In [1]:
```
import re
import pandas as pd
```

In [2]:
```
f = open("chat.txt",'r',encoding='utf-8')
```

In [3]:
```
data = f.read()
```

In [ ]:
```
```

In [5]:
```
pattern = '\d{1,2}/\d{1,2}/\d{2,4},\s\d{1,2}:\d{2}\s-\s'
```

In [6]:
```
messages= re.split(pattern,data)[1:]
```

In [8]:
```
dates= re.findall(pattern,data)
```

In [9]:

```
dates
```

```
Out[9]: ['18/01/21, 15:05 - ',
         '19/01/21, 15:05 - ',
         '27/08/22, 14:56 - ',
         '27/08/22, 16:28 - ',
         '27/08/22, 17:46 - ',
         '27/08/22, 17:46 - ',
         '27/08/22, 17:51 - ',
         '27/08/22, 18:07 - ',
         '27/08/22, 18:07 - ',
         '27/08/22, 18:09 - ',
         '27/08/22, 20:46 - ',
         '27/08/22, 23:06 - ',
         '28/08/22, 08:40 - ',
         '28/08/22, 08:50 - ',
         '28/08/22, 08:54 - ',
         '28/08/22, 10:31 - ',
         '28/08/22, 10:44 - ',
         '28/08/22, 10:45 - ',
         '28/08/22, 11:20 - ',
         '28/08/22, 12:26 - ',
         '28/08/22, 12:42 - ',
         '28/08/22, 12:44 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 17:23 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 19:57 - ',
         '28/08/22, 23:43 - ',
         '28/08/22, 23:44 - ',
         '30/08/22, 15:39 - ',
         '30/08/22, 15:41 - ',
         '30/08/22, 15:41 - ',
         '30/08/22, 15:41 - ',
         '30/08/22, 15:41 - ',
         '30/08/22, 15:41 - ',
         '30/08/22, 15:41 - ',
         '30/08/22, 15:42 - ',
         '30/08/22, 15:42 - ',
 '30/08/22, 15:42 - ',
 '30/08/22, 16:33 - ',
 '30/08/22, 19:04 - ',
 '30/08/22, 19:04 - ',
 '30/08/22, 19:04 - ']
```

In [10]:
```
df = pd.DataFrame({"user_message":messages, "message_date":dates})
```

In [12]:
```
df["message_date"]
```

Out[12]:
```
0     18/01/21, 15:05 - 
1     19/01/21, 15:05 - 
2     27/08/22, 14:56 - 
3     27/08/22, 16:28 - 
4     27/08/22, 17:46 - 
             ...        
57    30/08/22, 15:42 - 
58    30/08/22, 16:33 - 
59    30/08/22, 19:04 - 
60    30/08/22, 19:04 - 
61    30/08/22, 19:04 - 
Name: message_date, Length: 62, dtype: object
```

In [13]:
```
df["message_date"] =pd.to_datetime(df['message_date'], format = "%d/%m/%y, %H:%
```

In [14]:
```
df['message_date']
```

Out[14]:
```
0    2021-01-18 15:05:00
1    2021-01-19 15:05:00
2    2022-08-27 14:56:00
3    2022-08-27 16:28:00
4    2022-08-27 17:46:00
             ...        
57   2022-08-30 15:42:00
58   2022-08-30 16:33:00
59   2022-08-30 19:04:00
60   2022-08-30 19:04:00
61   2022-08-30 19:04:00
Name: message_date, Length: 62, dtype: datetime64[ns]
```

In [15]:
```
df.shape
```

Out[15]:
```
(62, 2)
```

In [17]:
```
df.rename(columns={'message_date':'date'}, inplace = True)
```

In [19]:
```
users = []
messages= []

for message in df['user_message']:
    entry = re.split('([\w\W]+?):\s',message)
    if entry[1:]:
        users.append(entry[1])
        messages.append(" ".join(entry[2:]))
    else:
        users.append('group notification')
        messages.append(entry[0])

df['user']= users
df['message']= messages
df.drop(columns=['user_message'], inplace= True)
```

In [21]:
```
df['year']= df['date'].dt.year
```

In [22]:
```
df
```

Out[22]:

	date	user	message	year
0	2021-01-18 15:05:00	group notification	Akanksha "Team"\n	2021
1	2021-01-19 15:05:00	group notification	You were added\n	2021
2	2022-08-27 14:56:00	group notification	Monika Hope you candidates are ready to atten...	2022
3	2022-08-27 16:28:00	Shweta	<Media omitted>\n\n	2022
4	2022-08-27 17:46:00	Aknasha	<Media omitted>\n	2022
...	...	...	...	...
57	2022-08-30 15:42:00	Ashish	<Media omitted>\n\n	2022
58	2022-08-30 16:33:00	Bhaviya	Syllabus \n\n	2022
59	2022-08-30 19:04:00	Bhaviya	👍🏻 \n	2022
60	2022-08-30 19:04:00	Ashish	👌🏻 \n	2022
61	2022-08-30 19:04:00	Bhaviya	👌🏻	2022

62 rows × 4 columns

In [23]:
```
df['month']= df['date'].dt.month_name()
```

In [25]:
```
df['day']= df['date'].dt.day
```

In [26]:
```
df
```

Out[26]:

	date	user	message	year	month	day
0	2021-01-18 15:05:00	group notification	Akanksha "Team"\n	2021	January	18
1	2021-01-19 15:05:00	group notification	You were added\n	2021	January	19
2	2022-08-27 14:56:00	group notification	Monika Hope you candidates are ready to atten...	2022	August	27
3	2022-08-27 16:28:00	Shweta	<Media omitted>\n\n	2022	August	27
4	2022-08-27 17:46:00	Aknasha	<Media omitted>\n	2022	August	27
...	...	...	...	...	...	...
57	2022-08-30 15:42:00	Ashish	<Media omitted>\n\n	2022	August	30
58	2022-08-30 16:33:00	Bhaviya	Syllabus \n\n	2022	August	30
59	2022-08-30 19:04:00	Bhaviya	👍🏻\n	2022	August	30
60	2022-08-30 19:04:00	Ashish	👌🏻 \n	2022	August	30
61	2022-08-30 19:04:00	Bhaviya	👌🏻	2022	August	30

62 rows × 6 columns

In [27]:
```
df['hour']= df['date'].dt.hour
```

In [29]:
```
df['minute']= df['date'].dt.minute
```

In [31]:
```
df.shape[0]
```

Out[31]:
```
62
```

In [32]:
```
df[df['user']=='Bhaviya']
```

Out[32]:

	date	user	message	year	month	day	hour	minute
58	2022-08-30 16:33:00	Bhaviya	Syllabus \n\n	2022	August	30	16	33
59	2022-08-30 19:04:00	Bhaviya	👍🏻\n	2022	August	30	19	4
61	2022-08-30 19:04:00	Bhaviya	👌🏻	2022	August	30	19	4

In [33]:
```
words=[]
for message in df['message']:
    words.extend(message.split())
```

In [34]:
```
len(words)
```

Out[34]:
```
355
```

In [35]:
```
pip install urlextract
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: urlextract in c:\users\lenovo\appdata\roaming
\python\python39\site-packages (1.8.0)
Requirement already satisfied: idna in c:\programdata\anaconda3\lib\site-pack
ages (from urlextract) (3.3)
Requirement already satisfied: filelock in c:\programdata\anaconda3\lib\site-
packages (from urlextract) (3.6.0)
Requirement already satisfied: uritools in c:\users\lenovo\appdata\roaming\py
thon\python39\site-packages (from urlextract) (4.0.1)
Requirement already satisfied: platformdirs in c:\programdata\anaconda3\lib\s
ite-packages (from urlextract) (2.5.2)
Note: you may need to restart the kernel to use updated packages.
```

In [36]:
```
from urlextract import URLExtract

extractor = URLExtract()
urls= extractor.find_urls("Text with url : www.gmail.com")
print(urls)
```

```
['www.gmail.com']
```

In [37]:
```
links = []
for message in df['message']:
    links.extend(extractor.find_urls(message))
```

In [38]:
```
links
```

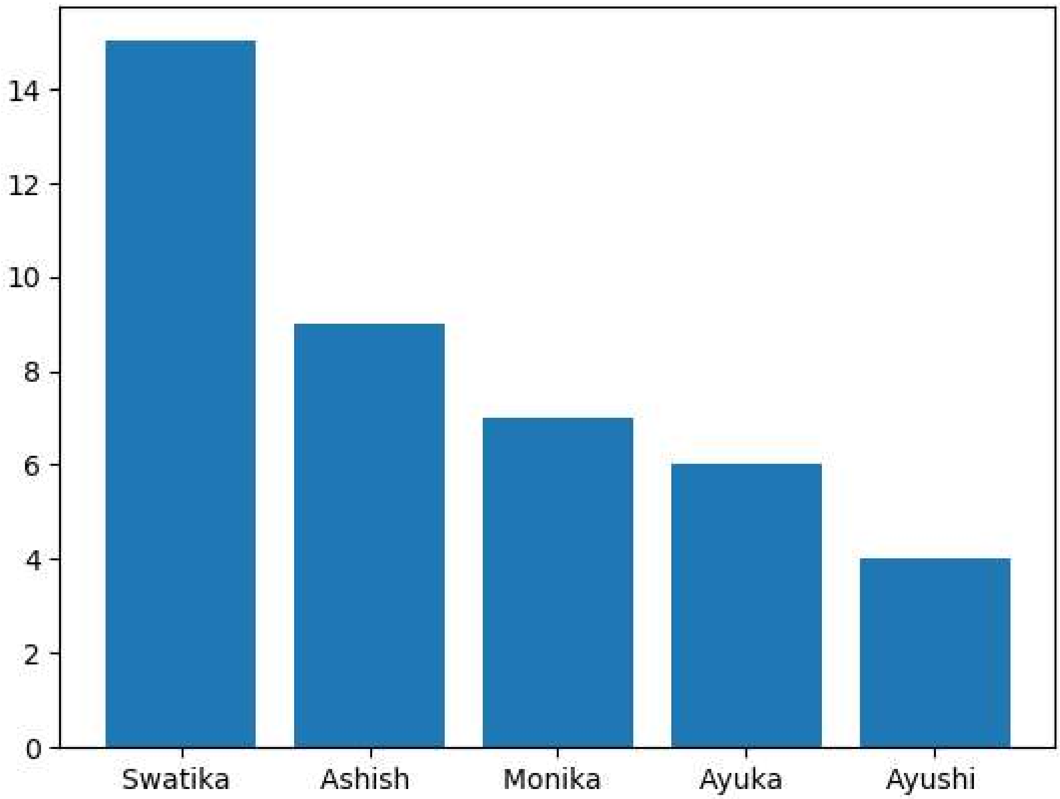
Out[38]:
```
['https://forms.gle/dNBpdL5NMsHmhF9k7',
 'https://www.youtube.com/watch?v=23fQ9-XUSCU',
 'https://www.youtube.com/watch?v=iBIcCGpSpeM&t=574s',
 'https://tinyurl.com/24vee9jt',
 'https://youtu.be/ZWlyGYWw7Cw',
 'https://forms.gle/Ux6YWQkjMAkhvthR6',
 'https://chat.whatsapp.com/D4VpUnfyafU0dS2Hp3vjuJ',
 'https://forms.gle/Ux6YWQkjMAkhvthR6',
 'https://chat.whatsapp.com/D4VpUnfyafU0dS2Hp3vjuJ',
 'https://forms.gle/Ux6YWQkjMAkhvthR6',
 'https://forms.gle/Ux6YWQkjMAkhvthR6']
```

In [40]:
```
x = df['user'].value_counts().head()
```

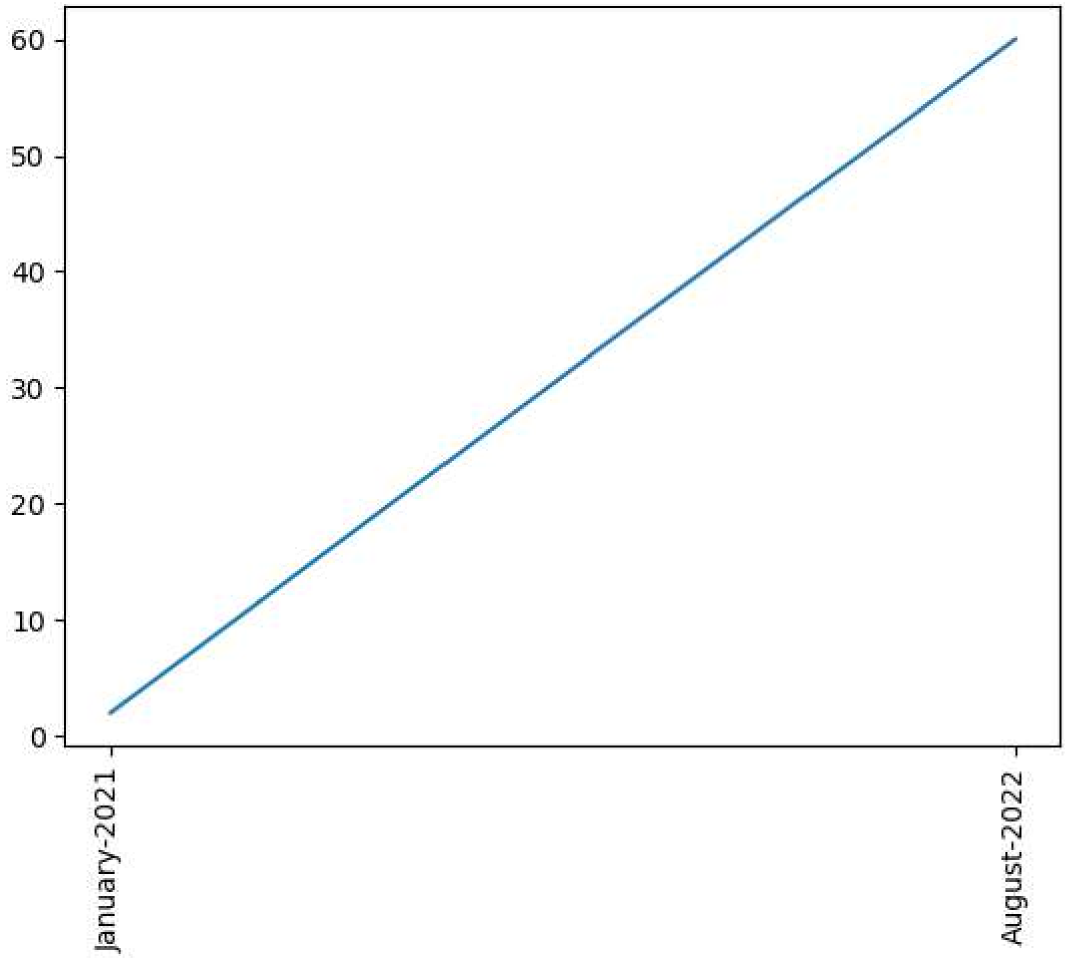
In [41]:
```
import matplotlib.pyplot as plt
```

In [42]:
```
name = x.index
count = x.values
```

In [43]:
```
plt.bar(name,count)
plt.show()
```



In [86]:
```
plt.plot(timeline['time'], timeline['message'])
plt.xticks(rotation = 'vertical')
plt.show()
```


January-2021
August-2022