastern Crisis ...
1   <p>What language(s) were considered the primar...
2   <p>When in history do we first have record of ...
3   <p>The title above is fairly broad, so more sp...
10  <p>Political discussions leading up to the Bay...

                                                Title  \
0   What factors related to the Eastern Crisis con...
1   What language(s) were spoken within the Holy R...
2                       When did steel first appear?
3   What events led to the fall of the Zhou Dynast...
10  Who, if any, in JFK's inner circle argued agai...

                                          Clean Body Url Counts Body  \
0   What factors related to the Eastern Crisis con...              {}
1   What language(s) were considered the primary l...              {}
2   When in history do we first have record of ste...              {}
3   The title above is fairly broad, so more speci...              {}
10  Political discussions leading up to the Bay of...              {}

   Url Counts Title
0              {}
1              {}
2              {}
3              {}
10             {}

C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\1957545844.py:2: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

```
# Convert the 'url_counts' column to a DataFrame with separate rows for each
 ↪  URL and its count
url_counts_title = pd.DataFrame(url_df["Url Counts Title"].tolist(),
 ↪  index=url_df.index).stack().reset_index()
url_counts_title.columns = ["index", "domain", "count_title"]
url_counts_title = url_counts_title.drop(columns = ["index"])

print(url_counts_title)
```

```
Empty DataFrame
Columns: [domain, count_title]
Index: []
```

Now we can see that it is evident that there are no URLs referenced in the headers of any posts. Let's determine if there are any URLs referenced in the comments of a specific post.

```
# Create a new dataframe for the cmt
url_cmt = cmt_df[["Text"]]
print(url_cmt.head())
```

```
                                                Text
0  Please elaborate a bit on "Eastern Crisis". I ...
1  Some elaboration would be handy. This question...
2  Seconding the request for elaboration on what ...
3  Copying from Wikipedia doesn't seem correct, e...
4  @Jacob, I agree, see my meta post regarding ci...
```

```
# Apply the function to the Text column and create a new column represents
 ↪  the urls counts of that text comments
url_cmt["Url Counts Cmt"] = url_cmt["Text"].apply(extract_and_count_urls)
print(url_cmt.head())
```

```
                                                Text Url Counts Cmt
0  Please elaborate a bit on "Eastern Crisis". I ...             {}
1  Some elaboration would be handy. This question...             {}
2  Seconding the request for elaboration on what ...             {}
3  Copying from Wikipedia doesn't seem correct, e...             {}
4  @Jacob, I agree, see my meta post regarding ci...             {}
```

```
C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\3656202012.py:2: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

Since the Comments dataframe is significantly larger than the post dataframe, we can't simply count the number of URLs appearing in each comment as we did with the post dataframe. Instead, we need to filter out rows with empty dictionaries (as these dictionaries represent the number of appearances of specific URLs in that comment) to reduce the size of the dataframe.

```
# Apply the function to filter out the rows that have empty dictionary in
 ↪  the DataFrame
url_cmt = url_cmt[url_cmt["Url Counts Cmt"].apply(filter_empty_dicts)]

url_cmt = url_cmt.reset_index(drop = True)
print(url_cmt)
```

```
                                                    Text  \
0      Also some of this question might be better in ...
1      @GPierce - I have started a [discussion in met...
2      or [linguistics.stackexchange.com](http://ling...
3      When you say "... the USA actually wanted..." ...
4      You've left out one of the big ones: the [Mugh...
...                                                  ...
20725  Does this answer your question? [Have there be...
20726  @ItalianPhilosophers4Monica See [Aliyah](https...
20727  I suspect that the "brake to Skilby Castle" wa...
20728  One clear example is Germany attempting to inf...
20729  Not what you're asking for but in terms of the...

                                      Url Counts Cmt
0                      {'french.stackexchange.com': 1}
1                  {'meta.history.stackexchange.com': 1}
2                   {'linguistics.stackexchange.com': 1}
3                               {'en.wikipedia.org': 1}
4                               {'en.wikipedia.org': 4}
...                                                 ...
```

```
20725                   {'history.stackexchange.com': 1}
20726                            {'en.wikipedia.org': 1}
20727                            {'en.wikipedia.org': 1}
20728  {'en.wikipedia.org': 2, 'www.archives.gov': 1,...
20729                           {'library.oapen.org': 1}

[20730 rows x 2 columns]
```

```
url_list_cmt = url_cmt["Url Counts Cmt"].tolist()
url_counts = {}

# Aggregate counts from list of dictionaries
for url_dict in url_list_cmt:
    for url, count in url_dict.items():
        if url in url_counts:
            url_counts[url] += count
        else:
            url_counts[url] = count

# Create DataFrame from dictionary
url_counts_cmt= pd.DataFrame(list(url_counts.items()), columns=["domain",
 ↪  "Time Appears Cmt"])

print(url_counts_cmt)
```

```
                            domain  Time Appears Cmt
0           french.stackexchange.com                 7
1     meta.history.stackexchange.com               126
2      linguistics.stackexchange.com                16
3                   en.wikipedia.org              9751
4                   www.stratfor.com                 1
...                              ...               ...
4793          americanliterature.com                 1
4794              lollardsociety.org                 2
4795       www.railway-technical.com                 1
4796                 www.tedrail.com                 1
4797               library.oapen.org                 1

[4798 rows x 2 columns]
```

```
# Sort out to find out which urls referenced the most in the comments
url_counts_cmt = url_counts_cmt.sort_values(by = "Time Appears Cmt",ascending
 ↪  = False).reset_index(drop = True)
print(url_counts_cmt)
```

```
                                  domain  Time Appears Cmt
0                         en.wikipedia.org              9751
1                 history.stackexchange.com              1738
2                    chat.stackexchange.com               597
3            history.meta.stackexchange.com               493
4     interpersonal.meta.stackexchange.com               455
...                                    ...               ...
4793               www.natureworldnews.com                 1
4794                            old.memo.ru                1
4795                         www.mesoweb.com                1
4796               www.britishempire.me.uk                1
4797                     library.oapen.org                1

[4798 rows x 2 columns]
```

Let's revisit the DataFrames we created to represent the occurrence of URLs referenced in the titles, bodies, and comments of the posts.

```
print("Occurences of the Urls in Tiles:")
print(url_counts_title)

print("Occurences of the Urls in Body:")
print(url_counts_body)

print("Occurences of the Urls in the Comments:")
print(url_counts_cmt)
```

```
Occurences of the Urls in Tiles:
Empty DataFrame
Columns: [domain, count_title]
Index: []
Occurences of the Urls in Body:
                         domain  Times appear
0                en.wikipedia.org           137
1                 www.youtube.com            17
```

```
2              i.stack.imgur.com            12
3             en.m.wikipedia.org            11
4                       youtu.be             7
..                           ...           ...
343           military.wikia.org             1
344  military-history.fandom.com             1
345                   medium.com             1
346    media-cdn.tripadvisor.com             1
347      zoeperissos.blogspot.de             1

[348 rows x 2 columns]
Occurences of the Urls in the Comments:
                                    domain  Time Appears Cmt
0                         en.wikipedia.org              9751
1                 history.stackexchange.com              1738
2                    chat.stackexchange.com               597
3            history.meta.stackexchange.com               493
4     interpersonal.meta.stackexchange.com               455
...                                    ...               ...
4793               www.natureworldnews.com                 1
4794                             old.memo.ru               1
4795                          www.mesoweb.com              1
4796               www.britishempire.me.uk                 1
4797                     library.oapen.org                 1

[4798 rows x 2 columns]
```

Now, let's combine our dataframes into a single dataframe that represents the frequency of URLs in both the post bodies and their comments. (The post titles will not be included as they do not contain any URLs).

```
url_complete = pd.merge(url_counts_body,url_counts_cmt,on = "domain",how =
 ↪  "inner")
print(url_complete)
```

```
                        domain  Times appear  Time Appears Cmt
0                en.wikipedia.org          137              9751
1                 www.youtube.com           17               418
2              i.stack.imgur.com            12                24
3             en.m.wikipedia.org            11               346
4                       youtu.be             7               111
```

```
..                            ...           ...               ...
142                myarmoury.com             1                 1
143     movies.stackexchange.com             1                15
144           military.wikia.org             1                 2
145  military-history.fandom.com             1                 2
146                   medium.com             1                 6

[147 rows x 3 columns]
```

From our combined dataframe, it's evident that some URLs are referenced in the post titles but not in the comments, and vice versa. This observation is notable as it has significantly reduced the number of rows in our merged dataframe. An important observation is that the Wikipedia page is the most frequently referenced site, indicating that its content likely provides valuable information and details that users find relevant and worthy of citation.

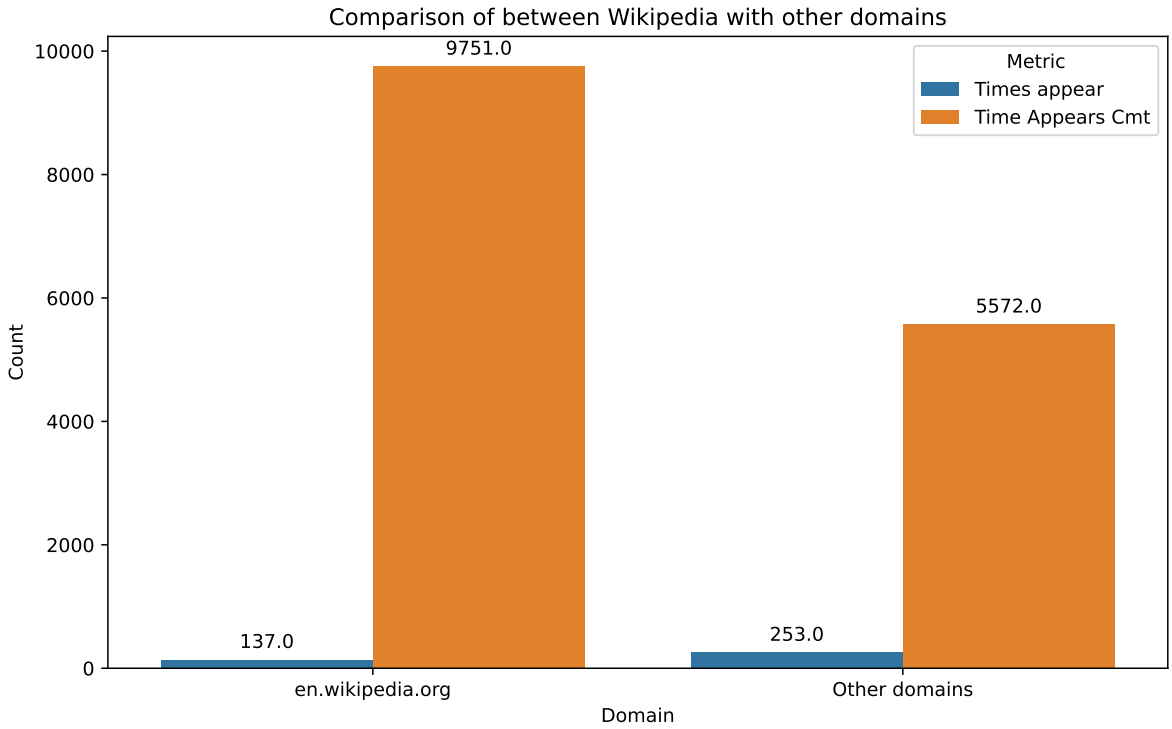
```
# Sum up the rows column-wise excluding the first row
sum_values = np.sum(url_complete.iloc[1:], axis=0)
print(sum_values)
```

```
domain              www.youtube.comi.stack.imgur.comen.m.wikipedia...
Times appear                                                      253
Time Appears Cmt                                                 5572
dtype: object
```

```
# Sum up the rows column-wise excluding the first row
sum_values = np.sum(url_complete.iloc[1:], axis = 0 )

# The first row which is the wikipedia page
first_row = url_complete.iloc[0].to_frame().T

# Create a DataFrame with the summed values and 'Total' as the domain
total_row = pd.DataFrame({"domain": ["Other domains"], "Times appear":
 ↪  [sum_values["Times appear"]], "Time Appears Cmt": [sum_values["Time
 ↪  Appears Cmt"]]})

# Concatenate the original DataFrame and the DataFrame with the total row
url_complete_with_total = pd.concat([first_row, total_row])


url_complete_with_total = url_complete_with_total.reset_index(drop = True)

print(url_complete_with_total)
```

```
              domain Times appear Time Appears Cmt
0  en.wikipedia.org          137              9751
1     Other domains          253              5572
```

```
# Melt the DataFrame to make it suitable for plotting
df_melted = url_complete_with_total.melt(id_vars="domain", var_name="Metric",
 ↪  value_name="Value")

# Plot
plt.figure(figsize=(10, 6))
axes = sns.barplot(data=df_melted, x="domain", y="Value", hue="Metric")

# Add annotations
for p in axes.patches:
    axes.annotate(format(p.get_height(), ".1f"),
                  (p.get_x() + p.get_width() / 2., p.get_height()),
                  ha = "center", va = "center",
                  xytext = (0, 9),
                  textcoords = "offset points")

plt.title("Comparison of between Wikipedia with other domains")
plt.xlabel("Domain")
plt.ylabel("Count")
plt.legend(title="Metric")
plt.show()
```



The bar plot unmistakably illustrates that URLs are more commonly present in comments than in the main posts. This trend is logical since individuals frequently incorporate references in their comments to reinforce their ideas and perspectives. It underscores the interactive and collaborative nature of discussions within online communities, where participants strive to enrich their contributions by providing external sources for credibility and further exploration.

More interestingly , the fact that Wikipedia is mentioned much more frequently in comments compared to all other websites combined suggests that people find it extremely useful. They rely on it for detailed and trustworthy information on a wide range of topics. When they include Wikipedia links in their comments, it's a way of showing that they've done their research and want to back up what they're saying with solid evidence.

In essence, the high frequency of Wikipedia mentions in comments reflects its central role as a source of knowledge and a catalyst for online interaction.

Next, we'll focus on analyzing how the sentiments expressed in the titles and bodies relate to the sentiments of the associated comments, as well as how the length of the content in both titles and bodies influences these sentiments.

Let's determine which topic receives the most mentions in our Stack Exchange data.

```
tag_df.sort_values(by = "Count",ascending = False)
```

	Id	TagName	Count	ExcerptPostId	WikiPostId
7	13	world-war-two	1591	44.0	43.0
81	162	united-states	1548	1485.0	1484.0
45	87	military	963	950.0	949.0
36	74	middle-ages	877	1487.0	1486.0
107	207	ancient-history	750	958.0	957.0
...	...	...	...	...	...
783	3101	post-war-japan	1	60156.0	60155.0
777	3088	nassau	1	59820.0	59819.0
774	3072	teutonic-order	1	59244.0	59243.0
710	2846	asdic	1	50056.0	50055.0
855	3363	middle-english	1	NaN	NaN

It appears that World War II is the most frequently discussed topic in our dataset. Now, let's delve into its analysis.

```
# Initialize a pre-trained sentiment analysis model called
 ↪  "cardiffnlp/twitter-roberta-base-sentiment"
# using the Hugging Face Transformers library
MODEL = f"cardiffnlp/twitter-roberta-base-sentiment"
tokenizer = AutoTokenizer.from_pretrained(MODEL)
model = AutoModelForSequenceClassification.from_pretrained(MODEL)
```

```
C:\Users\thinh\Documents\ANACONDA\Lib\site-packages\transformers\utils\generic.py:260: UserWa

torch.utils._pytree._register_pytree_node is deprecated. Please use torch.utils._pytree.regis
```

```
# Function to get the sentiment score
def polarity_scores_roberta(text):
    if isinstance(text, str):
        encoded_text = tokenizer(text, return_tensors='pt',max_length = 512)
        output = model(**encoded_text)
        scores = output[0][0].detach().numpy()
        scores = softmax(scores)

        # Calculate the compound score of the sentiment score
        # Basically the -1 represent the extreme negative , 0 : neutral , 1
         ↪  is extreme postive
        return (scores[0] * -1) + (scores[1] * 0) + (scores[2] * 1 )

    # If the text is nan return it as neutral
    else:
        return 0

# Function to remove special keywords
def remove_special_keywords(text):
    cleaned_text = re.sub(r'[!@#$%^&*]', '', text)
    return cleaned_text
```

```
# Test the function
negative_text = "Hi I really really hate you"
print("Negative score: ",polarity_scores_roberta(negative_text))
print("=" * 20)

neutral_text = "Hi"
print("Neutral score: ",polarity_scores_roberta(neutral_text))
print("="* 20)

positive_text = "Hi I incredibly like you a lot"
print("Positive score: ",polarity_scores_roberta(positive_text))
```

```
Truncation was not explicitly activated but `max_length` is provided a specific value, please

Negative score:  -0.9662588038481772
====================
Neutral score:  0.10823990404605865
====================
Positive score:  0.967550496570766
```

```
# Create a new dataframe represents represent the posts that related to WW2
ww2_df_post =
 ↪  post_df[post_df["Tags"].str.contains(r'\bworld-war-two\b')].reset_index(drop
 ↪  = True)
print(ww2_df_post)
```

```
         Id  PostTypeId             CreationDate  Score  ViewCount  \
0        22           1  2011-10-11T20:14:35.147     16     4688.0
1        23           1  2011-10-11T20:15:14.500     38    13491.0
2        25           1  2011-10-11T20:23:09.860     28   523405.0
3        27           1  2011-10-11T20:28:09.197     97    62285.0
4        46           1  2011-10-11T20:57:50.930      9    17162.0
...     ...         ...                      ...    ...        ...
1584  74388           1  2024-02-12T20:32:32.257      9     1857.0
1585  75526           1  2024-03-17T21:37:24.857      4      195.0
1586  75538           1  2024-03-19T04:17:46.123     -1      166.0
1587  75564           1  2024-03-21T11:22:46.497      1      240.0
1588  75581           1  2024-03-23T19:28:04.900      1       65.0

                                                   Body  OwnerUserId  \
0     <p>During World War 2, what government branche...         45.0
1     <p>There are accounts that British Prime Minis...         18.0
2     <p>How many troops (allied and axis) died at N...         16.0
3     <p>Japan and the Soviet Union shared a common ...          9.0
4     <p>I am interested in mapping the pre-war (WWI...         62.0
...                                                 ...          ...
1584  <p>In Hitler’s last will and testament, he ass...      22453.0
1585  <p>Recently I found a Bellingcat article title...      38316.0
1586  <p>Many Germans, even before World War I, beli...      52417.0
1587  <p>There's a <a href="https://cjil.uchicago.ed...      27431.0
1588  <p>I am attempting to do some research on my p...      65371.0

      LastEditorUserId             LastEditDate         LastActivityDate  \
0                961.0  2013-02-08T15:30:36.553  2019-07-31T02:03:50.747
1                 85.0  2012-06-14T17:45:02.680  2019-09-28T15:53:47.793
2              17887.0  2020-04-26T08:21:49.820  2020-04-26T08:21:49.820
3              17887.0  2020-04-17T13:39:52.473  2020-04-17T13:39:52.473
4                 62.0  2015-08-29T23:37:50.827  2017-07-06T21:12:26.617
...                ...                      ...                      ...
1584               NaN                      NaN  2024-02-13T10:25:44.313
1585           38316.0  2024-03-17T22:33:21.500  2024-03-17T22:33:21.500
```

```
1586          52417.0  2024-03-19T17:01:45.413  2024-03-19T17:45:23.620
1587          27431.0  2024-03-21T11:28:04.713  2024-04-06T19:23:31.240
1588              NaN                      NaN  2024-03-23T19:28:04.900

                                                  Title  \
0         German Government branches during World War 2
1     Why did Chamberlain act to appease Hitler lead...
2                         How many troops died on D-day?
3     Why didn't Imperial Japan attack the Soviet Un...
4     Where was the pre-war (ww2) border between Pol...
...                                                 ...
1584  What exactly was the position of ‘Party Minist...
1585  Does this WWII photo show a double-track railway?
1586  Why didn't the promise of the imminent Green R...
1587  When and where did the Allies use starvation t...
1588  Clarification on WW 2 US Army Discharge Papers...

                                                   Tags  AnswerCount  \
0     |20th-century|world-war-two|nazi-germany|germany|          1.0
1                           |world-war-two|20th-century|          8.0
2     |world-war-two|20th-century|dday|ww2-european-...          2.0
3     |world-war-two|20th-century|soviet-union|japan...         12.0
4     |20th-century|world-war-two|germany|poland|geo...          6.0
...                                                 ...          ...
1584  |world-war-two|nazi-germany|germany|hitler|naz...          2.0
1585   |world-war-two|photography|railroads|photograph|          0.0
1586                         |world-war-two|agriculture|          2.0
1587                 |world-war-two|military|starvation|          2.0
1588  |world-war-two|ww2-european-theater|historical...          0.0

      CommentCount ContentLicense
0                0   CC BY-SA 3.0
1                1   CC BY-SA 3.0
2                1   CC BY-SA 3.0
3                9   CC BY-SA 3.0
4                3   CC BY-SA 3.0
...            ...           ...
1584             5   CC BY-SA 4.0
1585             1   CC BY-SA 4.0
1586             5   CC BY-SA 4.0
1587            12   CC BY-SA 4.0
1588             1   CC BY-SA 4.0
```

```
[1589 rows x 15 columns]
```

```
# Again remove html tags for the Body
ww2_df_post["Clean_Body"] =ww2_df_post["Body"].apply(remove_html_tags)
print(ww2_df_post)
```

```
         Id  PostTypeId             CreationDate  Score  ViewCount  \
0        22           1  2011-10-11T20:14:35.147     16     4688.0
1        23           1  2011-10-11T20:15:14.500     38    13491.0
2        25           1  2011-10-11T20:23:09.860     28   523405.0
3        27           1  2011-10-11T20:28:09.197     97    62285.0
4        46           1  2011-10-11T20:57:50.930      9    17162.0
...     ...         ...                      ...    ...        ...
1584  74388           1  2024-02-12T20:32:32.257      9     1857.0
1585  75526           1  2024-03-17T21:37:24.857      4      195.0
1586  75538           1  2024-03-19T04:17:46.123     -1      166.0
1587  75564           1  2024-03-21T11:22:46.497      1      240.0
1588  75581           1  2024-03-23T19:28:04.900      1       65.0

                                                   Body  OwnerUserId  \
0     <p>During World War 2, what government branche...         45.0
1     <p>There are accounts that British Prime Minis...         18.0
2     <p>How many troops (allied and axis) died at N...         16.0
3     <p>Japan and the Soviet Union shared a common ...          9.0
4     <p>I am interested in mapping the pre-war (WWI...         62.0
...                                                 ...          ...
1584  <p>In Hitler’s last will and testament, he ass...      22453.0
1585  <p>Recently I found a Bellingcat article title...      38316.0
1586  <p>Many Germans, even before World War I, beli...      52417.0
1587  <p>There's a <a href="https://cjil.uchicago.ed...      27431.0
1588  <p>I am attempting to do some research on my p...      65371.0

      LastEditorUserId             LastEditDate         LastActivityDate  \
0                961.0  2013-02-08T15:30:36.553  2019-07-31T02:03:50.747
1                 85.0  2012-06-14T17:45:02.680  2019-09-28T15:53:47.793
2              17887.0  2020-04-26T08:21:49.820  2020-04-26T08:21:49.820
3              17887.0  2020-04-17T13:39:52.473  2020-04-17T13:39:52.473
4                 62.0  2015-08-29T23:37:50.827  2017-07-06T21:12:26.617
...                ...                      ...                      ...
1584               NaN                      NaN  2024-02-13T10:25:44.313
1585           38316.0  2024-03-17T22:33:21.500  2024-03-17T22:33:21.500
1586           52417.0  2024-03-19T17:01:45.413  2024-03-19T17:45:23.620
```

```
1587          27431.0  2024-03-21T11:28:04.713  2024-04-06T19:23:31.240
1588              NaN                      NaN  2024-03-23T19:28:04.900

                                                  Title  \
0         German Government branches during World War 2
1     Why did Chamberlain act to appease Hitler lead...
2                       How many troops died on D-day?
3     Why didn't Imperial Japan attack the Soviet Un...
4     Where was the pre-war (ww2) border between Pol...
...                                                 ...
1584  What exactly was the position of ‘Party Minist...
1585  Does this WWII photo show a double-track railway?
1586  Why didn't the promise of the imminent Green R...
1587  When and where did the Allies use starvation t...
1588  Clarification on WW 2 US Army Discharge Papers...

                                                  Tags  AnswerCount  \
0     |20th-century|world-war-two|nazi-germany|germany|          1.0
1                          |world-war-two|20th-century|          8.0
2     |world-war-two|20th-century|dday|ww2-european-...          2.0
3     |world-war-two|20th-century|soviet-union|japan...         12.0
4     |20th-century|world-war-two|germany|poland|geo...          6.0
...                                                 ...          ...
1584  |world-war-two|nazi-germany|germany|hitler|naz...          2.0
1585   |world-war-two|photography|railroads|photograph|          0.0
1586                        |world-war-two|agriculture|          2.0
1587                |world-war-two|military|starvation|          2.0
1588  |world-war-two|ww2-european-theater|historical...          0.0

      CommentCount ContentLicense  \
0                0   CC BY-SA 3.0
1                1   CC BY-SA 3.0
2                1   CC BY-SA 3.0
3                9   CC BY-SA 3.0
4                3   CC BY-SA 3.0
...            ...           ...
1584             5   CC BY-SA 4.0
1585             1   CC BY-SA 4.0
1586             5   CC BY-SA 4.0
1587            12   CC BY-SA 4.0
1588             1   CC BY-SA 4.0

                                             Clean_Body
```

```
0     During World War 2, what government branches, ...
1     There are accounts that British Prime Minister...
2     How many troops (allied and axis) died at Norm...
3     Japan and the Soviet Union shared a common bor...
4     I am interested in mapping the pre-war (WWII) ...
...                                                 ...
1584  In Hitler's last will and testament, he assign...
1585  Recently I found a Bellingcat article titled S...
1586  Many Germans, even before World War I, believe...
1587  There's a paper that claims in relation to the...
1588  I am attempting to do some research on my pate...

[1589 rows x 16 columns]
```

```
# Remove special keywords for the Cleanned Body and the Title
ww2_df_post["Clean_Body"] =
 ↪  ww2_df_post["Clean_Body"].apply(remove_special_keywords)
ww2_df_post["Title"] =  ww2_df_post["Title"].apply(remove_special_keywords)
print(ww2_df_post.head())
```

```
   Id  PostTypeId             CreationDate  Score  ViewCount  \
0  22           1  2011-10-11T20:14:35.147     16     4688.0
1  23           1  2011-10-11T20:15:14.500     38    13491.0
2  25           1  2011-10-11T20:23:09.860     28   523405.0
3  27           1  2011-10-11T20:28:09.197     97    62285.0
4  46           1  2011-10-11T20:57:50.930      9    17162.0

                                                Body  OwnerUserId  \
0  <p>During World War 2, what government branche...         45.0
1  <p>There are accounts that British Prime Minis...         18.0
2  <p>How many troops (allied and axis) died at N...         16.0
3  <p>Japan and the Soviet Union shared a common ...          9.0
4  <p>I am interested in mapping the pre-war (WWI...         62.0

   LastEditorUserId             LastEditDate         LastActivityDate  \
0             961.0  2013-02-08T15:30:36.553  2019-07-31T02:03:50.747
1              85.0  2012-06-14T17:45:02.680  2019-09-28T15:53:47.793
2           17887.0  2020-04-26T08:21:49.820  2020-04-26T08:21:49.820
3           17887.0  2020-04-17T13:39:52.473  2020-04-17T13:39:52.473
4              62.0  2015-08-29T23:37:50.827  2017-07-06T21:12:26.617

                                              Title  \
```

```
0      German Government branches during World War 2
1  Why did Chamberlain act to appease Hitler lead...
2                    How many troops died on D-day?
3  Why didn't Imperial Japan attack the Soviet Un...
4  Where was the pre-war (ww2) border between Pol...

                                              Tags  AnswerCount  \
0  |20th-century|world-war-two|nazi-germany|germany|          1.0
1                      |world-war-two|20th-century|          8.0
2  |world-war-two|20th-century|dday|ww2-european-...          2.0
3  |world-war-two|20th-century|soviet-union|japan...         12.0
4  |20th-century|world-war-two|germany|poland|geo...          6.0

   CommentCount ContentLicense  \
0             0   CC BY-SA 3.0
1             1   CC BY-SA 3.0
2             1   CC BY-SA 3.0
3             9   CC BY-SA 3.0
4             3   CC BY-SA 3.0

                                          Clean_Body
0  During World War 2, what government branches, ...
1  There are accounts that British Prime Minister...
2  How many troops (allied and axis) died at Norm...
3  Japan and the Soviet Union shared a common bor...
4  I am interested in mapping the pre-war (WWII) ...
```

```
# Narrow down our dataframe to make it contains the Title the body and the
 ↪  PostTypeId
ww2_df_post = ww2_df_post[["Id","Clean_Body","Title"]]
print(ww2_df_post)
```

```
         Id                                         Clean_Body  \
0        22  During World War 2, what government branches, ...
1        23  There are accounts that British Prime Minister...
2        25  How many troops (allied and axis) died at Norm...
3        27  Japan and the Soviet Union shared a common bor...
4        46  I am interested in mapping the pre-war (WWII) ...
...     ...                                                ...
1584  74388  In Hitler’s last will and testament, he assign...
1585  75526  Recently I found a Bellingcat article titled S...
1586  75538  Many Germans, even before World War I, believe...
```

```
1587  75564  There's a paper that claims in relation to the...
1588  75581  I am attempting to do some research on my pate...

                                                  Title
0         German Government branches during World War 2
1     Why did Chamberlain act to appease Hitler lead...
2                         How many troops died on D-day?
3     Why didn't Imperial Japan attack the Soviet Un...
4     Where was the pre-war (ww2) border between Pol...
...                                                 ...
1584  What exactly was the position of ‘Party Minist...
1585  Does this WWII photo show a double-track railway?
1586  Why didn't the promise of the imminent Green R...
1587  When and where did the Allies use starvation t...
1588  Clarification on WW 2 US Army Discharge Papers...

[1589 rows x 3 columns]
```

```
# Let's have a look at the newly cleanned dataframe to check if there are
#any null values in our title and body
print(ww2_df_post.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1589 entries, 0 to 1588
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Id          1589 non-null   int64
 1   Clean_Body  1589 non-null   object
 2   Title       1589 non-null   object
dtypes: int64(1), object(2)
memory usage: 37.4+ KB
None
```

Let’s determine the polarity score for the titles and their corresponding bodies.

```
# Body Score
ww2_df_post["Body Score"] =
 ↪  ww2_df_post["Clean_Body"].apply(polarity_scores_roberta)
print(ww2_df_post.head())
```

```
   Id                                         Clean_Body  \
0  22  During World War 2, what government branches, ...
1  23  There are accounts that British Prime Minister...
2  25  How many troops (allied and axis) died at Norm...
3  27  Japan and the Soviet Union shared a common bor...
4  46  I am interested in mapping the pre-war (WWII) ...

                                          Title  Body Score
0     German Government branches during World War 2   -0.462669
1  Why did Chamberlain act to appease Hitler lead...   -0.590581
2                    How many troops died on D-day?   -0.766205
3  Why didn't Imperial Japan attack the Soviet Un...   -0.335204
4  Where was the pre-war (ww2) border between Pol...    0.063665
```

```
# Calculate the polarity score of the title
ww2_df_post["Title Score"] =
 ↪  ww2_df_post["Title"].apply(polarity_scores_roberta)
print(ww2_df_post.head())
```

```
   Id                                         Clean_Body  \
0  22  During World War 2, what government branches, ...
1  23  There are accounts that British Prime Minister...
2  25  How many troops (allied and axis) died at Norm...
3  27  Japan and the Soviet Union shared a common bor...
4  46  I am interested in mapping the pre-war (WWII) ...

                                          Title  Body Score  Title Score
0     German Government branches during World War 2   -0.462669    -0.440161
1  Why did Chamberlain act to appease Hitler lead...   -0.590581    -0.573242
2                    How many troops died on D-day?   -0.766205    -0.735738
3  Why didn't Imperial Japan attack the Soviet Un...   -0.335204    -0.421336
4  Where was the pre-war (ww2) border between Pol...    0.063665    -0.135447
```

Once we've computed the polarity scores for the titles and bodies, our next step involves generating columns that depict the length of the text in both titles and their respective bodies.

```
# Titles
ww2_df_post["Length Title"] = ww2_df_post["Title"].apply(lambda text :
 ↪  len(text))

# Bodies
```

```
ww2_df_post["Length Body"] = ww2_df_post["Clean_Body"].apply(lambda text:
 ↪  len(text))
print(ww2_df_post)
```

```
         Id                                         Clean_Body  \
0        22  During World War 2, what government branches, ...
1        23  There are accounts that British Prime Minister...
2        25  How many troops (allied and axis) died at Norm...
3        27  Japan and the Soviet Union shared a common bor...
4        46  I am interested in mapping the pre-war (WWII) ...
...     ...                                                ...
1584  74388  In Hitler’s last will and testament, he assign...
1585  75526  Recently I found a Bellingcat article titled S...
1586  75538  Many Germans, even before World War I, believe...
1587  75564  There's a paper that claims in relation to the...
1588  75581  I am attempting to do some research on my pate...

                                                  Title  Body Score  \
0         German Government branches during World War 2   -0.462669
1     Why did Chamberlain act to appease Hitler lead...   -0.590581
2                         How many troops died on D-day?  -0.766205
3     Why didn't Imperial Japan attack the Soviet Un...   -0.335204
4     Where was the pre-war (ww2) border between Pol...    0.063665
...                                                 ...         ...
1584  What exactly was the position of ‘Party Minist...   -0.417163
1585  Does this WWII photo show a double-track railway?   -0.192908
1586  Why didn't the promise of the imminent Green R...   -0.270615
1587  When and where did the Allies use starvation t...   -0.293305
1588  Clarification on WW 2 US Army Discharge Papers...   -0.487920

      Title Score  Length Title  Length Body
0       -0.440161            45          179
1       -0.573242            84          444
2       -0.735738            30          148
3       -0.421336            69          378
4       -0.135447            62          281
...           ...           ...          ...
1584    -0.328712            94          656
1585    -0.033086            49         1874
1586    -0.433834           119         2305
1587    -0.349883            60          567
```

```
1588    -0.095875            63         1313

[1589 rows x 7 columns]
```

Next, we'll generate word clouds to identify the most commonly appeared words in the bodies of posts categorized under different sentiments: negative, neutral, and positive.

Initially, we'll create a new dataframe that includes bodies along with their polarity scores. Then, we'll classify the sentiments of these bodies based on their polarity scores.

```
# Extract appropriate columns
word_cloud_df = ww2_df_post[["Clean_Body","Body Score"]]
print(word_cloud_df)
```

```
                                             Clean_Body  Body Score
0     During World War 2, what government branches, ...   -0.462669
1     There are accounts that British Prime Minister...   -0.590581
2     How many troops (allied and axis) died at Norm...   -0.766205
3     Japan and the Soviet Union shared a common bor...   -0.335204
4     I am interested in mapping the pre-war (WWII) ...    0.063665
...                                                 ...         ...
1584  In Hitler's last will and testament, he assign...   -0.417163
1585  Recently I found a Bellingcat article titled S...   -0.192908
1586  Many Germans, even before World War I, believe...   -0.270615
1587  There's a paper that claims in relation to the...   -0.293305
1588  I am attempting to do some research on my pate...   -0.487920

[1589 rows x 2 columns]
```

```
# Funtion to return the sentiment of the text based on its sentiment score
def classify_sentiment(score):
    if score < -0.2:
        return "Negative"
    elif score > 0.2:
        return "Positive"
    else:
        return "Neutral"
```

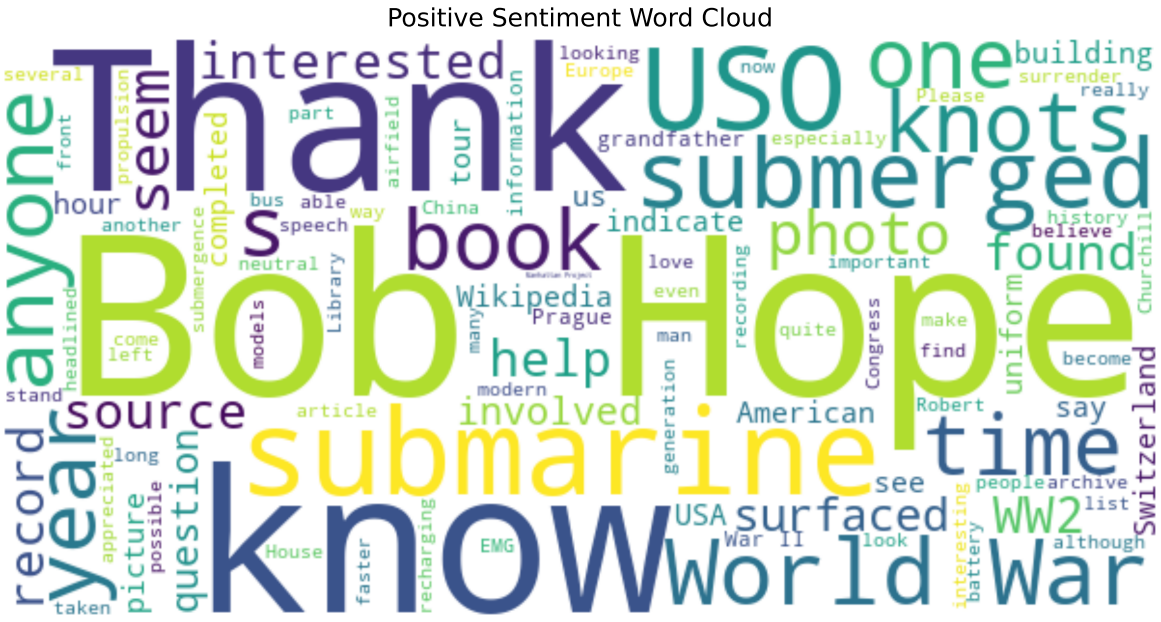
```
word_cloud_df["Sentiment"] = word_cloud_df["Body
 ↪  Score"].apply(classify_sentiment)
print(word_cloud_df)
```

```
                                             Clean_Body  Body Score Sentiment
0     During World War 2, what government branches, ...   -0.462669  Negative
1     There are accounts that British Prime Minister...   -0.590581  Negative
2     How many troops (allied and axis) died at Norm...   -0.766205  Negative
3     Japan and the Soviet Union shared a common bor...   -0.335204  Negative
4     I am interested in mapping the pre-war (WWII) ...    0.063665   Neutral
...                                                 ...         ...       ...
1584  In Hitler’s last will and testament, he assign...   -0.417163  Negative
1585  Recently I found a Bellingcat article titled S...   -0.192908   Neutral
1586  Many Germans, even before World War I, believe...   -0.270615  Negative
1587  There's a paper that claims in relation to the...   -0.293305  Negative
1588  I am attempting to do some research on my pate...   -0.487920  Negative

[1589 rows x 3 columns]

C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\1049559315.py:1: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

```
# Function to generate word cloud
def generate_word_cloud(text, title):
    wordcloud = WordCloud(width=600, height=300, background_color="white",
 ↪  prefer_horizontal=0.9).generate(text)
    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.title(title)
    plt.axis('off')
    plt.show()
```

```
# Generate word clouds for each sentiment group
# Positive group
positive_posts_text = " ".join(word_cloud_df[word_cloud_df["Sentiment"] ==
 ↪  "Positive"]["Clean_Body"])
```

```
generate_word_cloud(positive_posts_text, "Positive Sentiment Word Cloud")

# Neutral group
neutral_posts_text = " ".join(word_cloud_df[word_cloud_df["Sentiment"] ==
 ↪  "Neutral"]["Clean_Body"])
generate_word_cloud(neutral_posts_text, "Neutral Sentiment Word Cloud")

# Negative group
negative_posts_text = " ".join(word_cloud_df[word_cloud_df["Sentiment"] ==
 ↪  "Negative"]["Clean_Body"])
generate_word_cloud(negative_posts_text, "Negative Sentiment Word Cloud")
```



Neutral Sentiment Word Cloud

Negative Sentiment Word Cloud

From the word clouds, we observe that the size of words corresponds to their frequency in the text. Larger words indicate higher occurrence. In the positive sentiment group, prominent words like "Thank" and "Hope" signify positive emotions. These words, expressing gratitude and optimism, are more prevalent in

texts with positive sentiments. Thus, analyzing these word clouds offers insights into the prevailing positive themes and emotions conveyed in the text.

In the neutral sentiment word cloud, we notice that country names and general terms like "World War" are more prominent. This makes sense, as these words convey general information without expressing specific emotions.

Interestingly, in the negative sentiment word cloud, words like "War" are prominent, reflecting themes of fatality and death, which clearly indicate negative emotions. Additionally, country names such as Germany and Japan, along with nationalities like German and Japanese, are frequently mentioned. This is likely because these countries were associated with the Nazis, thus evoking negative sentiments.

Next, we'll analyze the comments by calculating their sentiment scores.

```
# Get the Ids of the posts related to World War 2
ww2_post_ids = ww2_df_post["Id"]

# Filter out comments related to World War 2 posts
ww2_cmt_df = cmt_df[cmt_df["PostId"].isin(ww2_post_ids)].reset_index(drop =
 ↪  True)
print(ww2_cmt_df)
```

```
          Id  PostId  Score  \
0         83      46      3
1        122      64      9
2        228     215     13
3        327     199      1
4        332     215     12
...      ...     ...    ...
7223  242695   75564      1
7224  242697   75564      0
7225  242713   75564      0
7226  242746   75581      2
7227  242860   74186      0

                                                   Text  \
0     Please keep in mind that Poland also had a bor...
1     Note that the US was already fairly involved i...
2     As a simple non-researched answer (I will atte...
```

```
3     I must admit, this is the most interesting que...
4           As an American, I have never heard of this.
...                                                   ...
7223  I know you said "Allies" but let's not forget ...
7224  And had there been no atomic bombs and no inva...
7225  In Europe both sides used naval blockades in w...
7226  You are looking at the War Department general ...
7227  @Steve: I had a similar idea, but I know that ...

                 CreationDate   UserId
0     2011-10-11T21:18:36.073     26.0
1     2011-10-12T07:01:48.573     22.0
2     2011-10-13T01:16:32.633     12.0
3     2011-10-13T22:28:57.897     16.0
4     2011-10-13T22:42:15.993     54.0
...                       ...      ...
7223  2024-03-22T08:52:16.947   1372.0
7224  2024-03-22T12:59:04.500  33565.0
7225  2024-03-23T08:17:31.607   3164.0
7226  2024-03-24T14:34:48.410  33565.0
7227  2024-03-29T06:26:11.207    994.0

[7228 rows x 6 columns]
```

```
# Filter out special keywords in the comments text
ww2_cmt_df = ww2_cmt_df[["PostId","Text"]]
ww2_cmt_df["Text"] = ww2_cmt_df["Text"].apply(remove_special_keywords)
print(ww2_cmt_df)
```

```
      PostId                                               Text
0         46  Please keep in mind that Poland also had a bor...
1         64  Note that the US was already fairly involved i...
2        215  As a simple non-researched answer (I will atte...
3        199  I must admit, this is the most interesting que...
4        215        As an American, I have never heard of this.
...      ...                                                ...
7223   75564  I know you said "Allies" but let's not forget ...
7224   75564  And had there been no atomic bombs and no inva...
7225   75564  In Europe both sides used naval blockades in w...
7226   75581  You are looking at the War Department general ...
7227   74186  Steve: I had a similar idea, but I know that d...
```

```
[7228 rows x 2 columns]
```

```
# Calculate the polarity score of the filtered comments
ww2_cmt_df["Comment Score"] =
 ↪  ww2_cmt_df["Text"].apply(polarity_scores_roberta)
print(ww2_cmt_df.head())
```

```
   PostId                                               Text  Comment Score
0      46  Please keep in mind that Poland also had a bor...      -0.284147
1      64  Note that the US was already fairly involved i...      -0.206081
2     215  As a simple non-researched answer (I will atte...      -0.478339
3     199  I must admit, this is the most interesting que...       0.954065
4     215        As an American, I have never heard of this.      -0.563590
```

```
# Calcualte the Average polarity score for each post
# Since one post can have many comments so we need to calculate the mean
 ↪  values
avg_ww2_cmt = ww2_cmt_df.groupby("PostId")[["Comment
 ↪  Score"]].mean().reset_index()
print(avg_ww2_cmt)
```

```
      PostId  Comment Score
0         23       0.977588
1         25       0.934179
2         27      -0.462330
3         46       0.533650
4         64      -0.383542
...      ...            ...
1369   74388       0.198546
1370   75526      -0.594684
1371   75538       0.314779
1372   75564      -0.465143
1373   75581      -0.369456

[1374 rows x 2 columns]
```

Let's have a look again about the two dataframes we've created—one for posts and one for comments.

```
print("Dataframe represents the Post Sentiment:")
print(ww2_df_post)

print("Dataframe represent the Comment Sentiment:")
print(avg_ww2_cmt)
```

```
Dataframe represents the Post Sentiment:
         Id                                         Clean_Body  \
0        22  During World War 2, what government branches, ...
1        23  There are accounts that British Prime Minister...
2        25  How many troops (allied and axis) died at Norm...
3        27  Japan and the Soviet Union shared a common bor...
4        46  I am interested in mapping the pre-war (WWII) ...
...     ...                                                ...
1584  74388  In Hitler’s last will and testament, he assign...
1585  75526  Recently I found a Bellingcat article titled S...
1586  75538  Many Germans, even before World War I, believe...
1587  75564  There's a paper that claims in relation to the...
1588  75581  I am attempting to do some research on my pate...

                                                  Title  Body Score  \
0         German Government branches during World War 2   -0.462669
1     Why did Chamberlain act to appease Hitler lead...   -0.590581
2                         How many troops died on D-day?   -0.766205
3     Why didn't Imperial Japan attack the Soviet Un...   -0.335204
4     Where was the pre-war (ww2) border between Pol...    0.063665
...                                                 ...         ...
1584  What exactly was the position of ‘Party Minist...   -0.417163
1585  Does this WWII photo show a double-track railway?   -0.192908
1586  Why didn't the promise of the imminent Green R...   -0.270615
1587  When and where did the Allies use starvation t...   -0.293305
1588  Clarification on WW 2 US Army Discharge Papers...   -0.487920

      Title Score  Length Title  Length Body
0       -0.440161            45          179
1       -0.573242            84          444
2       -0.735738            30          148
3       -0.421336            69          378
4       -0.135447            62          281
...           ...           ...          ...
1584    -0.328712            94          656
```

```
1585    -0.033086            49         1874
1586    -0.433834           119         2305
1587    -0.349883            60          567
1588    -0.095875            63         1313

[1589 rows x 7 columns]
Dataframe represent the Comment Sentiment:
      PostId  Comment Score
0         23       0.977588
1         25       0.934179
2         27      -0.462330
3         46       0.533650
4         64      -0.383542
...      ...            ...
1369   74388       0.198546
1370   75526      -0.594684
1371   75538       0.314779
1372   75564      -0.465143
1373   75581      -0.369456

[1374 rows x 2 columns]
```

Next, let's combine the two dataframes by merging them on the PostId column.

```
# First we need to rename our post dataframe to make it the same name Id to
 ↪ be
# able to merge the 2 dataframes
ww2_df_post.rename(columns = {"Id":"PostId"},inplace = True)
print(ww2_df_post)
```

```
      PostId                                         Clean_Body  \
0         22  During World War 2, what government branches, ...
1         23  There are accounts that British Prime Minister...
2         25  How many troops (allied and axis) died at Norm...
3         27  Japan and the Soviet Union shared a common bor...
4         46  I am interested in mapping the pre-war (WWII) ...
...      ...                                                ...
1584   74388  In Hitler's last will and testament, he assign...
1585   75526  Recently I found a Bellingcat article titled S...
1586   75538  Many Germans, even before World War I, believe...
1587   75564  There's a paper that claims in relation to the...
```

```
1588   75581  I am attempting to do some research on my pate...

                                                  Title  Body Score  \
0          German Government branches during World War 2   -0.462669
1     Why did Chamberlain act to appease Hitler lead...   -0.590581
2                         How many troops died on D-day?   -0.766205
3     Why didn't Imperial Japan attack the Soviet Un...   -0.335204
4     Where was the pre-war (ww2) border between Pol...    0.063665
...                                                 ...         ...
1584  What exactly was the position of ‘Party Minist...   -0.417163
1585  Does this WWII photo show a double-track railway?   -0.192908
1586  Why didn't the promise of the imminent Green R...   -0.270615
1587  When and where did the Allies use starvation t...   -0.293305
1588  Clarification on WW 2 US Army Discharge Papers...   -0.487920

      Title Score  Length Title  Length Body
0       -0.440161            45          179
1       -0.573242            84          444
2       -0.735738            30          148
3       -0.421336            69          378
4       -0.135447            62          281
...           ...           ...          ...
1584    -0.328712            94          656
1585    -0.033086            49         1874
1586    -0.433834           119         2305
1587    -0.349883            60          567
1588    -0.095875            63         1313

[1589 rows x 7 columns]
```

```
complete_sentiment_df = pd.merge(ww2_df_post, avg_ww2_cmt , on = "PostId",
 ↪  how = "left" )
print(complete_sentiment_df)
```

```
      PostId                                         Clean_Body  \
0         22  During World War 2, what government branches, ...
1         23  There are accounts that British Prime Minister...
2         25  How many troops (allied and axis) died at Norm...
3         27  Japan and the Soviet Union shared a common bor...
4         46  I am interested in mapping the pre-war (WWII) ...
...      ...                                                ...
1584   74388  In Hitler’s last will and testament, he assign...
```

```
1585   75526  Recently I found a Bellingcat article titled S...
1586   75538  Many Germans, even before World War I, believe...
1587   75564  There's a paper that claims in relation to the...
1588   75581  I am attempting to do some research on my pate...

                                                  Title  Body Score  \
0         German Government branches during World War 2   -0.462669
1     Why did Chamberlain act to appease Hitler lead...   -0.590581
2                         How many troops died on D-day?   -0.766205
3     Why didn't Imperial Japan attack the Soviet Un...   -0.335204
4     Where was the pre-war (ww2) border between Pol...    0.063665
...                                                 ...         ...
1584  What exactly was the position of ‘Party Minist...   -0.417163
1585  Does this WWII photo show a double-track railway?   -0.192908
1586  Why didn't the promise of the imminent Green R...   -0.270615
1587  When and where did the Allies use starvation t...   -0.293305
1588  Clarification on WW 2 US Army Discharge Papers...   -0.487920

      Title Score  Length Title  Length Body  Comment Score
0       -0.440161            45          179            NaN
1       -0.573242            84          444       0.977588
2       -0.735738            30          148       0.934179
3       -0.421336            69          378      -0.462330
4       -0.135447            62          281       0.533650
...           ...           ...          ...            ...
1584    -0.328712            94          656       0.198546
1585    -0.033086            49         1874      -0.594684
1586    -0.433834           119         2305       0.314779
1587    -0.349883            60          567      -0.465143
1588    -0.095875            63         1313      -0.369456

[1589 rows x 8 columns]
```

```
# Check the general information of the merged dataframe as well as checking
 ↪  null values
print("General Information:")
print(complete_sentiment_df.describe())

print("Checking Null Values:")
print(complete_sentiment_df.info())
```

```
General Information:
```

```
              PostId   Body Score  Title Score  Length Title  Length Body  \
count   1589.000000  1589.000000  1589.000000   1589.000000  1589.000000
mean   39067.859031    -0.359960    -0.298876     71.979232   847.863436
std    20958.573784     0.231066     0.210853     24.740929   820.448687
min       22.000000    -0.963766    -0.919497     15.000000    32.000000
25%    22763.000000    -0.504959    -0.450360     55.000000   368.000000
50%    41095.000000    -0.372045    -0.264101     69.000000   617.000000
75%    56081.000000    -0.229769    -0.128669     87.000000  1050.000000
max    75581.000000     0.894528     0.512722    150.000000  8850.000000

       Comment Score
count    1374.000000
mean       -0.204779
std         0.257666
min        -0.958195
25%        -0.372827
50%        -0.232889
75%        -0.081376
max         0.986530
Checking Null Values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1589 entries, 0 to 1588
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   PostId         1589 non-null   int64
 1   Clean_Body     1589 non-null   object
 2   Title          1589 non-null   object
 3   Body Score     1589 non-null   float64
 4   Title Score    1589 non-null   float64
 5   Length Title   1589 non-null   int64
 6   Length Body    1589 non-null   int64
 7   Comment Score  1374 non-null   float64
dtypes: float64(3), int64(3), object(2)
memory usage: 99.4+ KB
None
```

First, we notice that the Comment Score column contains some null values. Instead of simply deleting these rows, we'll use a random sampling technique to replace the null values.

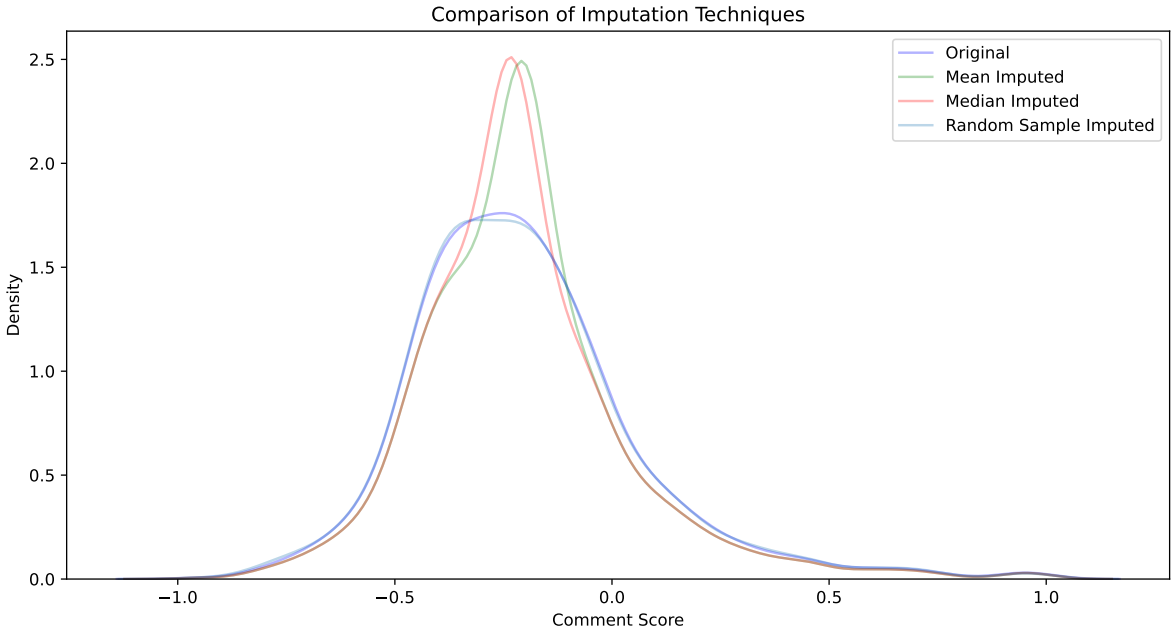
But we will firstly check the performance of it compared to other imputation techniques such as the mean and median imputation.

```
# Mean values
mean_value = complete_sentiment_df["Comment Score"].mean()
mean_impt = complete_sentiment_df["Comment Score"].fillna(mean_value)

# Median value
median = complete_sentiment_df["Comment Score"].median()
median_impt = complete_sentiment_df["Comment Score"].fillna(median)


# Random sample imputation
def random_sample_imputation(df):
    cols_with_missing_values = df.columns[df.isna().any()].tolist()

    for var in cols_with_missing_values:
        # extract a random sample
        random_sample_df = df[var].dropna().sample(df[var].isnull().sum(),
 ↪  replace=True, random_state=0)
        # re-index the randomly extracted sample
        random_sample_df.index = df[df[var].isnull()].index
        # replace the NA
        df.loc[df[var].isnull(), var] = random_sample_df.values

    return df

random_impt = random_sample_imputation(complete_sentiment_df[["Comment
 ↪  Score"]])
```

```
C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\1487536339.py:20: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

```
# Plotting part

# Original data
original_data = complete_sentiment_df["Comment Score"]

# Plotting the distributions
plt.figure(figsize=(12, 6))
```

```
# Original data
sns.kdeplot(original_data, label="Original", color="blue", alpha=0.3)

# Mean imputed data
sns.kdeplot(mean_impt, label="Mean Imputed", color="green",  alpha=0.3)

# Median imputed data
sns.kdeplot(median_impt, label="Median Imputed", color="red",alpha=0.3)

# Random sample imputed data
sns.kdeplot(random_impt, label="Random Sample Imputed",
 ↪  color="purple",alpha=0.3)

plt.title("Comparison of Imputation Techniques")
plt.xlabel("Comment Score")
plt.ylabel("Density")
plt.legend()
plt.show()
```



Based on the distribution of the data generated by different imputation techniques, we observe that the random sample imputation technique preserves the original data's distribution most accurately.

```
# Filling NA values
complete_sentiment_df["Comment Score"] =
 ↪  random_sample_imputation(complete_sentiment_df[["Comment Score"]])
```

```
C:\Users\thinh\AppData\Local\Temp\ipykernel_26212\1487536339.py:20: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide
```

```
# Check the Null values again
print(complete_sentiment_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1589 entries, 0 to 1588
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   PostId         1589 non-null   int64
 1   Clean_Body     1589 non-null   object
 2   Title          1589 non-null   object
 3   Body Score     1589 non-null   float64
 4   Title Score    1589 non-null   float64
 5   Length Title   1589 non-null   int64
 6   Length Body    1589 non-null   int64
 7   Comment Score  1589 non-null   float64
dtypes: float64(3), int64(3), object(2)
memory usage: 99.4+ KB
None
```

Another important point is that some columns, like the length of the body and titles, have a wide range of values. Therefore, we need to standardize our dataframe to ensure the values are on a similar scale.

```
# Simply remove our unnecessary columns
complete_sentiment_df.drop(columns = ["PostId","Clean_Body","Title"],inplace
 ↪  = True)
print(complete_sentiment_df)
```

```
      Body Score  Title Score  Length Title  Length Body  Comment Score
0      -0.462669    -0.440161            45          179      -0.031729
1      -0.590581    -0.573242            84          444       0.977588
2      -0.766205    -0.735738            30          148       0.934179
3      -0.335204    -0.421336            69          378      -0.462330
4       0.063665    -0.135447            62          281       0.533650
...          ...          ...           ...          ...            ...
1584   -0.417163    -0.328712            94          656       0.198546
1585   -0.192908    -0.033086            49         1874      -0.594684
1586   -0.270615    -0.433834           119         2305       0.314779
1587   -0.293305    -0.349883            60          567      -0.465143
1588   -0.487920    -0.095875            63         1313      -0.369456

[1589 rows x 5 columns]
```

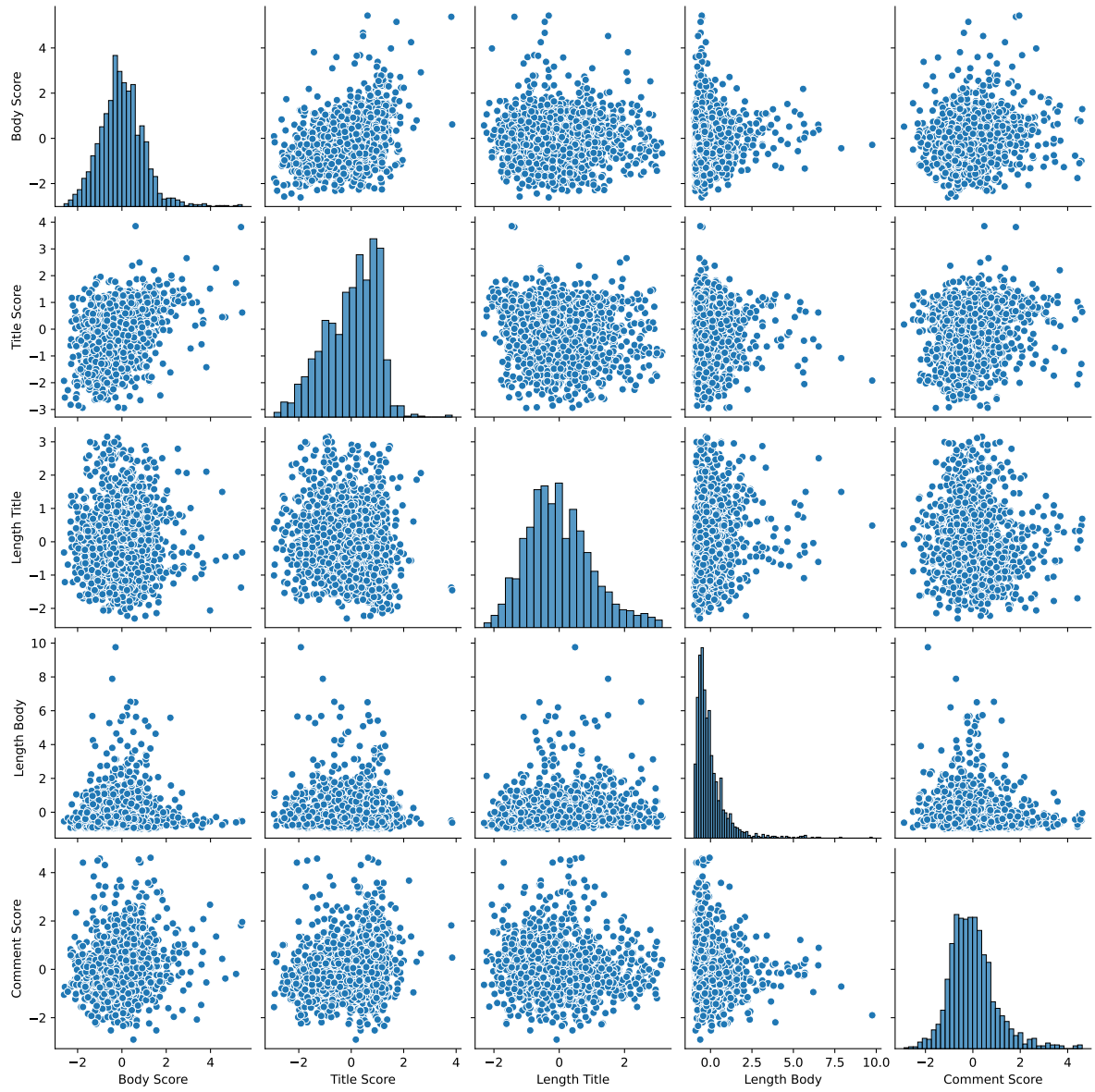
```
# Initialize the StandardScaler
scaler = StandardScaler()

# Apply the scaler to the selected columns
std_df = scaler.fit_transform(complete_sentiment_df)
std_df = pd.DataFrame(std_df,columns = complete_sentiment_df.columns)
print(std_df)
```

```
      Body Score  Title Score  Length Title  Length Body  Comment Score
0      -0.444645    -0.670276     -1.090813    -0.815498       0.677590
1      -0.998393    -1.301630      0.486019    -0.492402       4.580185
2      -1.758692    -2.072532     -1.697287    -0.853294       4.412341
3       0.107169    -0.580968     -0.120455    -0.572871      -0.987355
4       1.833929     0.775323     -0.403476    -0.691136       2.863669
...          ...          ...           ...          ...            ...
1584   -0.247643    -0.141549      0.890334    -0.233925       1.567965
1585    0.723189     1.260941     -0.929087     1.251095      -1.499113
1586    0.386784    -0.640261      1.901124     1.776583       2.017387
1587    0.288555    -0.241988     -0.484339    -0.342437      -0.998235
1588   -0.553958     0.963059     -0.363045     0.567108      -0.628252

[1589 rows x 5 columns]
```

Let's examine the relationships between variables as well as the distribution of them by utilizing the pairplot function.

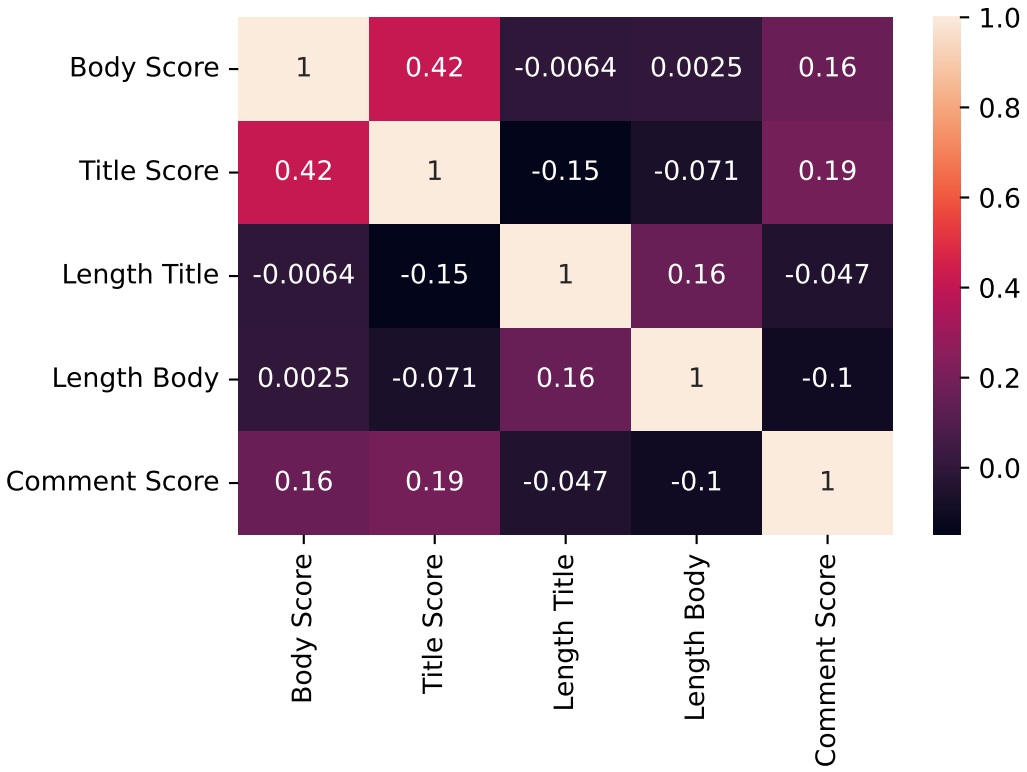
```
sns.pairplot(std_df)
```

```
C:\Users\thinh\Documents\ANACONDA\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning:

The figure layout has changed to tight
```



Now ,let's explore the relationship between variables by calculating the coefficient correlation.

```
sns.heatmap(std_df.corr(),annot = True)
```



Upon initial observation, it appears that there is a correlation between the sentiment score of the title and its associated body. However, there doesn't seem to be any noticeable correlation between the other variables in the dataset.

To strengthen our assertion, We'll construct a basic linear regression model to assess the relationship between variables. Given that posts are typically created before comments are made, we'll designate all attributes of the post as predictors, with the sentiment score of the comments serving as our target variable.

```
X = std_df.drop(columns = ["Comment Score"]) # predictor variables
y = std_df[["Comment Score"]]  # target variable


X = sm.add_constant(X)

# Create a linear regression model
model = sm.OLS(y, X)
```

```python
# Fit the model to the data
results = model.fit()

# print the summary of the model
print(results.summary())
```

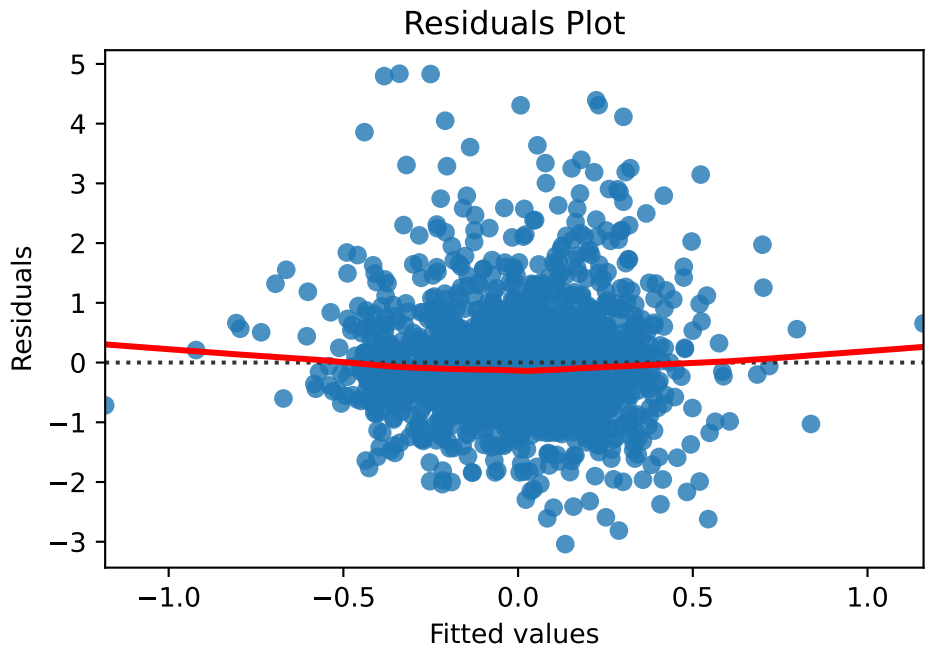
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Comment Score   R-squared:                       0.054
Model:                            OLS   Adj. R-squared:                  0.052
Method:                 Least Squares   F-statistic:                     22.63
Date:                Wed, 22 May 2024   Prob (F-statistic):           3.37e-18
Time:                        21:02:18   Log-Likelihood:                -2210.5
No. Observations:                1589   AIC:                             4431.
Df Residuals:                    1584   BIC:                             4458.
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         4.857e-17      0.024   1.99e-15      1.000      -0.048       0.048
Body Score       0.1035      0.027      3.844      0.000       0.051       0.156
Title Score      0.1427      0.027      5.234      0.000       0.089       0.196
Length Title    -0.0108      0.025     -0.432      0.665      -0.060       0.038
Length Body     -0.0894      0.025     -3.607      0.000      -0.138      -0.041
==============================================================================
Omnibus:                      299.935   Durbin-Watson:                   1.958
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              791.054
Skew:                           0.997   Prob(JB):                    1.68e-172
Kurtosis:                       5.823   Cond. No.                         1.62
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

From the summary of our linear regression model, we can deduce that the sentiment score of the body, the sentiment score of the title, and the length of the body of each post have a statistically significant impact on our target variable, which is the sentiment score of the comments for that post. However, it appears that the length of the header of each post does

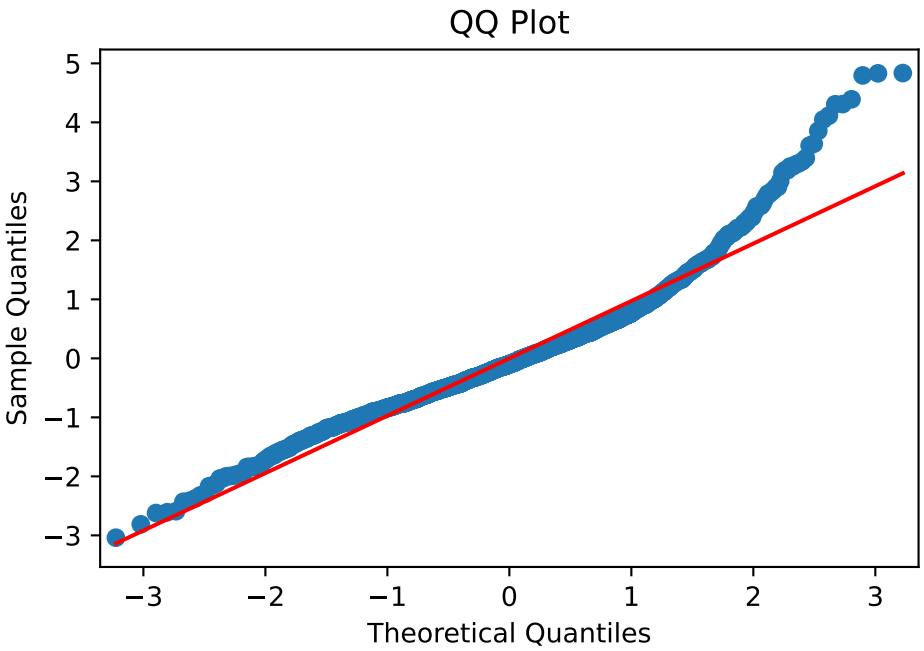
not have a statistically significant effect on our target variable (the p-value is much larger than 5%).

We'll create a residuals plot to assess the assumptions of our linear regression model.

```
# Residuals Plot

residuals = results.resid

# Step 2: Plot the residuals against the predicted values
sns.residplot(x=results.fittedvalues, y=residuals, lowess=True,
 ↪  line_kws={'color': 'red'})
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residuals Plot')
plt.show()
```




```
# QQ plot
qqplot(residuals, line='s')
plt.title('QQ Plot')
plt.show()
```



As observed in our residual plot, most data points are concentrated in the center, and the curved lowess line suggests a potential non-linear relationship between our predictors and the response variable that our linear model is not capturing.

Regarding the QQ plot, we notice that while the data points initially align with the linear line, they increasingly diverge as they progress. This indicates that the residuals do not follow a normal distribution.

From our OLS summary, we see that the R-squared value is relatively low at just 5.2%, indicating low accuracy. This suggests that while the sentiment scores of the title and body, along with their associated lengths, do have some relationship with the response variable, they are not strong enough to be considered robust predictors.

Interestingly, the more sentiments expressed in the titles, the more sentiments will be reflected in the bodies of the posts.

Regular expressions represent a cornerstone in text data processing, offering versatile capabilities for extracting and manipulating information from raw text. With regular expressions, we can delve into the intricacies of unstructured data, pinpointing specific patterns or phrases that hold significance within a vast sea of text. For instance, in our case , in sentiment analysis, where understanding the

emotional tone of text is paramount, regular expressions enable us to identify sentiment-bearing words or expressions, thus laying the groundwork for insightful analysis. Moreover, in tasks like named entity recognition or information extraction, regular expressions serve as invaluable tools for identifying and extracting entities such as locations, names, dates, or numerical data, contributing to the creation of structured datasets from unstructured text sources.

Furthermore, regular expressions play a crucial role in data preprocessing, particularly in the realm of natural language processing (NLP). By employing regular expressions, we can effectively clean text data by removing noise such as stop words, punctuation, or special characters that might obscure the underlying signal. This data cleaning process is essential for enhancing the quality and reliability of subsequent analyses, ensuring that our models are trained on clean, relevant data. Additionally, regular expressions facilitate the standardization of text data by enforcing consistent formatting conventions, thereby streamlining downstream tasks such as text classification, clustering, or information retrieval. In essence, the power of regular expressions lies not only in their ability to extract valuable insights from text but also in their capacity to transform raw text into structured, actionable data that fuels advanced analytics and decision-making processes.

In summary, while regular expressions provide powerful tools for analyzing text data, their use raises important considerations regarding data privacy and ethics. These concerns include the inadvertent disclosure of sensitive information during sentiment analysis or named entity recognition, the potential bias introduced by indiscriminate data cleaning, the homogenization of diverse voices through standardization, and the risk of reidentification from structured datasets. To address these issues, practitioners must prioritize ethical principles such as data minimization, anonymization, and transparency to ensure responsible and ethical data analysis practices.