



ALGO ALCHEMIST

PROJECT 1: TOPIC MODELLING

TEAM MEMBERS:

- 1. ADITHYA (220587)**
- 2. YASH GIRI (221218)**

A BRIEF INTIMIDATION OF OUR EXPERIENCE

- Honestly, it was tough ! Firstly to understand, Secondly to implement. But we hope that the outcome is good (I mean, fairly good).
- Working with a stranger would have given some extra points but Yeah! working with Adithya wasn't bad either. (Adithya: Lolol)
- Being one of the youngest in the course attendees, we were bound to take help from a few seniors. (Thanks to Shivam (Y21) and Shambhavi (Y20) for code suggestions).

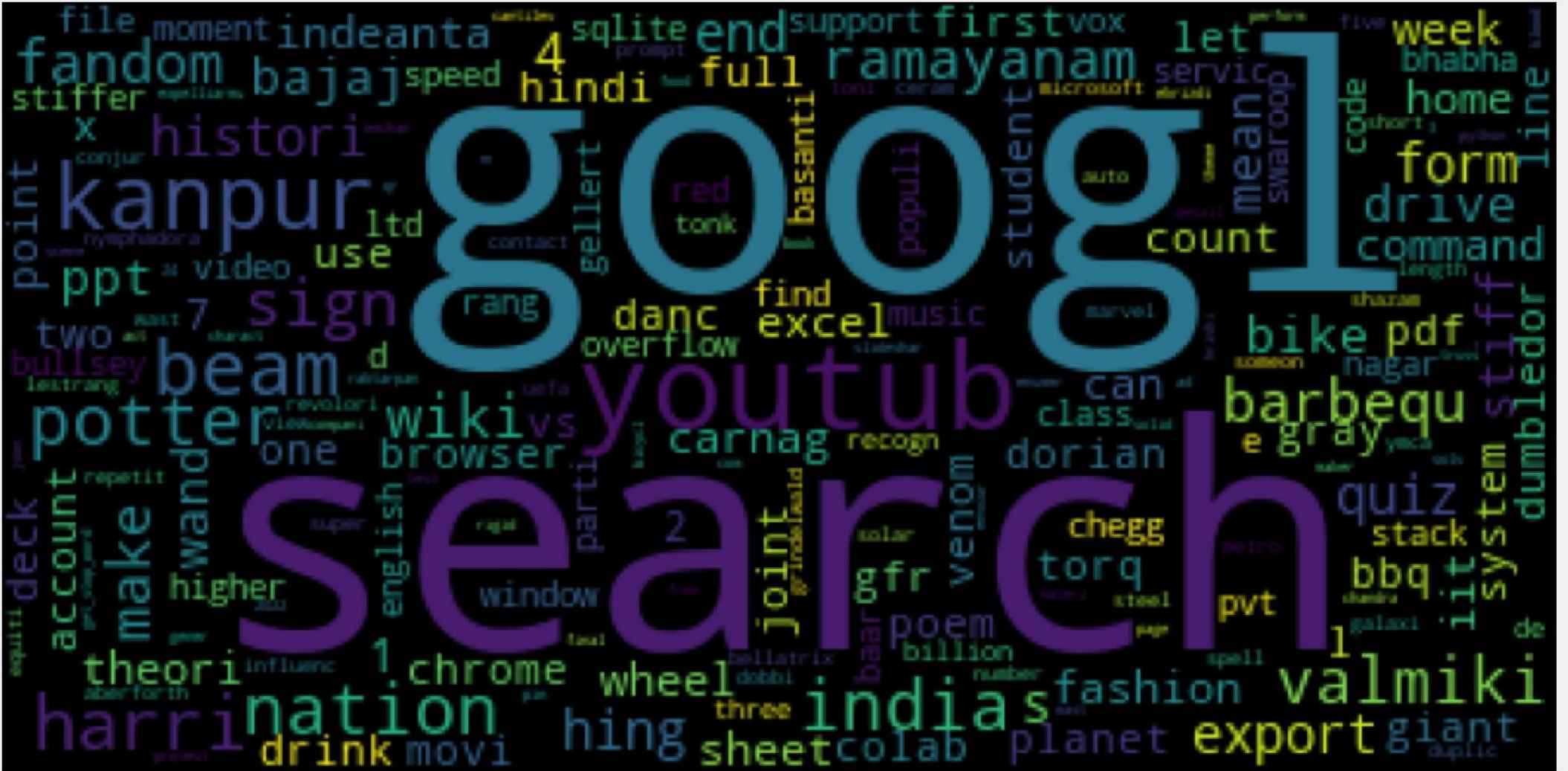
Let's Dive into it !!

STEPS WE FOLLOWED:

1. Exported Adithya's and Yash's web browsing data and got results in form of history_export files.
2. Used VS Code to run 'topic modelling' file and essentially found some errors in the code. (Possibly, errors in our execution)
3. Installed Necessary libraries such as gensim, nltk, stop_words and wordcloud.
4. Plotted Histogram and Line chart for both the history_export files covering Topics (#0, #1, and #2) and established a healthy comparison on productivity scales.

OUR OUTPUTS:

Topic #0

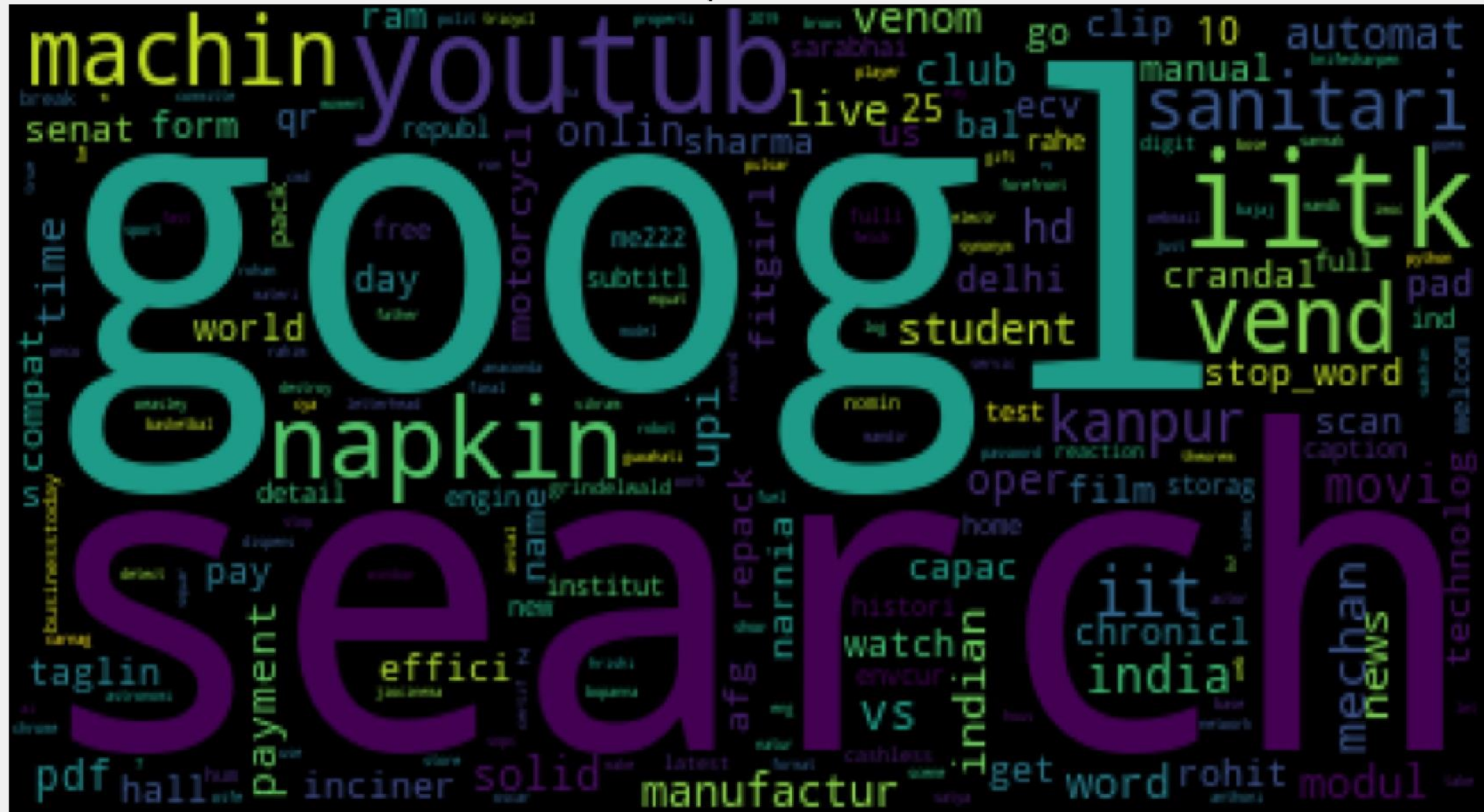


A word cloud visualization for Topic #0. The most prominent words are 'google', 'search', 'youtube', 'india', 'hindi', 'full', 'ramayanam', 'support', 'first', 'week', 'home', 'form', 'drive', 'command', 'pdf', 'barbequ', 'quiz', 'stack', 'bbq', 'it', 'system', 'dumbledor', 'valmiki', 'export', 'planet', 'sheet', 'colab', 'movi', 'drink', 'harrin', 'nation', 'chrome', 'wheel', 'hing', 'three', 'billion', 'poem', 'gfr', 'solar', 'joint', 'grind', 'laald', 'parti', 'window', 'english', 'super', 'wand', 'make', 'account', 'deck', 'repetit', 'theori', 'influenc', '1', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25', '26', '27', '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38', '39', '40', '41', '42', '43', '44', '45', '46', '47', '48', '49', '50', '51', '52', '53', '54', '55', '56', '57', '58', '59', '60', '61', '62', '63', '64', '65', '66', '67', '68', '69', '70', '71', '72', '73', '74', '75', '76', '77', '78', '79', '80', '81', '82', '83', '84', '85', '86', '87', '88', '89', '90', '91', '92', '93', '94', '95', '96', '97', '98', '99', '100'. The words are arranged in a dense, overlapping manner, with 'google' and 'search' being the largest and most central. The colors of the words vary, including shades of blue, green, yellow, and red.

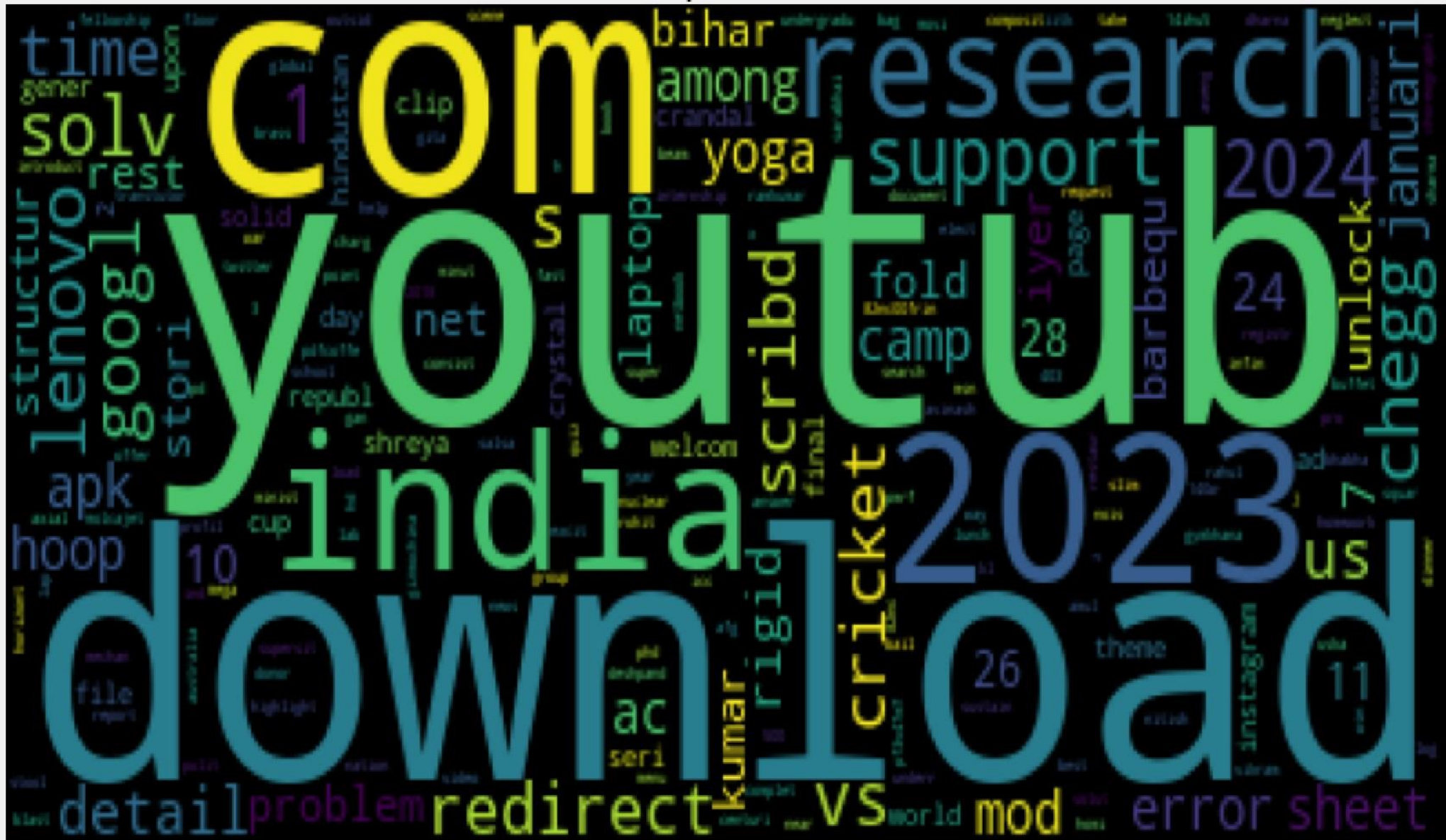
Topic #0



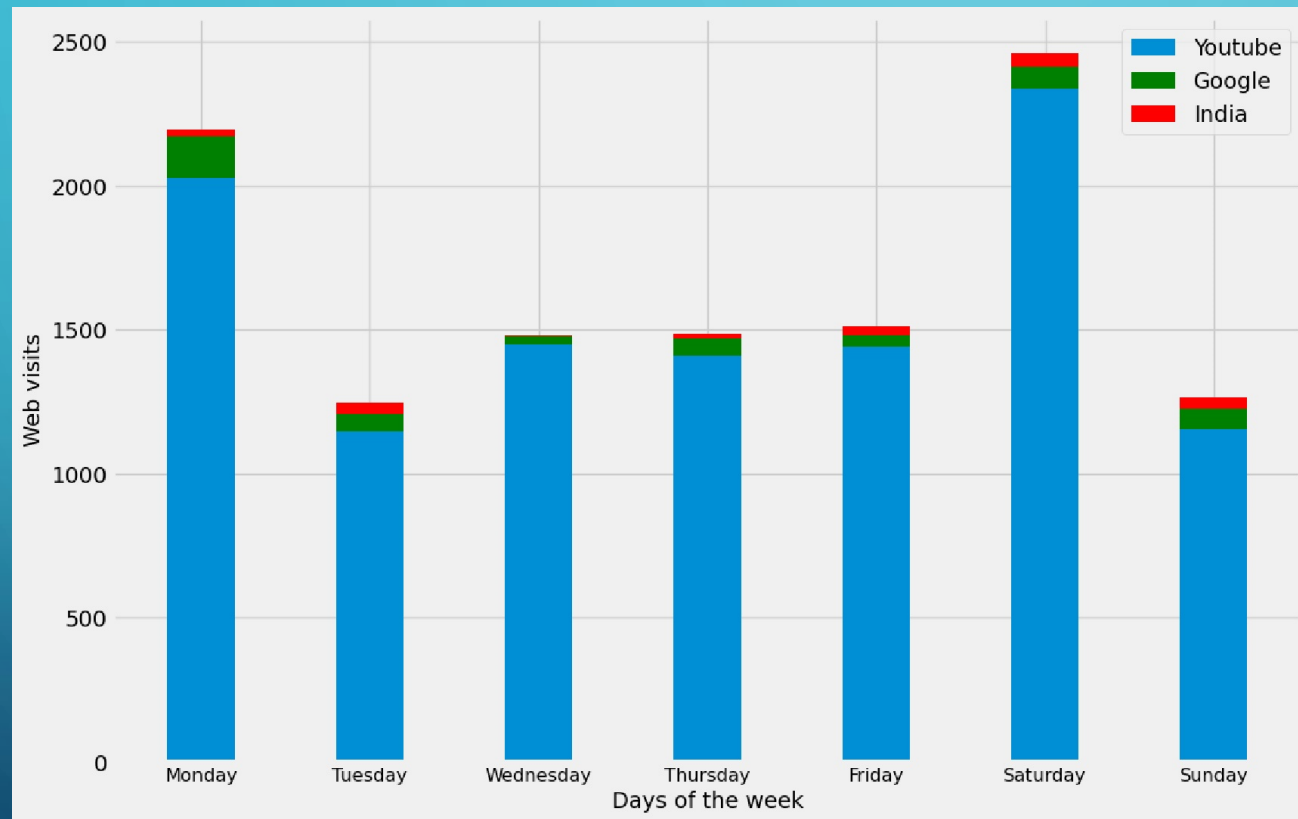
Topic #1



Topic #2



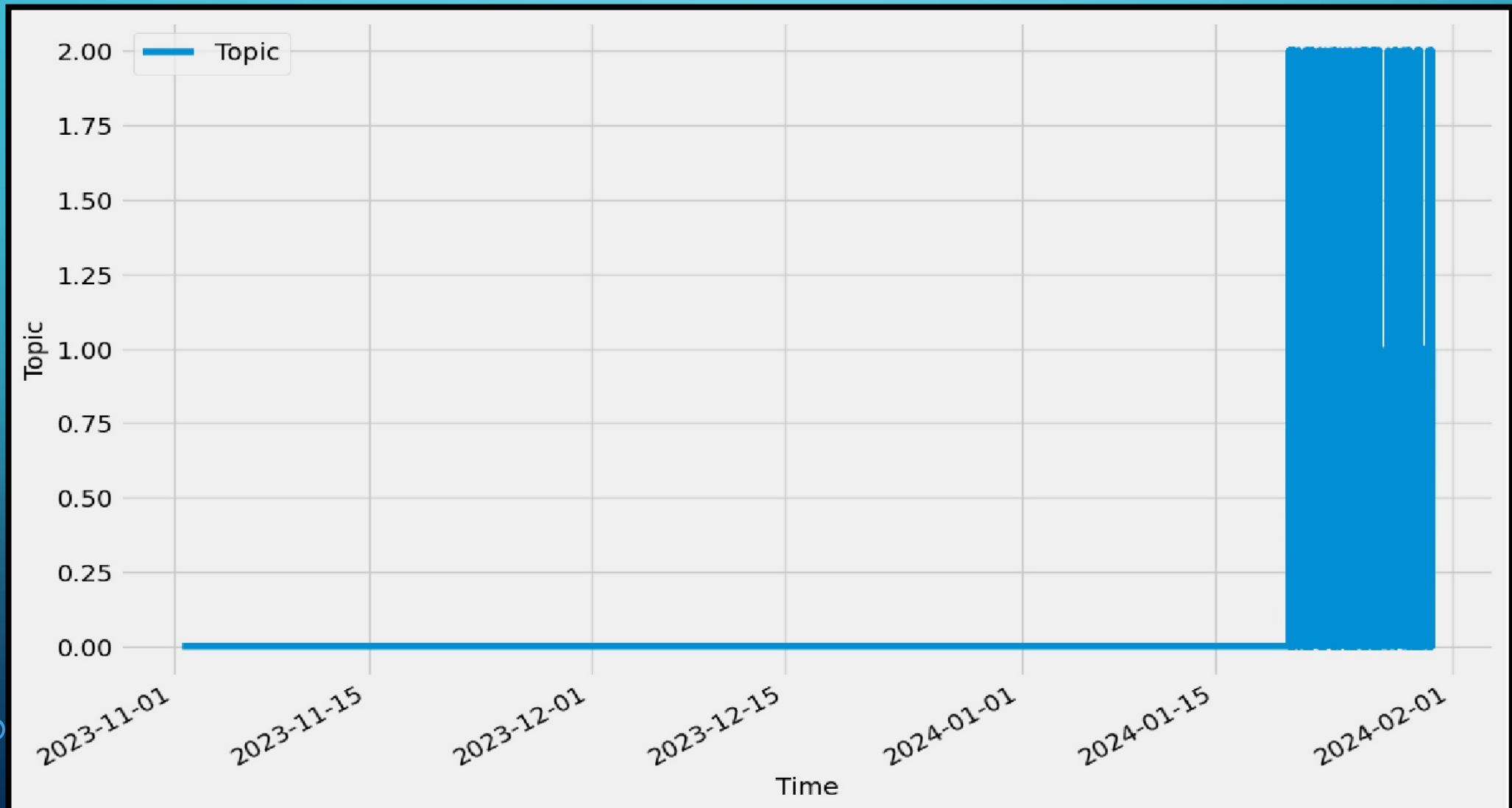
ELEMENTARY COUNTS ANALYSIS



Major Searches:

1. Youtube.com
2. Google.co.in
3. India (Google search)

ANOTHER SIDE OF THE PROJECT: UNEXPECTED OUTPUT



OPTION 1: EXTEND COUNTS ANALYSIS TO HOURS WITHIN THE DAY

1. Edited previously written code to include `hourly_occurences` list and provided a range of 24 hours.
2. Similarly, distinguished topic occurrences for each topic in separate list namely `hourly_t0`, `hourly_t1` and `hourly_t2`.
3. Replicated matplotlib commands to initiate graph plotting using `range=24` instead of `range=7`, naming topics as **YouTube**, **Google** and **India**

CODE

```
hourly_occurrences = []
hours = range(24)          #24 HOURS in One day

for hour in hours:
    hlist = cp_data[cp_data.index.hour == hour].Topic.tolist()
    res = np.histogram(hlist, bins=[0, 1, 2, 3])
    hourly_occurrences.append(list(res[0]))
```

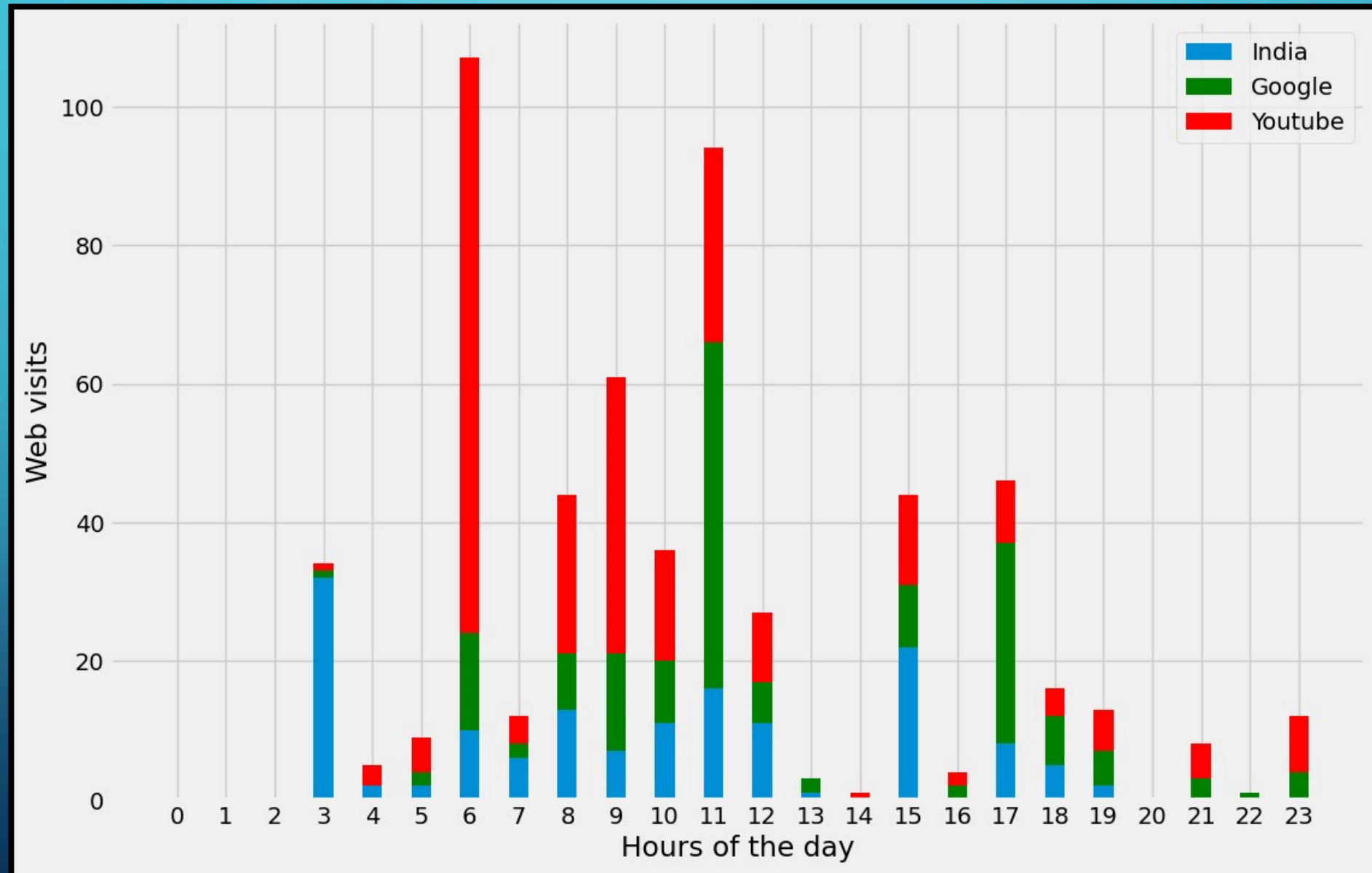
```
hourly_t0 = [hour[0] for hour in hourly_occurrences]
hourly_t1 = [hour[1] for hour in hourly_occurrences]
hourly_t2 = [hour[2] for hour in hourly_occurrences]

cum_hourly_t1 = [sum(x) for x in zip(hourly_t0, hourly_t1)]

plt.figure(figsize=(12, 8))
p0 = plt.bar(range(24), hourly_t0, 0.4, label='India')
p1 = plt.bar(range(24), hourly_t1, 0.4, bottom=hourly_t0, color='green', label='Google')
p2 = plt.bar(range(24), hourly_t2, 0.4, bottom=cum_hourly_t1, color='red', label='Youtube')

plt.xticks(range(24), hours)
plt.xlabel('Hours of the day')
plt.ylabel('Web visits')
plt.legend()
plt.show()
```

OUTPUT



OPTION N: DATA COMPARISON OF TEAM MEMBERS TO RATE ON PRODUCTIVITY

1. Chalked out 6 most visited topics for both the team members.
2. Mutually agreed on **Productivity Rules** for various website categories.
3. Rated web history dominators on various parameters.
4. Declared the Winner ! (The less 'less productive' one)

SOME OBVIOUS PRODUCTIVITY RULES:

1. YouTube, AI website and Entertainment domains are unproductive. (Rating: -5)
2. Infotainment and News to be considered partially productive. (Rating: +5)
3. Cricket and Harry Potter like stuff are unproductive (Rating: -8)
4. Overleaf, Hackerank, CodeChef, Research Interns like searches are productive. (Rating: +8)
5. 'Google Search' or 'Downloading anything' is Considered neutral

PARAMETERS TO JUDGE:

1. Amount of Distraction available on the Topic.
2. Usability of the topic in Academic Reference.
3. Learning + Practice oriented nature of the Topic

TOP 6 MOST VISITED TOPICS -A COMPARATIVE ANALYSIS

ADITHYA:

- Song (-5)
- India (+5)
- Google Search (0)
- Download (0)
- Bank (+5)
- Cricket 2023 (-8)

YASH:

- YouTube (-5)
- Google Search (0)
- India (+5)
- Download (0)
- Research (+8)
- Cricket 2023 (-8)

WINNER HERE !

Points Gained by:

- Adithya: -3
- Yash: 0

The background is a blue gradient with decorative white circuit-like lines in the corners. The text is centered in a bold, dark blue font.

THANK YOU !
IT WAS SERIOUSLY A NEW EXPERIENCE !