CS 6350
ASSIGNMENT ___3___________

Names of students in your group:

Venkata Kowsik Temididapathi – VXT200001
Veda Nandan Gandi – VXG200001

Number of free late days used: ____2________________

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

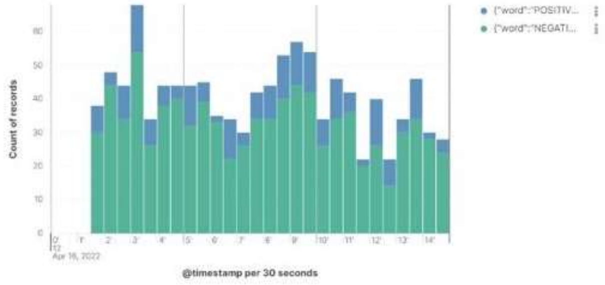
Note: The README files for each part are there in the part folder. It contains the details on how to run it and other details.

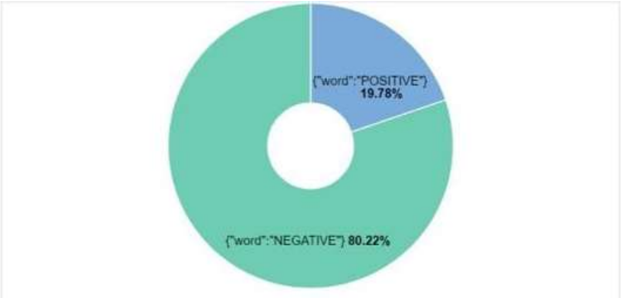
Please list clearly all the sources/references that you have used in this assignment.

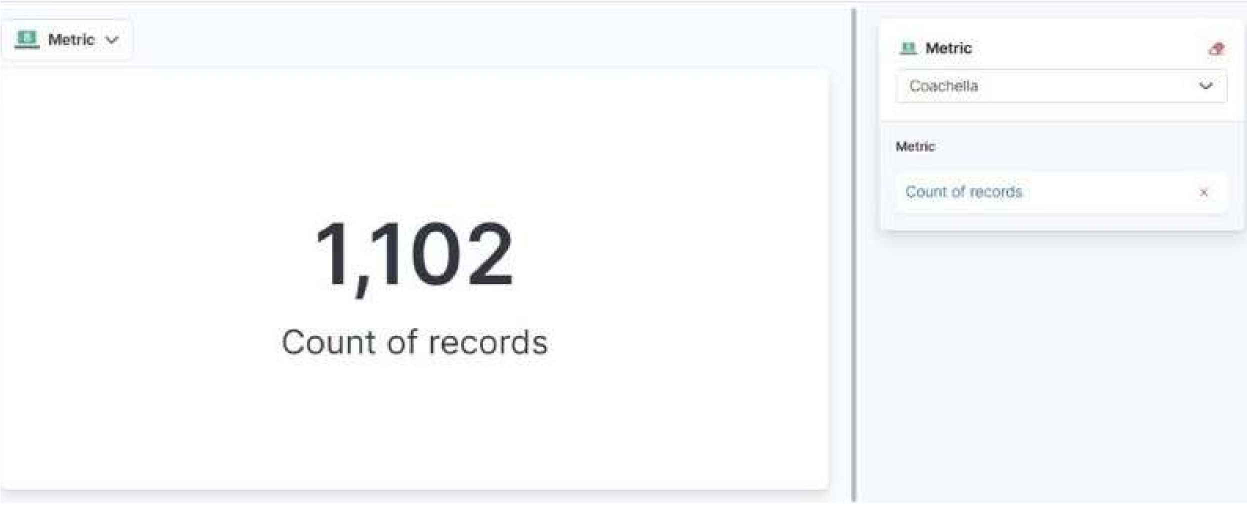
- https://www.elastic.co/guide/en/elasticsearch/reference/current/important-settings.html
- https://stackoverflow.com/questions/70581159/typeerror-init-missing-2-required-positional-arguments-access-token-an
- https://kafka.apache.org/quickstart

- https://docs.microsoft.com/en-us/windows/wsl/setup/environment
- https://snap.stanford.edu/data/index.html#socnets
- https://graphframes.github.io/graphframes/docs/_site/user-guide.html#pagerank
- https://graphframes.github.io/graphframes/docs/_site/user-guide.html#triangle-count
- https://graphframes.github.io/graphframes/docs/_site/user-guide.html#connected-components

Part 1:
The search term can be changed in the config file that is in the part 1 folder.
We are using the hashtag of #Coachella (as it is presently happening, and it is trending).





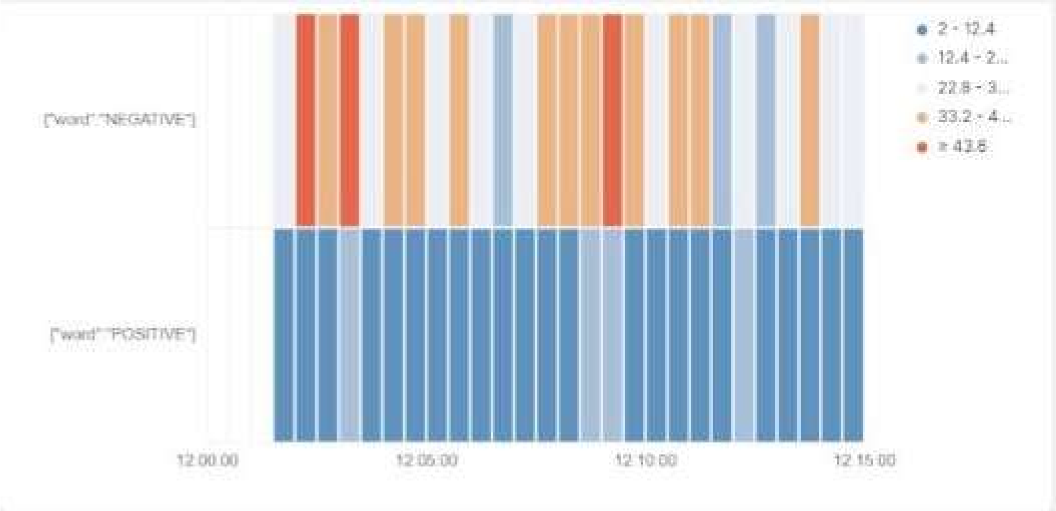


1,102

Count of records

Metric: Coachella

Metric: Count of records



From the timeseries graph, we can see that most of the tweets are negative. Since Coachella is a music festival, the language that is used is more negative in a positive sense. For example, “the festival was soooo bad”, the actual meaning is that the festival was cool or rad, but sentiment analysis will classify it as negative. The timeline is almost around the same range, and it was higher during the performance of the artists. It was higher when a particular performance was taking place

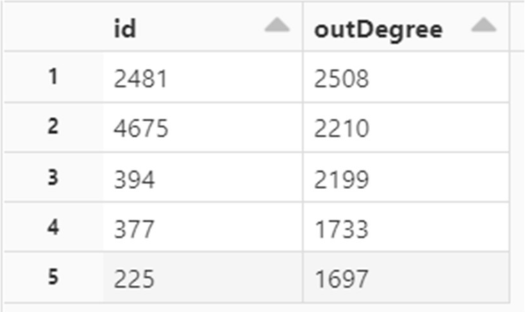
(https://pitchfork.com/news/2ne1-reunite-during-88risings-set-at-coachella-2022-watch/).
You can also see the distribution in the bar graph and it almost the same quantity around all windows (i.e. 30 seconds).

Part 2:
Dataset Used: Slash Dot Network.
Databrick Notebook: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4241627352815915/1853243403756171/1617445708688506/latest.html
Output: It is shown and labelled in databrick notebook.
Summary:

- Node 2481 has the highest in-degree and out-degree so ultimately this will have the highest page rank. This can be validated in the page rank produced by the code where node 2481 is the highest.
- Node 0 has the highest connections from other nodes with a count of 33457. This might be the default node created at the beginning. The next nodes have connections of 6.
- Node 46 has the highest “Triangle Count” with a count of 14888.

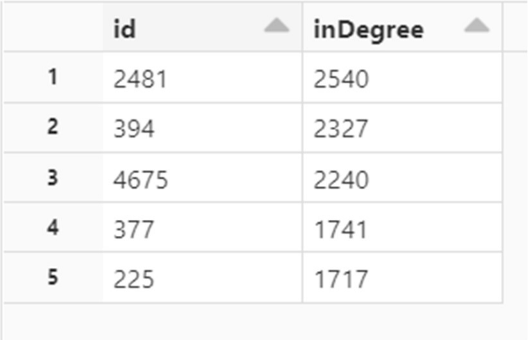
Output:

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



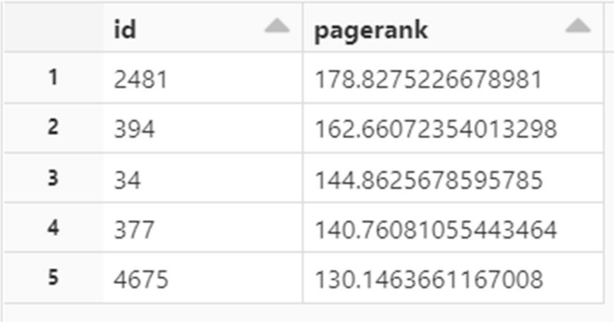
	id	outDegree
1	2481	2508
2	4675	2210
3	394	2199
4	377	1733
5	225	1697

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



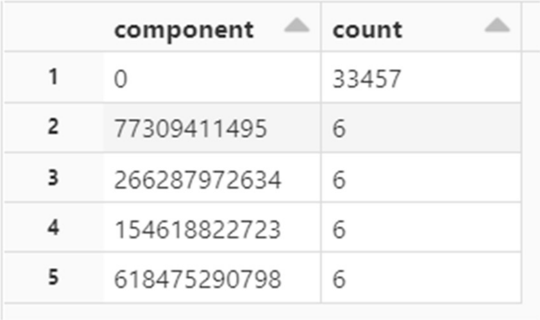
	id	inDegree
1	2481	2540
2	394	2327
3	4675	2240
4	377	1741
5	225	1717

3. 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.



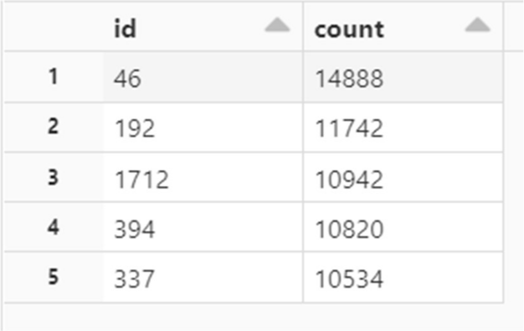
	id	pagerank
1	2481	178.8275226678981
2	394	162.66072354013298
3	34	144.8625678595785
4	377	140.76081055443464
5	4675	130.1463661167008

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



	component	count
1	0	33457
2	77309411495	6
3	266287972634	6
4	154618822723	6
5	618475290798	6

5. 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.



	id	count
1	46	14888
2	192	11742
3	1712	10942
4	394	10820
5	337	10534