Use the allocated data source (any resource given below) and build a utility to extract and curate data for the analytics tasks (Eg. Hate Speech Detection, etc) specified. Mention any keywords used for search, Ensure that this module provides a Live Stream of data. – (3 marks)

Using the data collected as part of question 1, perform the given task, without using bloom filter with the data stream and with the bloom filter applied to the data stream (Use an existing library if available for this part) – (2+2 Marks)

Finally, build a visualization module for data obtained from the task carried out in part 2 so that any changes in the stream are reflected in the visualization. Visualization will be done using time series plot, heat map and word clouds for both before and after application of bloom filter. Also compute execution time before and after applying bloom filter and give your reasons for the results obtained– (3 Marks).

Live Stream Data Sources List:

1. Wiki Logs API – How to create and access wiki logs
https://www.mediawiki.org/wiki/API:Logevents#GET_request and
https://www.mediawiki.org/wiki/Special:ApiSandbox#action=query&format=json&list=logevents

2. Reddit API- How to create and access https://www.reddit.com/wiki/api (Access from own private network as firewall policy on NITK will not allow students to access this site)

3. iMDb API – How to create and access https://developer.imdb.com/

4. YouTube API – How to create and access https://developers.google.com/youtube/v3/

5. Tumblr API – How to create and access https://www.tumblr.com/docs/en/api/v2

Task Allocation List:

Data Source Task

Wiki Logs Top Stats and Usage Identification (Nos. 1-19 + backlogs)

Reddit Hate Speech Detection (Nos. 20-41)

YouTube Hate Speech Detection (Nos. 66-87)

Tumblr Topic Identification