BDA ASSIGNMENT-3

1. Explain your methodology: approach and reason clearly in the report.

Our Methodology:

- Upload the CSV data to locally hosted MongoDB server
- Using the Pyspark API we import the data from the mongo server into PySpark Dataframe.
- Now as we had computational limits, we used 10000 rows only. All the questions were merged into a single column giving 165931 questions once the duplicates were removed. Elaborated below.
- User-Defined Functions were used on every row to create 5-Shingles as the length of the document is short.
- Next, we used CountVectoriser to represent the shingles as integers. In other words, mapping the shingles to a representative integer value.
- After this step, 100-dimensional signatures were created using 100 hash functions. The choice of hash functions is explained below:

  MinHash function family used: **(ax+1) mod n**, where a goes from 1 to 100, and n is a prime number (n=165931). X represents the shingle_id found by hashing the shingles from their string form to integers.
  A hash table size that is a prime number is the only way to guarantee that you do not accidentally re-probe a previously probed location.
  Mathematically, if the hash function h(k) is of the form ax+1 mod n where n is prime, then the probability of hashing 2 distinct keys to the same bucket is 1/n. So, when n is large like in our case, this probability decreases significantly and that is what we want from our hash function.

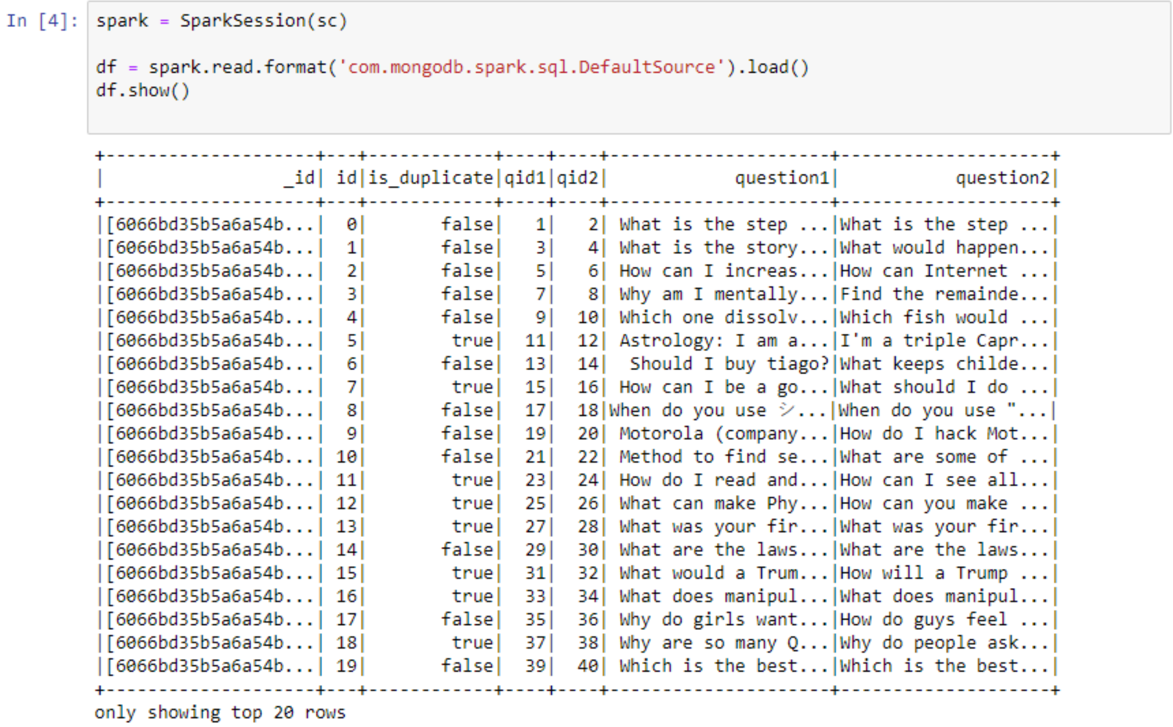
- In the LSH part we used bands=10 ⇒ r=#hash_fns/bands= 100/10 =10. Using more rows implies we'll incorporate more dimensions from the signature while choosing the bucket to put the document into. This improves the probability that more similar documents are hashed into the same bucket for a band.
- Next, we calculate the candidate pairs corresponding to a bucket as per the LSH algorithm.
- Using the "is_duplicate" field in train.csv as the ground truth labels, we then calculate how well our LSH algorithm has performed, using precision and recall.

1. Download dataset (train csv)

Done

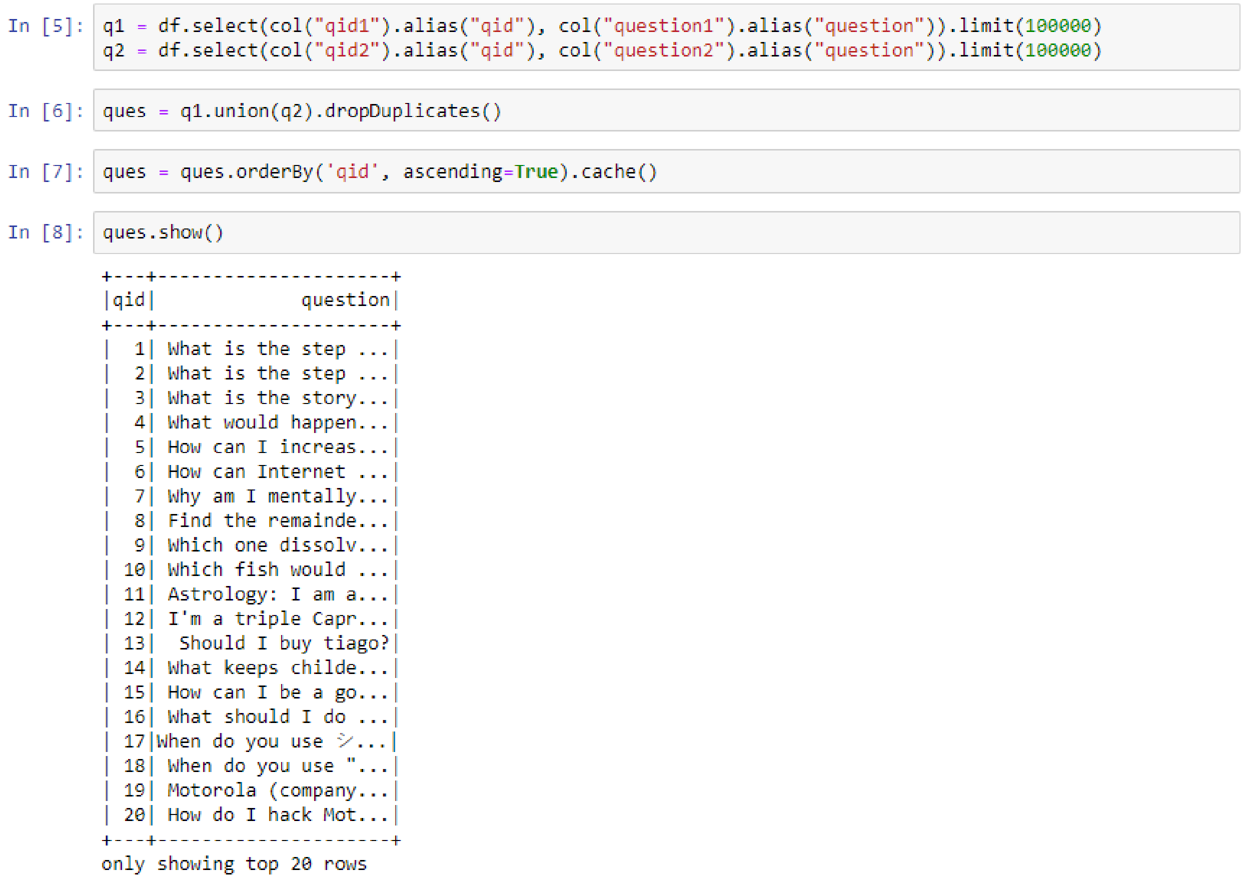
2. Do the preprocessing if necessary

As there were 2 rows in the train.csv file given, however for our LSH algorithm, we needed to merge the two rows into one and drop the duplicates. This can be seen below:



```
In [4]: spark = SparkSession(sc)

        df = spark.read.format('com.mongodb.spark.sql.DefaultSource').load()
        df.show()
```

```
+--------------------+---+------------+----+----+--------------------+--------------------+
|                 _id| id|is_duplicate|qid1|qid2|           question1|           question2|
+--------------------+---+------------+----+----+--------------------+--------------------+
|[6066bd35b5a6a54b...|  0|       false|   1|   2| What is the step ...|What is the step ...|
|[6066bd35b5a6a54b...|  1|       false|   3|   4| What is the story...|What would happen...|
|[6066bd35b5a6a54b...|  2|       false|   5|   6| How can I increas...|How can Internet ...|
|[6066bd35b5a6a54b...|  3|       false|   7|   8| Why am I mentally...|Find the remainde...|
|[6066bd35b5a6a54b...|  4|       false|   9|  10| Which one dissolv...|Which fish would ...|
|[6066bd35b5a6a54b...|  5|        true|  11|  12| Astrology: I am a...|I'm a triple Capr...|
|[6066bd35b5a6a54b...|  6|       false|  13|  14|  Should I buy tiago?|What keeps childe...|
|[6066bd35b5a6a54b...|  7|        true|  15|  16| How can I be a go...|What should I do ...|
|[6066bd35b5a6a54b...|  8|       false|  17|  18|When do you use シ...|When do you use "...|
|[6066bd35b5a6a54b...|  9|       false|  19|  20| Motorola (company...|How do I hack Mot...|
|[6066bd35b5a6a54b...| 10|       false|  21|  22| Method to find se...|What are some of ...|
|[6066bd35b5a6a54b...| 11|        true|  23|  24| How do I read and...|How can I see all...|
|[6066bd35b5a6a54b...| 12|        true|  25|  26| What can make Phy...|How can you make ...|
|[6066bd35b5a6a54b...| 13|        true|  27|  28| What was your fir...|What was your fir...|
|[6066bd35b5a6a54b...| 14|       false|  29|  30| What are the laws...|What are the laws...|
|[6066bd35b5a6a54b...| 15|        true|  31|  32| What would a Trum...|How will a Trump ...|
|[6066bd35b5a6a54b...| 16|        true|  33|  34| What does manipul...|What does manipul...|
|[6066bd35b5a6a54b...| 17|       false|  35|  36| Why do girls want...|How do guys feel ...|
|[6066bd35b5a6a54b...| 18|        true|  37|  38| Why are so many Q...|Why do people ask...|
|[6066bd35b5a6a54b...| 19|       false|  39|  40| Which is the best...|Which is the best...|
+--------------------+---+------------+----+----+--------------------+--------------------+
only showing top 20 rows
```

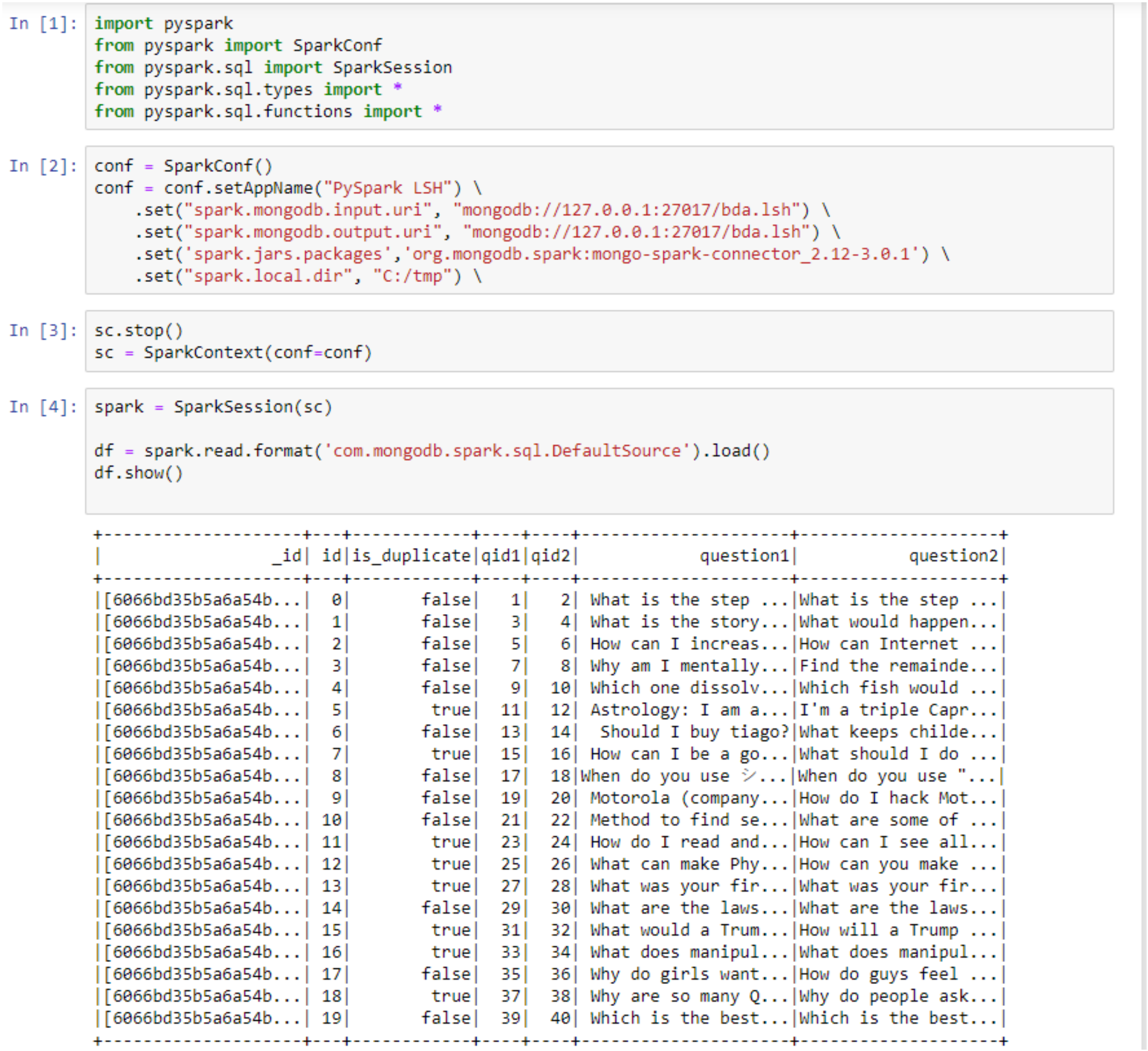


```
In [5]: q1 = df.select(col("qid1").alias("qid"), col("question1").alias("question")).limit(100000)
        q2 = df.select(col("qid2").alias("qid"), col("question2").alias("question")).limit(100000)

In [6]: ques = q1.union(q2).dropDuplicates()

In [7]: ques = ques.orderBy('qid', ascending=True).cache()

In [8]: ques.show()

        +---+--------------------+
        |qid|            question|
        +---+--------------------+
        |  1| What is the step ...|
        |  2| What is the step ...|
        |  3| What is the story...|
        |  4| What would happen...|
        |  5| How can I increas...|
        |  6| How can Internet ...|
        |  7| Why am I mentally...|
        |  8| Find the remainde...|
        |  9| Which one dissolv...|
        | 10| Which fish would ...|
        | 11| Astrology: I am a...|
        | 12| I'm a triple Capr...|
        | 13|  Should I buy tiago?|
        | 14| What keeps childe...|
        | 15| How can I be a go...|
        | 16| What should I do ...|
        | 17|When do you use シ...|
        | 18| When do you use "...|
        | 19| Motorola (company...|
        | 20| How do I hack Mot...|
        +---+--------------------+
        only showing top 20 rows
```

3. Since there are 404290 samples you may use the first 1,00,000 samples if you have compute constraints.

Yes

4. Store the data in mongodb. (20 marks)



```
In [1]: import pyspark
        from pyspark import SparkConf
        from pyspark.sql import SparkSession
        from pyspark.sql.types import *
        from pyspark.sql.functions import *
```

```
In [2]: conf = SparkConf()
        conf = conf.setAppName("PySpark LSH") \
            .set("spark.mongodb.input.uri", "mongodb://127.0.0.1:27017/bda.lsh") \
            .set("spark.mongodb.output.uri", "mongodb://127.0.0.1:27017/bda.lsh") \
            .set('spark.jars.packages','org.mongodb.spark:mongo-spark-connector_2.12-3.0.1') \
            .set("spark.local.dir", "C:/tmp") \
```

```
In [3]: sc.stop()
        sc = SparkContext(conf=conf)
```

```
In [4]: spark = SparkSession(sc)

        df = spark.read.format('com.mongodb.spark.sql.DefaultSource').load()
        df.show()
```

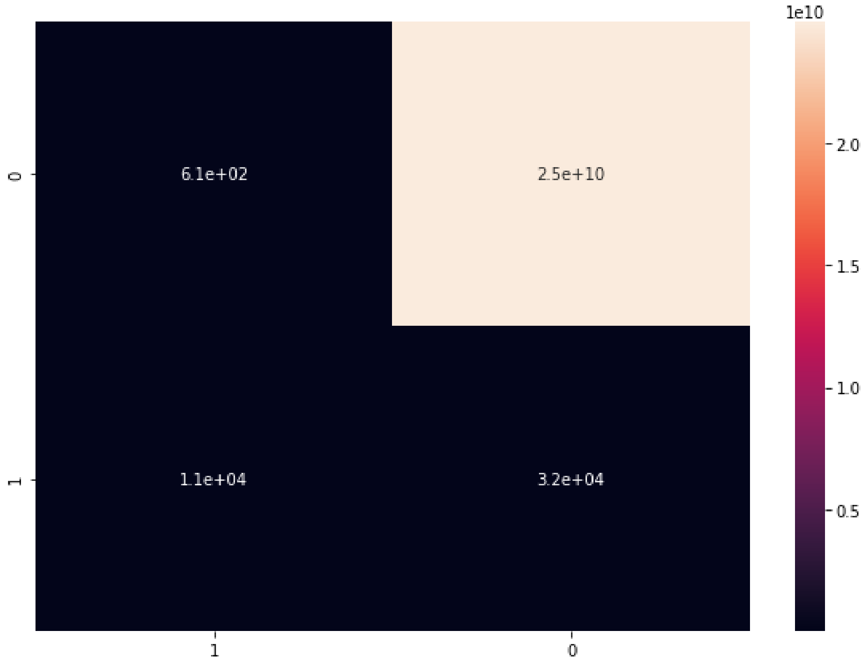
```
+--------------------+---+------------+----+----+--------------------+--------------------+
|                 _id| id|is_duplicate|qid1|qid2|           question1|           question2|
+--------------------+---+------------+----+----+--------------------+--------------------+
|[6066bd35b5a6a54b...|  0|       false|   1|   2|What is the step ...|What is the step ...|
|[6066bd35b5a6a54b...|  1|       false|   3|   4|What is the story...|What would happen...|
|[6066bd35b5a6a54b...|  2|       false|   5|   6|How can I increas...|How can Internet ...|
|[6066bd35b5a6a54b...|  3|       false|   7|   8|Why am I mentally...|Find the remainde...|
|[6066bd35b5a6a54b...|  4|       false|   9|  10|Which one dissolv...|Which fish would ...|
|[6066bd35b5a6a54b...|  5|        true|  11|  12|Astrology: I am a...|I'm a triple Capr...|
|[6066bd35b5a6a54b...|  6|       false|  13|  14| Should I buy tiago?|What keeps childe...|
|[6066bd35b5a6a54b...|  7|        true|  15|  16|How can I be a go...|What should I do ...|
|[6066bd35b5a6a54b...|  8|       false|  17|  18|When do you use シ...|When do you use "...|
|[6066bd35b5a6a54b...|  9|       false|  19|  20|Motorola (company...|How do I hack Mot...|
|[6066bd35b5a6a54b...| 10|       false|  21|  22|Method to find se...|What are some of ...|
|[6066bd35b5a6a54b...| 11|        true|  23|  24|How do I read and...|How can I see all...|
|[6066bd35b5a6a54b...| 12|        true|  25|  26|What can make Phy...|How can you make ...|
|[6066bd35b5a6a54b...| 13|        true|  27|  28|What was your fir...|What was your fir...|
|[6066bd35b5a6a54b...| 14|       false|  29|  30|What are the laws...|What are the laws...|
|[6066bd35b5a6a54b...| 15|        true|  31|  32|What would a Trum...|How will a Trump ...|
|[6066bd35b5a6a54b...| 16|        true|  33|  34|What does manipul...|What does manipul...|
|[6066bd35b5a6a54b...| 17|       false|  35|  36|Why do girls want...|How do guys feel ...|
|[6066bd35b5a6a54b...| 18|        true|  37|  38|Why are so many Q...|Why do people ask...|
|[6066bd35b5a6a54b...| 19|       false|  39|  40|Which is the best...|Which is the best...|
+--------------------+---+------------+----+----+--------------------+--------------------+
```

5. Implement LSH in spark (inbuilt methods implementing LSH are not allowed) (50 marks)
Done in Code

6. Detect duplicates using spark LSH implementation on data from mongodb and report the questions that are duplicates. Note you are using LSH to avoid comparing a question with all other questions by generating candidate pairs. After detecting duplicates use the “is duplicate” label to **get ground truth and report precision and Recall. (20 marks)**

```
Accuracy: 0.9997855571114223
Precision: 0.02767727930535456
Recall: 0.030797101449275364
F1-score: 0.029153963414634154
```

7. Plot confusion matrix. (10 marks)



Make a section “Learning”, which describes your learning in doing this assignment.

- Pyspark library and its functions
- Integrate these databases with apache spark to fetch data into RDD.
- RDD and SQL query engines and difference between the two and how to use them via the spark framework.
- Aggregate pipeline query mongodb engine queries.
- How to use spark RDD for computations on large datasets, and observe that it is faster.
- Use UDF to run functions on RDD, as it performs computations on all elements.