The description of the project can be found in this " Cloud-Real Time Event Monitoring.pdf " presentation.

Here, I will discuss how to run the code.

Step-1 : Collecting the twitter stream. For that, we used java twitter4j library. TwitterStreamer is folder which contains the java source code. For running the code, import the project in eclipse add corresponding jar files.

The code generates two output files:
hashtags1.txt and tweets1.txt

Step-2: This input is given to a spark job which detects the hot topics based on the count.

It generates outputs in the "test" folder which contains the tweet and corresponding text in the tweet. The code for this is "tweet.py"

```
test
----
  |
  ----part-00000
  ----Success
```

```
./bin/pyspark tweet.py hashtags1.txt
```

Step-3: This part-00000 file is given as an input to "tweet1.py" which clusters the tweets and puts them in a dictionary. This output is stored in another file.

The dictionary formart is tweetlist['tweet']=List[All tweets with that hashtag]

```
python tweet1.py test/part-00000 > temp
```

Step-4: temp file is given as an input to tweet2.py which computes the similarity and finds the tweet with highest content.

```
python tweet2.py
```