SCALA: 2.11.8

SPARK: 2.2.0

run from “seekwell” folder:
1) `sbt clean package` (to create executable jar)
2)
`spark-submit --class Seeker --deploy-mode client --num-executors 4 target/scala-2.11/seekwell_2.11-1.0.jar`

```
17/10/08 14:29:42 INFO SparkContext: Starting job: show at Seeker.scala:70
17/10/08 14:29:42 INFO DAGScheduler: Got job 4 (show at Seeker.scala:70) with 1 output partitions
17/10/08 14:29:42 INFO DAGScheduler: Final stage: ResultStage 6 (show at Seeker.scala:70)
17/10/08 14:29:42 INFO DAGScheduler: Parents of final stage: List()
17/10/08 14:29:42 INFO DAGScheduler: Missing parents: List()
17/10/08 14:29:42 INFO DAGScheduler: Submitting ResultStage 6 (MapPartitionsRDD[26] at show at Seeker.scala:70), which has no
 missing parents
17/10/08 14:29:42 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 4.0 KB, free 366.2 MB)
17/10/08 14:29:42 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 2.3 KB, free 366.2 MB)
17/10/08 14:29:42 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on 192.168.0.158:44057 (size: 2.3 KB, free: 366.3
 MB)
17/10/08 14:29:42 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1006
17/10/08 14:29:42 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 6 (MapPartitionsRDD[26] at show at Seeker.sc
ala:70) (first 15 tasks are for partitions Vector(0))
17/10/08 14:29:42 INFO TaskSchedulerImpl: Adding task set 6.0 with 1 tasks
17/10/08 14:29:42 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 12, localhost, executor driver, partition 0, PROCE
SS_LOCAL, 14041 bytes)
17/10/08 14:29:42 INFO Executor: Running task 0.0 in stage 6.0 (TID 12)
17/10/08 14:29:42 INFO Executor: Finished task 0.0 in stage 6.0 (TID 12). 11571 bytes result sent to driver
17/10/08 14:29:42 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 12) in 5 ms on localhost (executor driver) (1/1)
17/10/08 14:29:42 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
17/10/08 14:29:42 INFO DAGScheduler: ResultStage 6 (show at Seeker.scala:70) finished in 0.006 s
17/10/08 14:29:42 INFO DAGScheduler: Job 4 finished: show at Seeker.scala:70, took 0.025187 s
17/10/08 14:29:42 INFO CodeGenerator: Code generated in 58.645534 ms
17/10/08 14:29:42 INFO CodeGenerator: Code generated in 32.527057 ms
+--------+--------------------+--------------------+--------------------+--------------------+
|    name|             content|              tokens|         rawFeatures|            features|
+--------+--------------------+--------------------+--------------------+--------------------+
|doc1.txt|Chiang Mai (/'tʃj...|[chiang, mai, (/'...|(262144,[866,2030...|(262144,[866,2030...|
|doc2.txt|Bali (Balinese: □...|[bali, (balinese:...|(262144,[1461,353...|(262144,[1461,353...|
|doc3.txt|Hawaii (English: ...|[hawaii, (english...|(262144,[2956,367...|(262144,[2956,367...|
|doc4.txt|Bora Bora is a 30...|[bora, bora, is, ...|(262144,[2325,797...|(262144,[2325,797...|
|doc5.txt|Aruba (/ə'ru:bə/ ...|[aruba, (/ə'ru:bə...|(262144,[632,4927...|(262144,[632,4927...|
+--------+--------------------+--------------------+--------------------+--------------------+

search>hey
i'm here
search>hey
i'm here
search>wsuup
i'm here
search>:quit
17/10/08 14:29:52 INFO SparkContext: Invoking stop() from shutdown hook
```

notes

- Class *Builder* reads the documentscreates dataframe of the schema

```
case class Doc(name: String, content: String)

new DataFrame( Seq(Doc*))
```

- Then df.col(“content”) is tokenized, Basic tokenizer is used, it’s possible to customize the regex one to avoid accents, and none alphanumeric
- Then TF-IDF is calculated using spark’s tf features hashing (TFHashing)

Next step

- Get the the query sent by the use, calculate cosine similarity or other like LSH, maybe even Doc2Vec and retrieve required results