Twitter4J API can be used in collecting tweets but one of the major drawback that we realized in collecting those tweets was that stream cannot be filtered by location bounding box. So, our first major issue was setting up auto collection of geolocated tweets. We got around this problem by using twitter streaming API directly. Twitter Streaming APi is simple query which has to be fired on twitter servers at 'https://stream.twitter.com/1.1/statuses/filter.json'. This query can be changed appropriately to accommodate the type of filters one wants to apply.
Now, once a stream of geolocated stream has been established we needed to take it in spark as D-streams. D-Streams are collection of sets RDDs used by spark to manage incoming live data. So, one was getting it was pulling stream through a socket. Hence, we separated two processes of data collection and data processing and analysis.
We wrote our collection stage in Python, since it is very easy to work with sockets in Python. Reason being, levels of abstractions and available libraries. We pushed the incoming stream to a socket. This stream is then picked up in Spark to process the data.
The above part discusses on the architecture for how the data was collected.

Data Collected:
We setup two bounding boxes: one for New York and one for entire America. We also made sure all the tweets that came through the query were in English.

Queries used are:

1. NY:
   https://stream.twitter.com/1.1/statuses/filter.json?language=en&locations=-74,40,-73,41
2. USA:
   https://stream.twitter.com/1.1/statuses/filter.json?language=en&locations=-124,24,-67,49

Data that comes from this API has to be converted into byte stream and pushed over the socket.

Reported Bug in Twitter4J API:
While exploring how we could collect tweets directly using Spark Streaming, we can use Twitter4J API. But since there is no support for geolocated tweet level filtering as mentioned above. But while trying to find we saw that there were inconsistency in Twitter Streaming API and Twitter4J API. This inconsistency was in filtering tweets by keywords. The array structure used was wrong. So, we saw previous versions and new ones and there was inconsistency here as well. Bug was only in the new version.