CS 6350

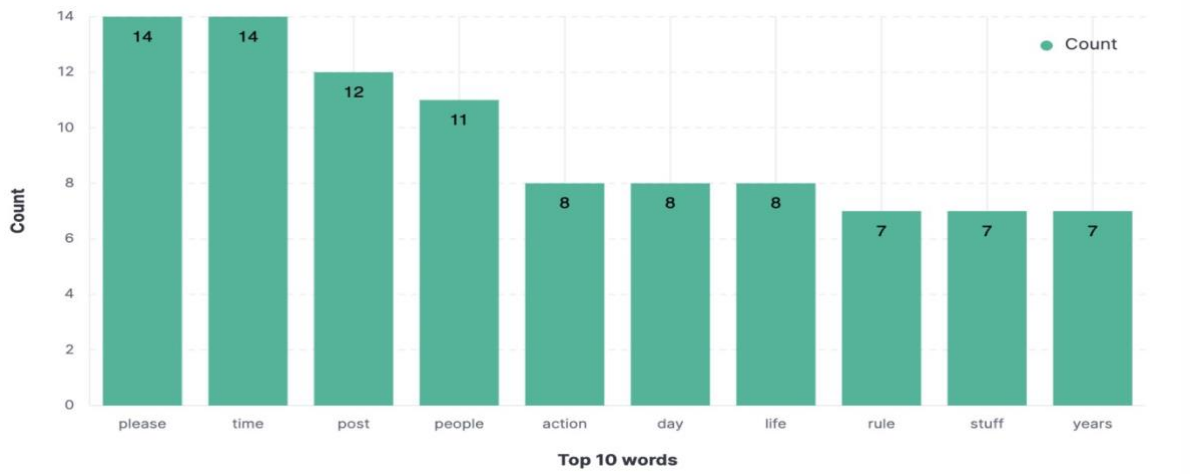
Names of students in your group:

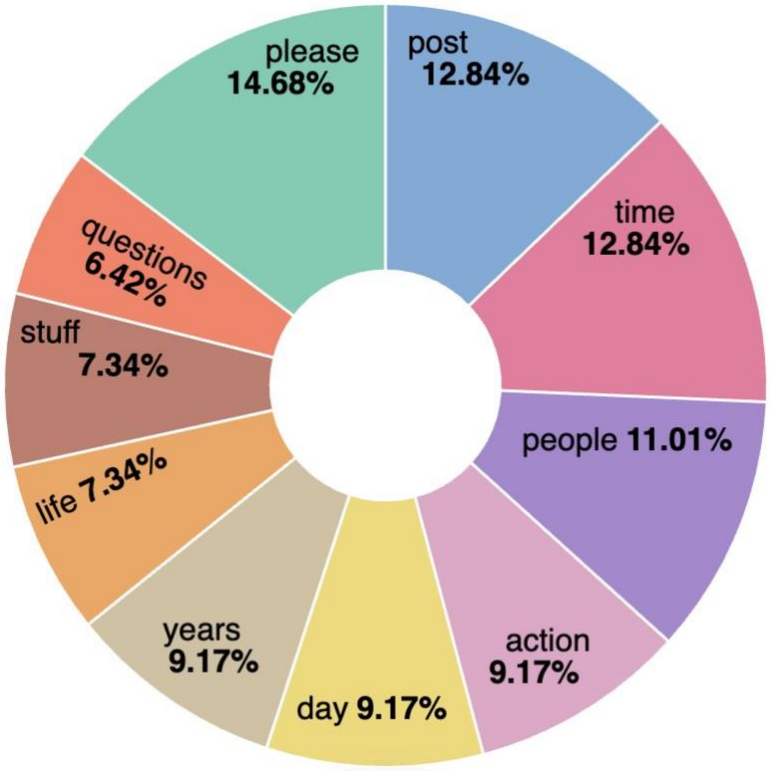
Charan Lokku – CXL220029

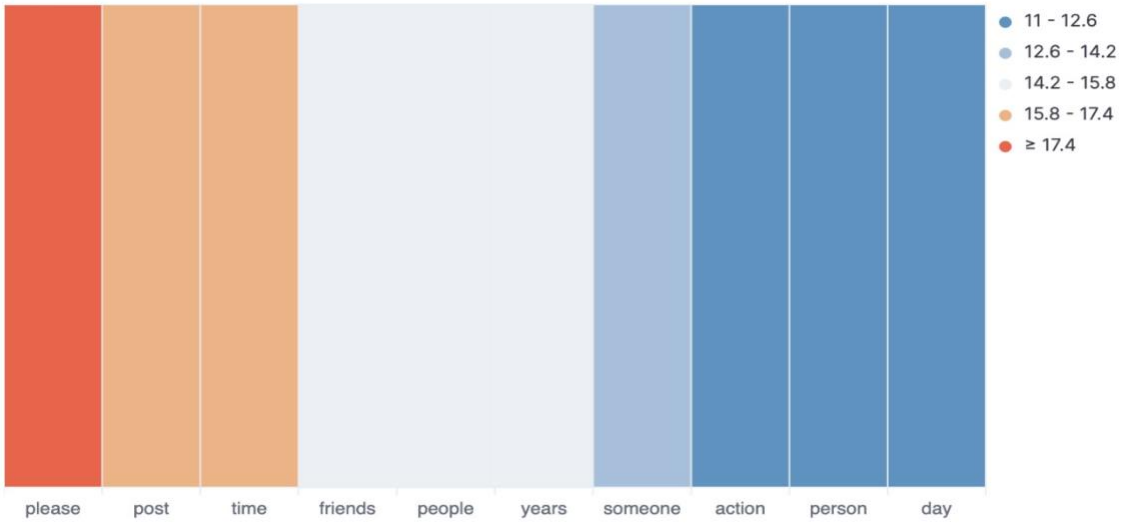
REFERENCES:

- https://www.nltk.org/
- https://kafka.apache.org/#:~:text=Apache%20Kafka%20is%20an%20open,%2C%20and%20mission%2Dcritical%20applications.
- https://dattell.com/data-architecture-blog/what-is-zookeeper-how-does-it-support-kafka/
- https://www.elastic.co/kibana
- https://docs.python.org/3/library/json.html
- https://snap.stanford.edu/data/wiki-Vote.html

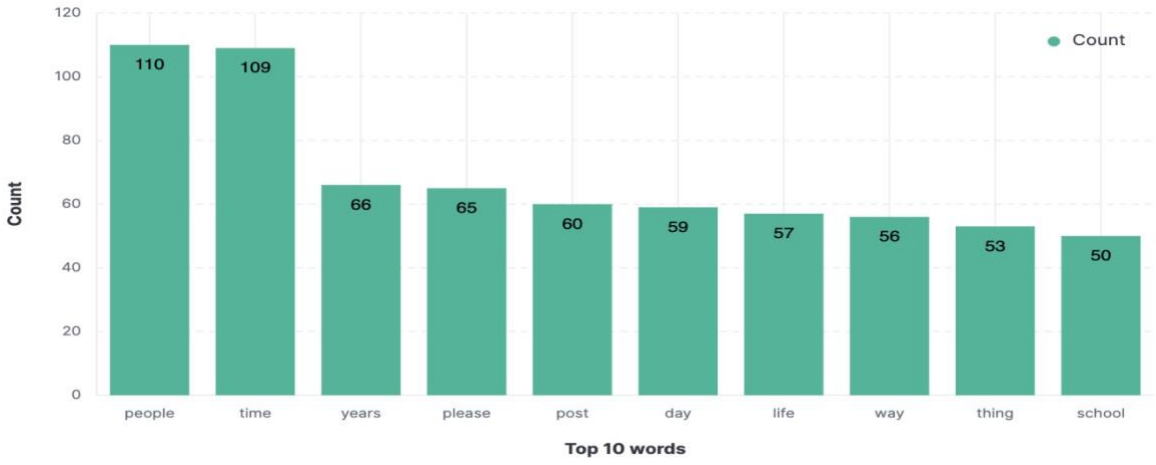
Question 1:
Plots:
15 min

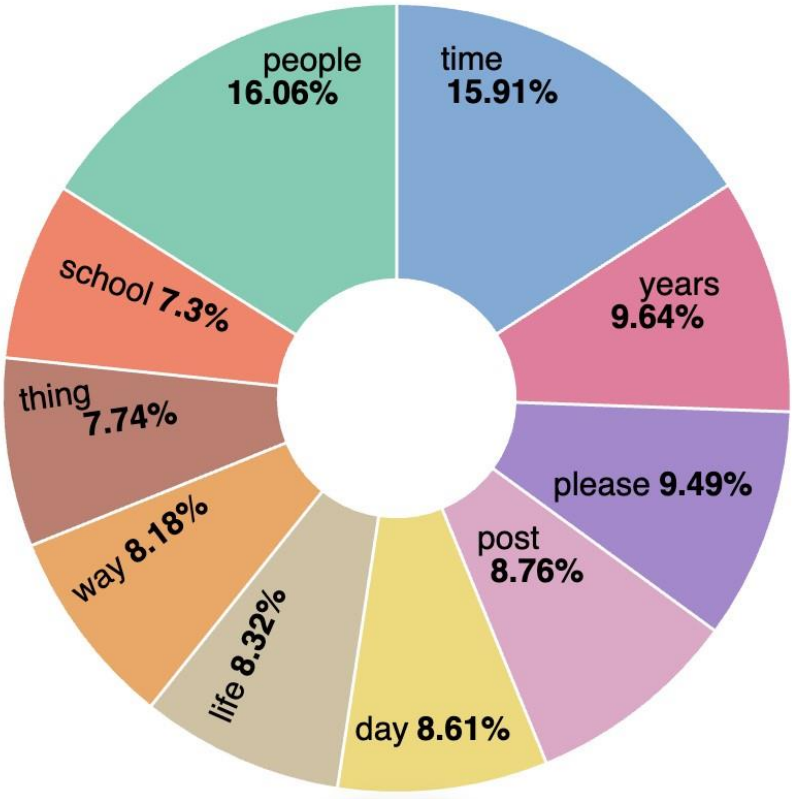

Count
14
12
10
8
6
4
2
0
14
14
12
11
8
8
8
7
7
7
please
time
post
people
action
day
life
rule
stuff
years
Top 10 words


please 14.68%
post 12.84%
time 12.84%
people 11.01%
action 9.17%
day 9.17%
years 9.17%
life 7.34%
stuff 7.34%
questions 6.42%

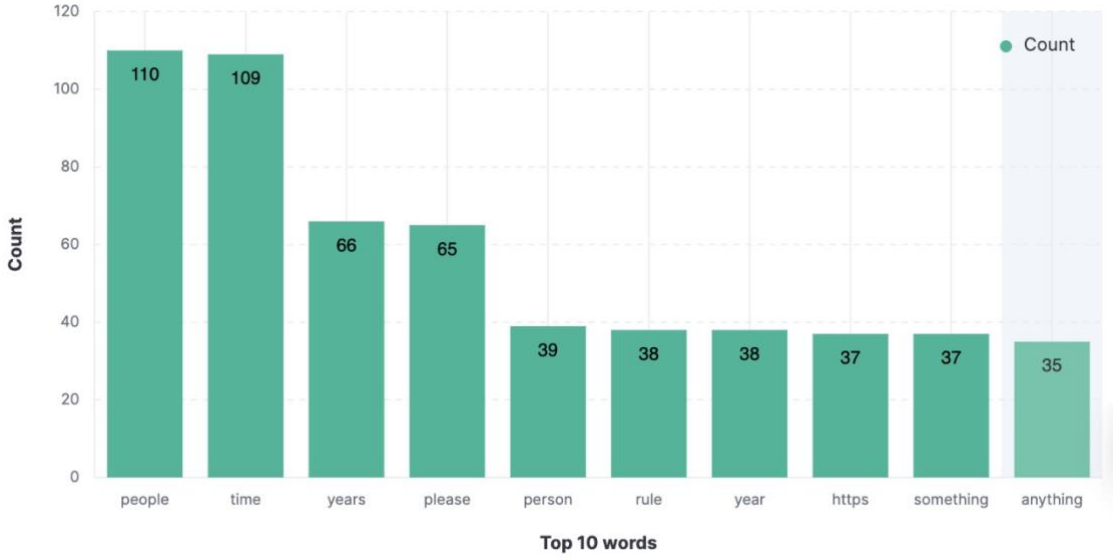

11 - 12.6
12.6 - 14.2
14.2 - 15.8
15.8 - 17.4
≥ 17.4
please
post
time
friends
people
years
someone
action
person
day

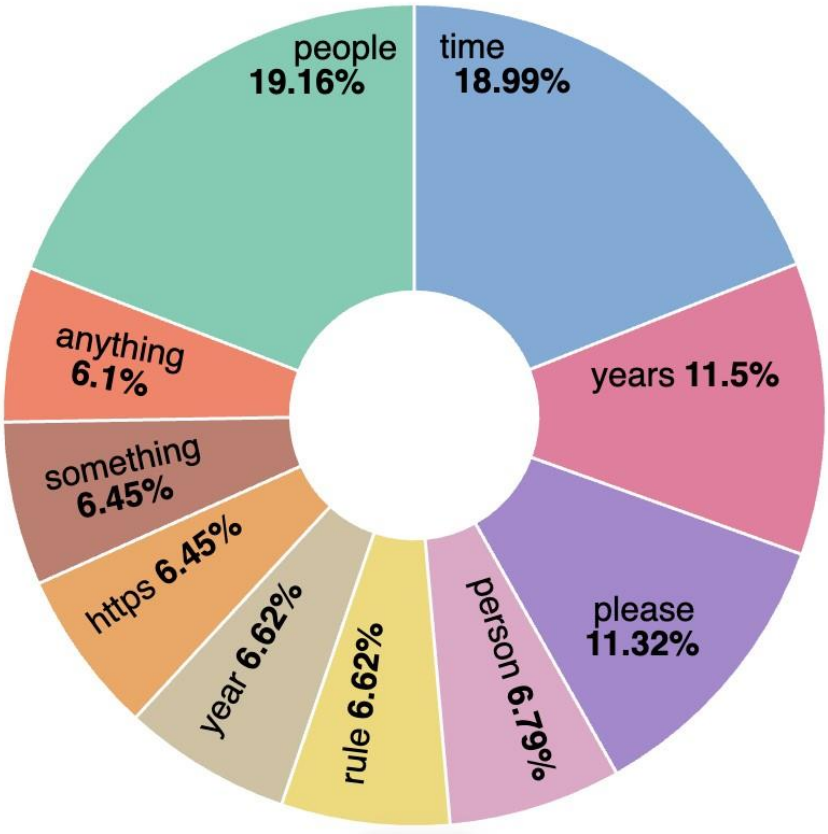
30 min


Count
120
100
80
60
40
20
0
110
109
66
65
60
59
57
56
53
50
people
time
years
please
post
day
life
way
thing
school
Top 10 words

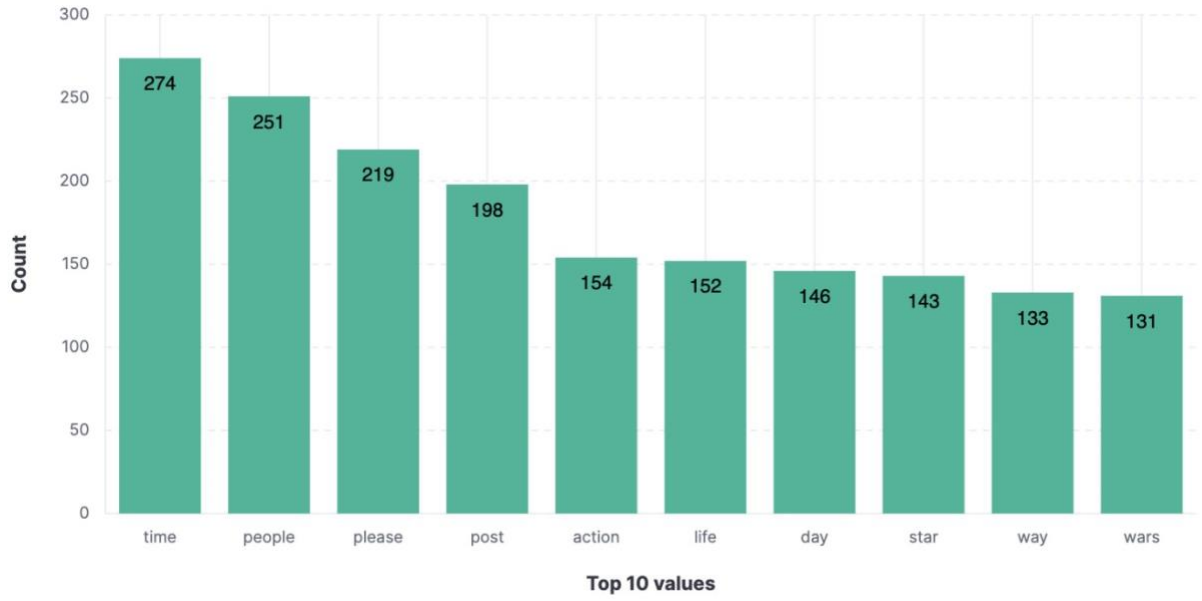

people
16.06%
time
15.91%
years
9.64%
please 9.49%
post
8.76%
day 8.61%
life 8.32%
way 8.18%
thing
7.74%
school 7.3%

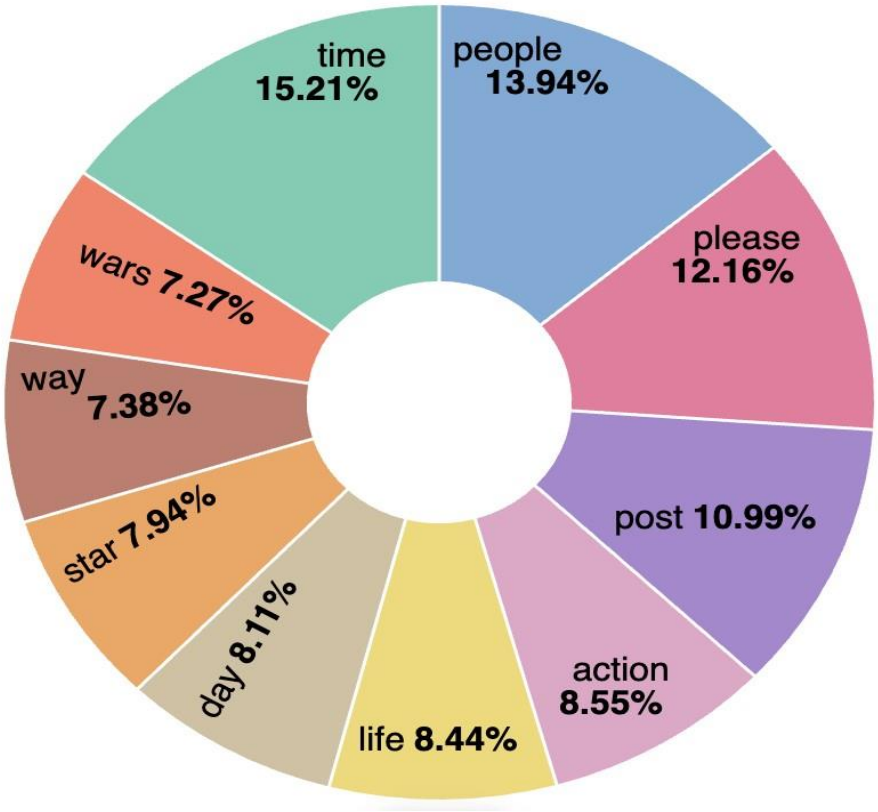
45 min


Count
120
100
80
60
40
20
0
110
109
66
65
39
38
38
37
37
35
people
time
years
please
person
rule
year
https
something
anything
Top 10 words


people
19.16%
time
18.99%
anything
6.1%
years 11.5%
something
6.45%
https 6.45%
year 6.62%
rule 6.62%
person 6.79%
please
11.32%

60 min


300
250
200
150
100
50
0
Count
274
251
219
198
154
152
146
143
133
131
time
people
please
post
action
life
day
star
way
wars
Top 10 values



Summary:

Data Source:

Data for this report originates from NewsAPI, a widely-used news aggregation platform and API offering an extensive array of articles from diverse and reliable sources. Covering various topics like politics, business, tech, sports, entertainment, and more, NewsAPI provides both real-time and historical news data. Its functionalities allow for tailored searches based on keywords, sources, language, and time, making it a valuable tool for monitoring and analyzing news trends.

Results:

The analysis conducted using NewsAPI data and graphs depicting named entity frequency illustrates a consistent increase in word frequency over time. Tracking the top 10 mentioned entities in news articles at 15, 30, 45, and 60-minute intervals revealed a continuous uptick in their references. This trend indicates that specific entities become more prominent and garner increased attention in news coverage as time progresses, potentially signaling emerging trends or significant news events. This data presents

valuable insights into the evolving news sphere and shifting media priorities, assisting businesses, policymakers, and individuals in making well-informed decisions.

Question - 2:

Input data: https://snap.stanford.edu/data/wiki-Vote.html

Databricks link:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/7537527909955009/1258545579046733/8520173309402241/latest.html

Output:

a. Find the top 5 nodes with the highest outdegree and nd the count of the number of outgoing edges in each

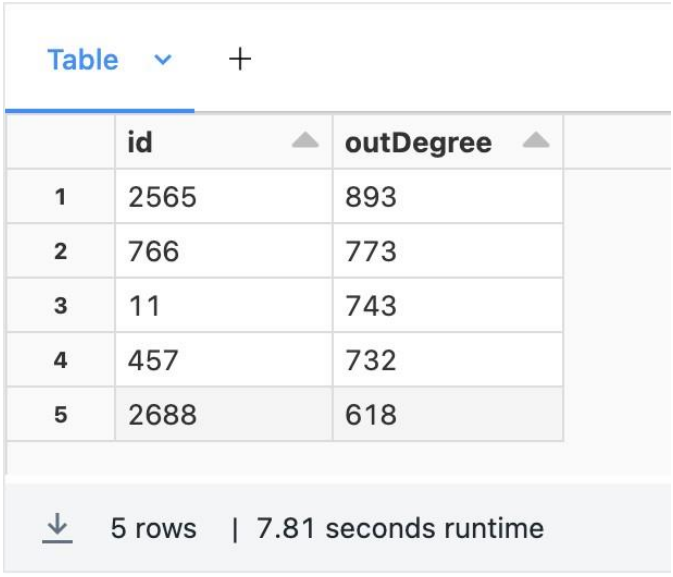


Table +

	id	outDegree
1	2565	893
2	766	773
3	11	743
4	457	732
5	2688	618

5 rows | 7.81 seconds runtime

b. Find the top 5 nodes with the highest indegree and nd the count of the number of incoming edges in each

Table +

	id	inDegree
1	4037	457
2	15	361
3	2398	340
4	2625	331
5	1297	309

5 rows | 3.21 seconds runtime

c. Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values. You are free to define any suitable parameters.

Table +

	id	pagerank
1	4037	32.761392590350795
2	15	26.25300495761947
3	6634	26.16452443488649
4	2625	23.51151593302638
5	2398	18.728389390669683

5 rows | 58.75 seconds runtime

d. Run the connected components algorithm on it and nd the top 5 components with the largest number of nodes.

Table +

	component	count
1	0	7066
2	532575944741	3
3	592705486870	3
4	936302870556	3
5	884763263008	2

5 rows | 19.99 minutes runtime

e. Run the triangle counts algorithm on each of the vertices and output the top 5 vertices with the largest triangle count. In case of ties, you can randomly select the top 5 vertices.

Table +

	id	count
1	2565	30940
2	1549	22003
3	766	18204
4	1166	17361
5	2688	14220

5 rows | 36.09 seconds runtime

Summary:

- Node with highest out-degree is node 2565 and the node with highest in-degree is node 4037.
- So finally, node 4037 will have the highest page rank.
- It can be observed and hence confirmed from the results that node 4037 has the highest page rank.

- With 7066 connections, node 0 has the most number of connections with other nodes. This may be the initial default node that was formed.
- Node 2565 has the highest Triangle Count which is 30940.