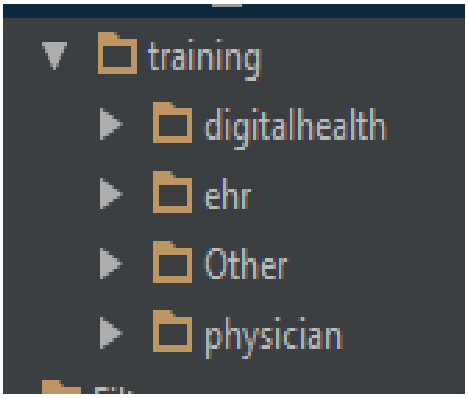
Lab Report for Assignment 8

Recommendation System:

Here , we are recommending the user some important tweets based on few popular medical hash tags. We have filtered some tags and collected tweets accordingly.

allfilter.txt

For now we have divided it into below mentioned 4 categories. These categories are on basis of hashtags:



We are training the data for this 4 categories and will test them against the future tweets. For now its just 4 categories but we will include more and plan is to ask user for his interest of tweet and on basis of that we will find appropriate tweet. The selection criteria for choosing tweet from streaming will be on basis of ratings. Program will decide ratings on basis of user's input and how influenced that tweet is to others people so far. We will include sentiment analysis as well for making sure that positive tweets will reach to user.

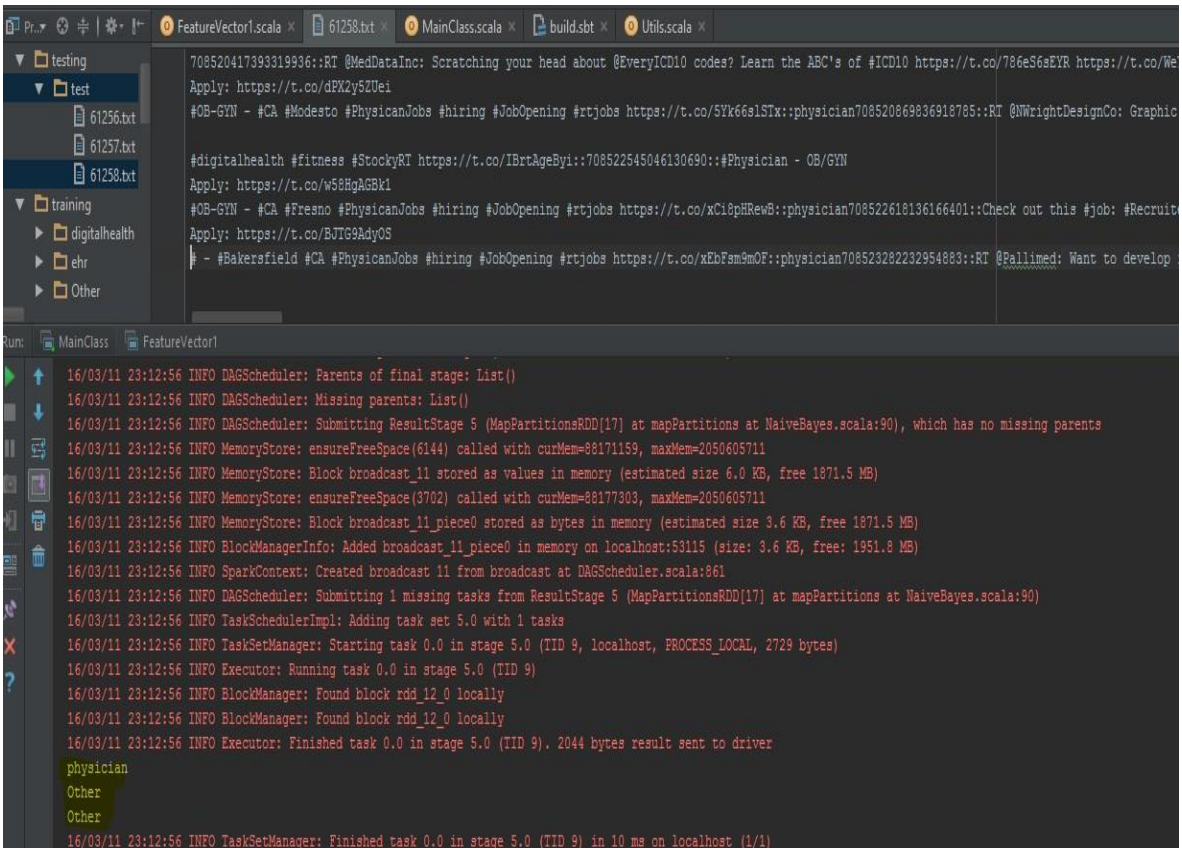
Below is the sample predicted results for few tweets.

61257.txt

61258.txt

61256.txt

Above three files are the collected tweets during streaming and below is the predicted output:


FeatureVector1.scala
61258.txt
MainClass.scala
build.sbt
Utils.scala
testing
test
61256.txt
61257.txt
61258.txt
training
digitalhealth
ehr
Other
708520417393319936::RT @MedDataInc: Scratching your head about @EveryICD10 codes? Learn the ABC's of #ICD10 https://t.co/786eS6sEYR https://t.co/We
Apply: https://t.co/dPX2y5ZUei
#OB-GYN - #CA #Modesto #PhysicanJobs #hiring #JobOpening #rtjobs https://t.co/5Yk66s1STx::physician708520869836918785::RT @NWrightDesignCo: Graphic
#digitalhealth #fitness #StockyRT https://t.co/IBrtAgeByi::708522545046130690::#Physician - OB/GYN
Apply: https://t.co/w58HgAGBk1
#OB-GYN - #CA #Fresno #PhysicanJobs #hiring #JobOpening #rtjobs https://t.co/xCi8pHRewB::physician708522618136166401::Check out this #job: #Recruit
Apply: https://t.co/BJTG9AdyOS
- #Bakersfield #CA #PhysicanJobs #hiring #JobOpening #rtjobs https://t.co/xEbFsm9mOF::physician708523282232954883::RT @Pallimed: Want to develop
Run:
MainClass
FeatureVector1
16/03/11 23:12:56 INFO DAGScheduler: Parents of final stage: List()
16/03/11 23:12:56 INFO DAGScheduler: Missing parents: List()
16/03/11 23:12:56 INFO DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[17] at mapPartitions at NaiveBayes.scala:90), which has no missing parents
16/03/11 23:12:56 INFO MemoryStore: ensureFreeSpace(6144) called with curMem=88171159, maxMem=2050605711
16/03/11 23:12:56 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 6.0 KB, free 1871.5 MB)
16/03/11 23:12:56 INFO MemoryStore: ensureFreeSpace(3702) called with curMem=88177303, maxMem=2050605711
16/03/11 23:12:56 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 3.6 KB, free 1871.5 MB)
16/03/11 23:12:56 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:53115 (size: 3.6 KB, free: 1951.8 MB)
16/03/11 23:12:56 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:861
16/03/11 23:12:56 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (MapPartitionsRDD[17] at mapPartitions at NaiveBayes.scala:90)
16/03/11 23:12:56 INFO TaskSchedulerImpl: Adding task set 5.0 with 1 tasks
16/03/11 23:12:56 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 9, localhost, PROCESS_LOCAL, 2729 bytes)
16/03/11 23:12:56 INFO Executor: Running task 0.0 in stage 5.0 (TID 9)
16/03/11 23:12:56 INFO BlockManager: Found block rdd_12_0 locally
16/03/11 23:12:56 INFO BlockManager: Found block rdd_12_0 locally
16/03/11 23:12:56 INFO Executor: Finished task 0.0 in stage 5.0 (TID 9). 2044 bytes result sent to driver
physician
Other
Other
16/03/11 23:12:56 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 9) in 10 ms on localhost (1/1)