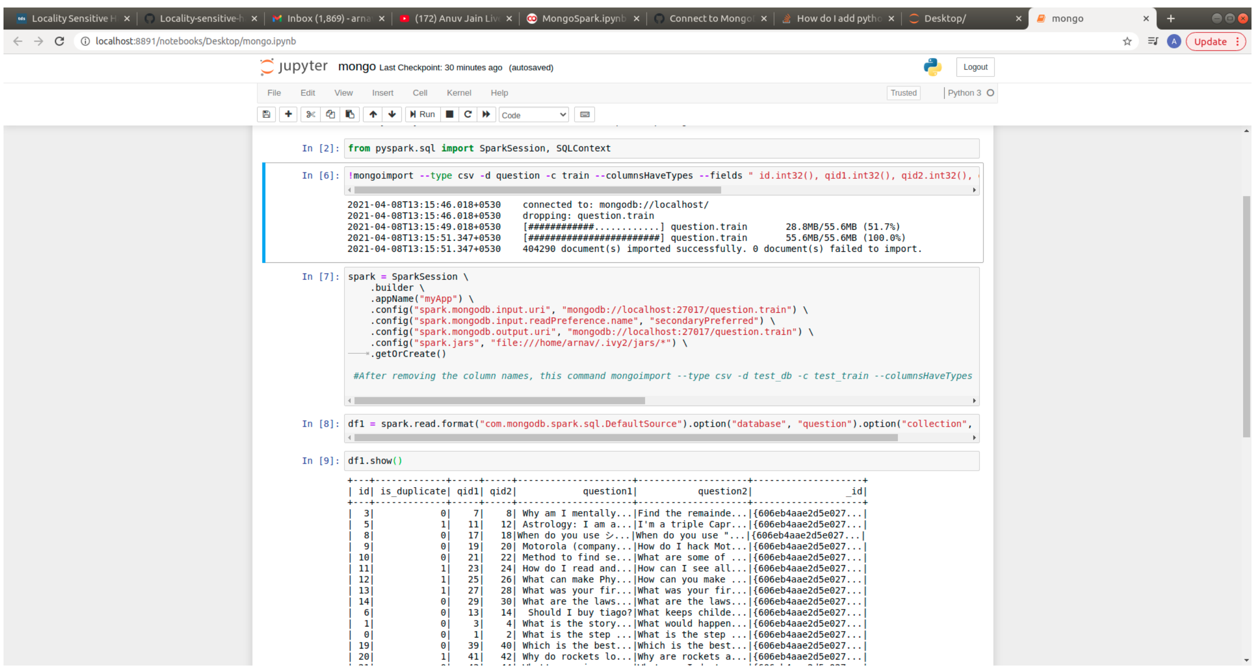
LSH using PySpark

Garvita Jain 2018034 | Arnav Tandon 2018278

Store data in MongoDB

Command : !mongoimport --type csv -d question -c train --columnsHaveTypes --fields " id.int32(), qid1.int32(), qid2.int32(), question1.string(), question2.string(), is_duplicate.int32()" --drop train.csv

Output



Reading data normally :
df1 = spark.read.format("com.mongodb.spark.sql.DefaultSource").option("database", "question").option("collection", "train").load()

df2_rdd = spark.read.format("com.mongodb.spark.sql.DefaultSource").option("database", "question").option("collection", "train").load().rdd jupyter mongo Last Checkpoint: 35 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Help Trusted Python 3

```
In [8]: .tSource").option("database", "question").option("collection", "train").load()).option("collection", "train").load()
```

```
In [9]: df1.show()
```

```
+---+------------+-----+-----+--------------------+--------------------+--------------------+
| id|is_duplicate| qid1| qid2|           question1|           question2|                 _id|
+---+------------+-----+-----+--------------------+--------------------+--------------------+
|  3|           0|    7|    8| Why am I mentally...|Find the remainde...|{606eb4aae2d5e027...|
|  5|           1|   11|   12| Astrology: I am a...|I'm a triple Capr...|{606eb4aae2d5e027...|
|  8|           0|   17|   18|When do you use シ...|When do you use "...|{606eb4aae2d5e027...|
|  9|           0|   19|   20| Motorola (company...|How do I hack Mot...|{606eb4aae2d5e027...|
| 10|           0|   21|   22| Method to find se...|What are some of ...|{606eb4aae2d5e027...|
| 11|           1|   23|   24| How do I read and...|How can I see all...|{606eb4aae2d5e027...|
| 12|           1|   25|   26| What can make Phy...|How can you make ...|{606eb4aae2d5e027...|
| 13|           1|   27|   28| What was your fir...|What was your fir...|{606eb4aae2d5e027...|
| 14|           0|   29|   30| What are the laws...|What are the laws...|{606eb4aae2d5e027...|
|  6|           0|   13|   14|  Should I buy tiago?|What keeps childe...|{606eb4aae2d5e027...|
|  1|           0|    3|    4| What is the story...|What would happen...|{606eb4aae2d5e027...|
|  0|           0|    1|    2| What is the step ...|What is the step ...|{606eb4aae2d5e027...|
| 19|           0|   39|   40| Which is the best...|Which is the best...|{606eb4aae2d5e027...|
| 20|           1|   41|   42| Why do rockets lo...|Why are rockets a...|{606eb4aae2d5e027...|
| 21|           0|   43|   44| What's causing so...|What can I do to ...|{606eb4aae2d5e027...|
| 22|           0|   45|   46| What are the ques...|Which question sh...|{606eb4aae2d5e027...|
| 23|           0|   47|   48| How much is 30 kV...|Where can I find ...|{606eb4aae2d5e027...|
| 24|           0|   49|   50| What does it mean...|How many times a ...|{606eb4aae2d5e027...|
| 25|           0|   51|   52| What are some tip...|What are some tip...|{606eb4aae2d5e027...|
| 26|           0|   53|   54| What is web appli...|What is the web a...|{606eb4aae2d5e027...|
+---+------------+-----+-----+--------------------+--------------------+--------------------+
only showing top 20 rows
```

```
In [12]: df2_rdd = spark.read.format("com.mongodb.spark.sql.DefaultSource").option("database", "question").option("collectio
```

```
In [13]: df2_rdd.collect()
```

```
Out[13]: [Row( id=3,  is_duplicate=0,  qid1=7,  qid2=8,  question1='Why am I mentally very lonely? How can I solve it?',
question2='Find the remainder when [math]23^{24}[/math] is divided by 24,23?', _id=Row(oid='606eb4aae2d5e027f20d
4d31')),
 Row( id=5,  is_duplicate=1,  qid1=11,  qid2=12,  question1='Astrology: I am a Capricorn Sun Cap moon and cap ri
sing...what does that say about me?',  question2="I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn)
What does this say about me?",  _id=Row(oid='606eb4aae2d5e027f20d4d32')),
 Row( id=8,  is_duplicate=0,  qid1=17,  qid2=18,  question1='When do you use シ instead of し?',  question2='When
do you use "&" instead of "and"?',  _id=Row(oid='606eb4aae2d5e027f20d4d33')),
 Row( id=9,  is_duplicate=0,  qid1=19,  qid2=20,  question1='Motorola (company): Can I hack my Charter Motorolla
DCX3400?',  question2='How do I hack Motorola DCX3400 for free internet?', _id=Row(oid='606eb4aae2d5e027f20d4d3
```

GitHub - nicol × | Desktop/ × | python 3.x - H × | LSH Pyspark × | Desktop/ × | mongo × | Downloads/ × | code × | spark access × | Locality-sensi × | LSH on Spark ×

localhost:8889/notebooks/Downloads/code.ipynb

jupyter code (autosaved)

Logout

File Edit View Insert Cell Kernel Help Connecting to kernel Trusted Python 3

```
In [3]: spark = SparkSession \
            .builder \
            .appName("Assignment3") \
            .config("spark.mongodb.input.uri", "mongodb://localhost:27017/question.train") \
            .config("spark.mongodb.output.uri", "mongodb://localhost:27017/question.train") \
            .config("spark.jars", "file:///home/arnav/.ivy2/jars/*") \
            .getOrCreate()
```

```
In [4]: !mongoimport --type csv -d question -c train --headerline --maintainInsertionOrder --drop train.csv
```

```
2021-04-09T20:16:13.067+0530    connected to: mongodb://localhost/
2021-04-09T20:16:13.067+0530    dropping: question.train
2021-04-09T20:16:16.067+0530    [#############..........] question.train       33.1MB/60.5MB (54.7%)
2021-04-09T20:16:19.067+0530    [######################.] question.train       59.0MB/60.5MB (97.5%)
2021-04-09T20:16:19.213+0530    [#######################] question.train       60.5MB/60.5MB (100.0%)
2021-04-09T20:16:19.213+0530    404290 document(s) imported successfully. 0 document(s) failed to import.
```

```
In [5]: spark.sparkContext.uiWebUrl
```

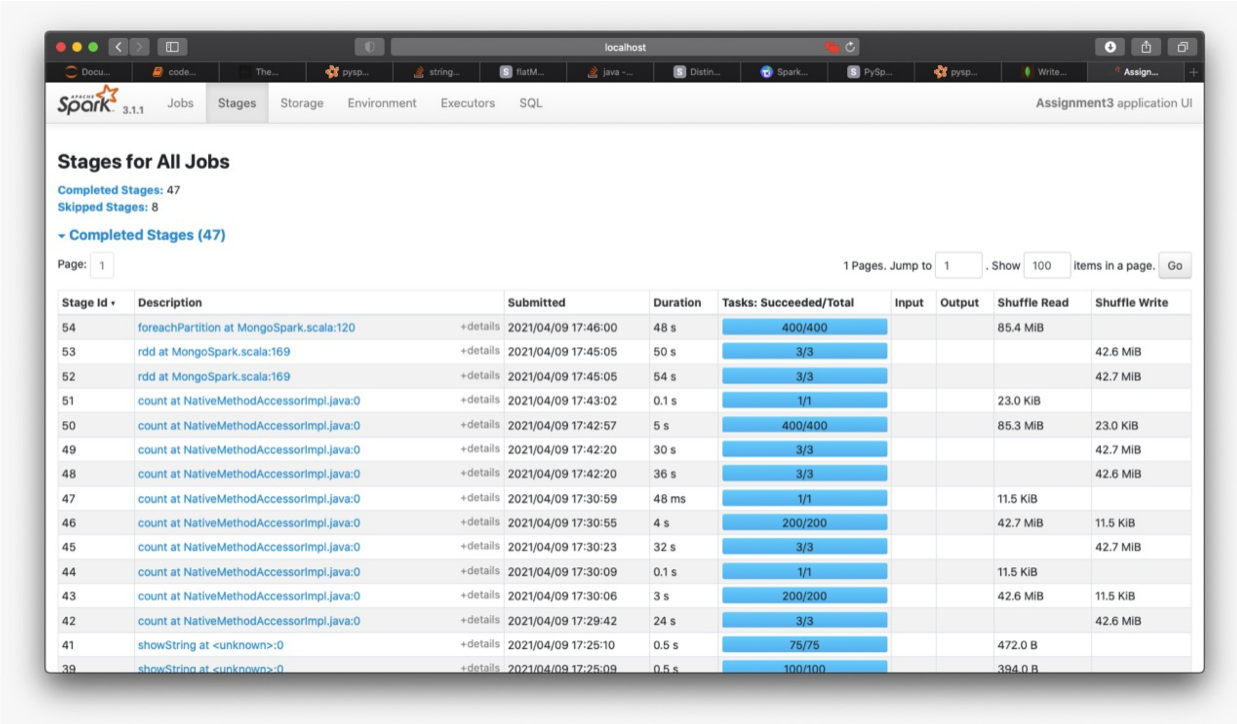
```
Out[5]: 'http://192.168.1.43:4041'
```

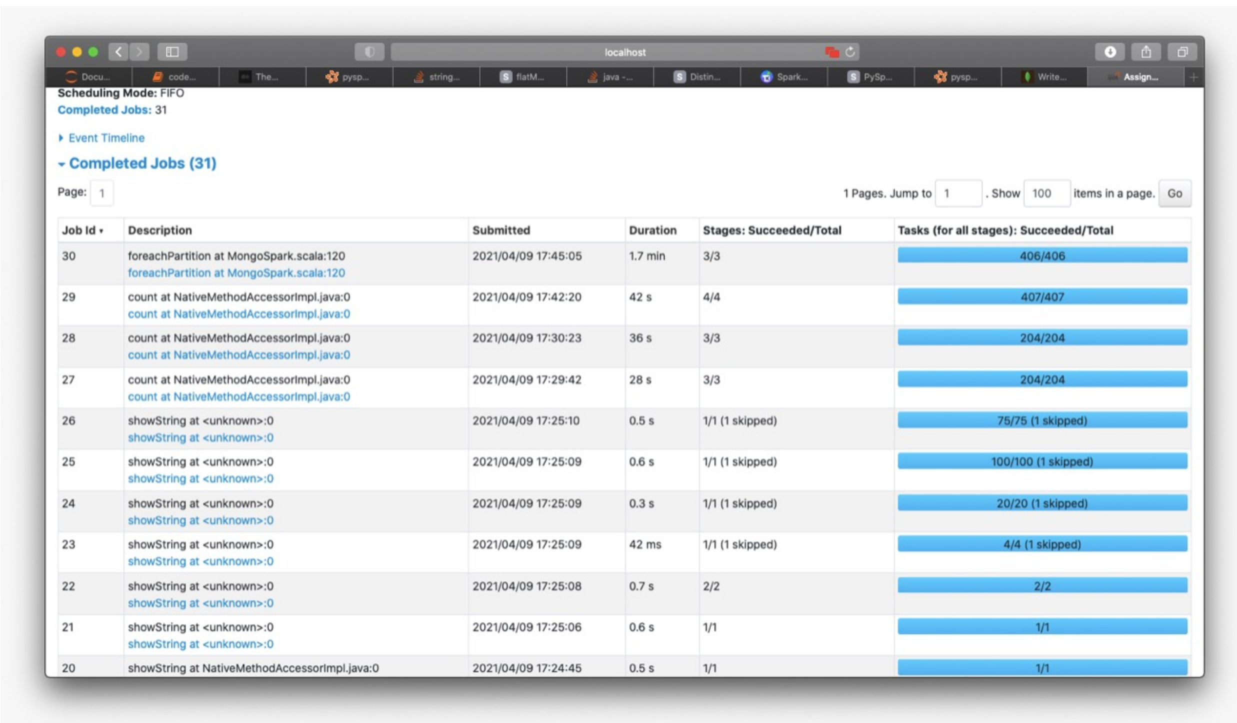
```
In [6]: df1 = spark.read.format("com.mongodb.spark.sql.DefaultSource").option("database", "question") \
            .option("collection", "train").load().limit(100000)
```

```
In [7]: # df1.count()
        df1.show()
```

```
+--------------------+---+------------+----+----+--------------------+--------------------+
|                 _id| id|is_duplicate|qid1|qid2|           question1|           question2|
+--------------------+---+------------+----+----+--------------------+--------------------+
|{607068b589c0716f...|  0|           0|   1|   2| What is the step ...|What is the step ...|
|{607068b589c0716f...|  1|           0|   3|   4| What is the story...|What would happen...|
|{607068b589c0716f...|  2|           0|   5|   6| How can I increas...|How can Internet ...|
|{607068b589c0716f...|  3|           0|   7|   8| Why am I mentally...|Find the remainde...|
|{607068b589c0716f...|  4|           0|   9|  10| Which one dissolv...|Which fish would ...|
|{607068b589c0716f...|  5|           1|  11|  12| Astrology: I am a...|I'm a triple Capr...|
|{607068b589c0716f...|  6|           0|  13|  14|  Should I buy tiago?|What keeps childe...|
|{607068b589c0716f...|  7|           1|  15|  16| How can I be a go...|What should I do ...|
|{607068b589c0716f...|  8|           0|  17|  18|When do you use シ...|When do you use "...|
|{607068b589c0716f...|  9|           0|  19|  20| Motorola (company...|How do I hack Mot...|
|{607068b589c0716f...| 10|           0|  21|  22| Method to find se...|What are some of ...|
|{607068b589c0716f...| 11|           1|  23|  24| How do I read and...|How can I see all...|
|{607068b589c0716f...| 12|           1|  25|  26| What can make Phy...|How can you make ...|
|{607068b589c0716f...| 13|           1|  27|  28| What was your fir...|What was your fir...|
|{607068b589c0716f...| 14|           0|  29|  30| What are the laws...|What are the laws...|
|{607068b589c0716f...| 15|           1|  31|  32| What would a Trum...|How will a Trump ...|
```

Apache Spark Web UI screenshots :





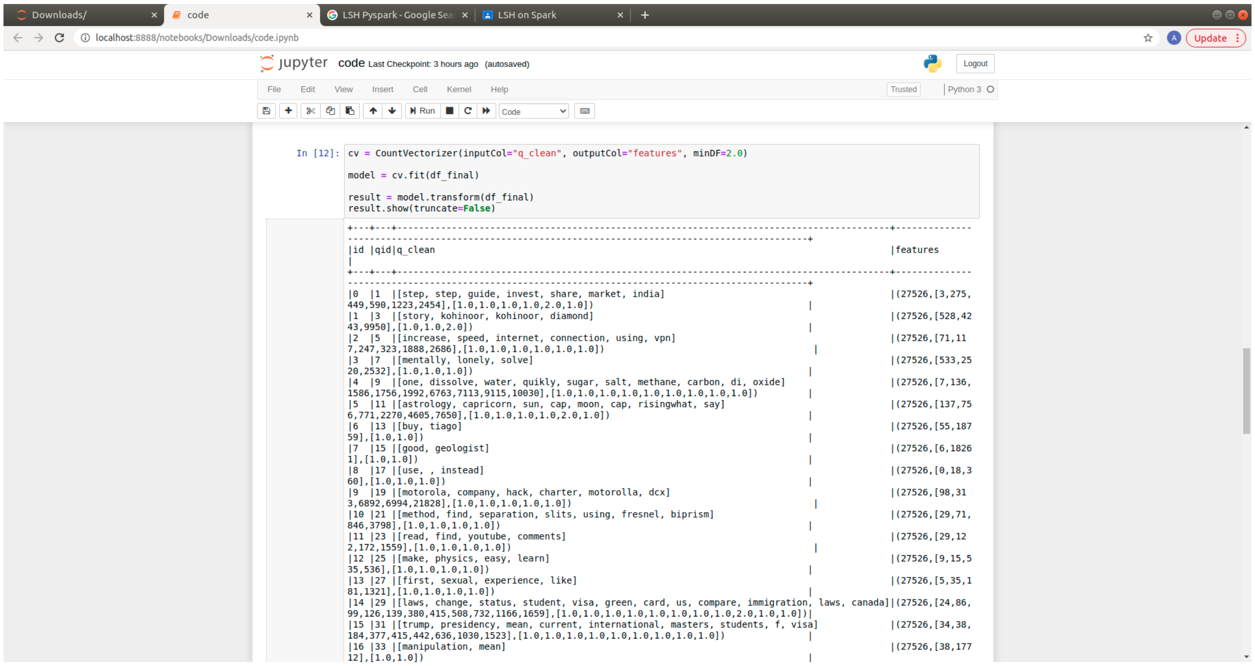
Preprocessing the data :

1) We have cleaned the data by first lowering all the document content and then by using regexp_replace for removal of integers and symbols.
2) We have then tokenized the data on the dataframe which was obtained after cleaning the data in the above step.

3) Stopwords are removed from the pyspark.ml.feature library to remove the unwanted words which do not contribute significantly to the text obtained.

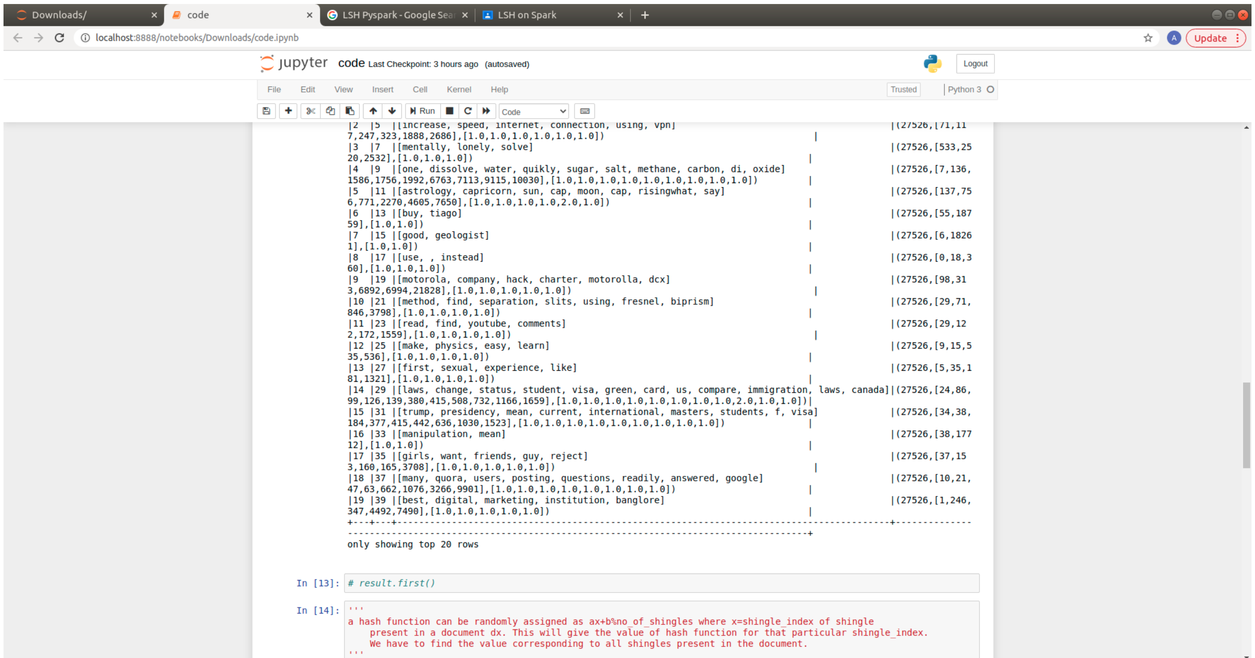
Shingling :

1) Shingles were created by splitting by whitespaces.
2) For each document, a unique set of shingles was created and added to a new table.



Sparse Vector of the document :

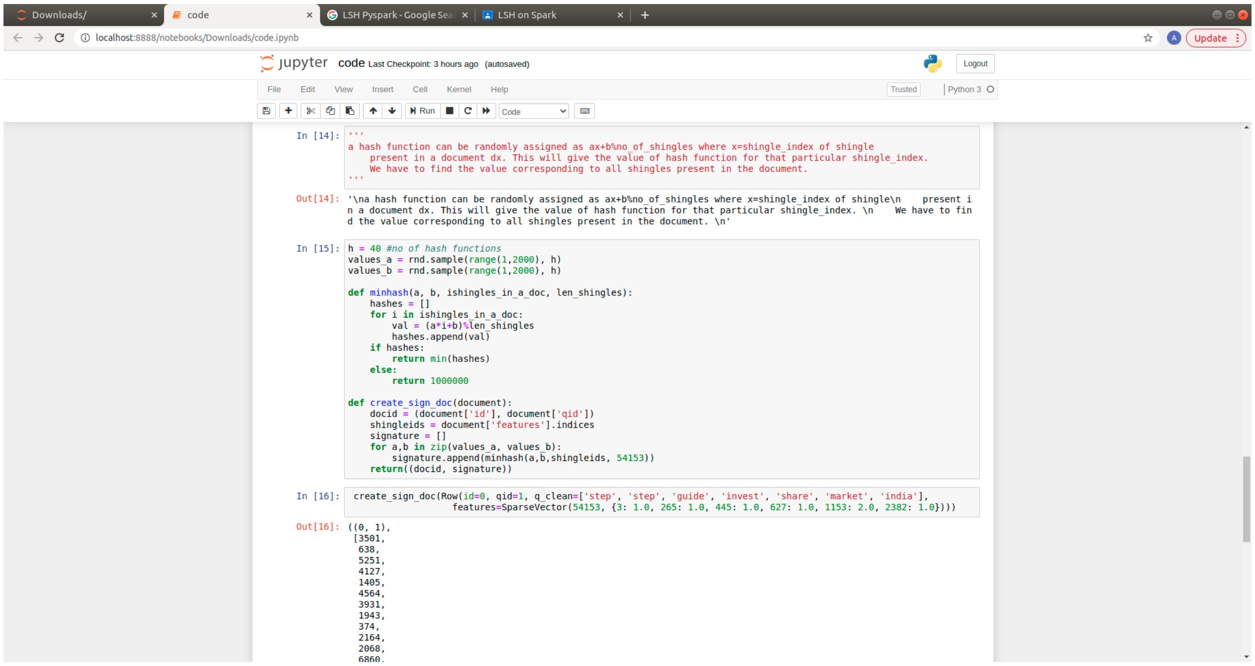
1) For each document, a sparse matrix was created to represent the presence/absence of each shingle present in the entire dataset.
2) The SparseVector provided the number of shingles/rows in the vector, the list of indices at which the shingle was present for a particular question, and its frequency.



Made Signature Matrix using random Hash functions :

We have taken h = 30 which are our number of hash functions and we have defined our hash function as val = (a*i+b) % (length of the shingles) where a,b are random numbers.

A hash function can be randomly assigned as ax+b%no_of_shingles where x=shingle_index of shingle present in a document dx. This will give the value of hash function for that particular shingle_index. We have to find the value corresponding to all shingles present in the document.



LSH Implementation

When we have to compare two things and the size is too it takes O(n^2) time, but there is a better way of comparing every possible pair by **Locality Sensititve Hashing** (LSH) in O(n).

As there are 2 lac short documents for comparison, so we are taking k as 5 (commonly used size for shorter documents).

We use the band structure on the signature matrix made by us earlier, An if there are n rows, so it into b bands and with each row having width w, so we get :

1) N = b*w
2) Let's say the true similarity score is p when comparison is done between a pair,

   Probability of one band not matching = $1-p^w$

   Probability of number of bands that match = $(1-p^w)^b$

   So Probability of at least one band matches = $1-(1-p^w)^b$

After implementing the LSH in PySpark we got the candidate pairs, meaning they have a similarity score above the threshold and there are multiple candidate pairs with same band_hash as shown in the figure below.

B and W make sense here and they do the opposite things as when we increase b it gives documents more chances to match, so we get pairs with lower similarity scores in the candidate pair list. But increasing W makes the match criteria stricter, so it is restricted to higher similarity scores.

```
In [89]: lsh_df.groupby('band_no', 'band_hash').agg(collect_list("doc_id").alias("candidate_pairs")).show()
```

```
+-------+--------------------+--------------------+
|band_no|           band_hash|     candidate_pairs|
+-------+--------------------+--------------------+
|      0|-9098964674786157832|      [{1013, 2021}]|
|      0|-9097076610132250771|[{92908, 155452},...|
|      0|-9075903486970713091|[{11085, 21433}, ...|
|      0|-9075133033449877333|    [{45196, 81020}]|
|      0|-9063007082634758990|   [{94866, 158331}]|
|      0|-9036711468021138188|        [{463, 924}]|
|      0|-8991721065649749160|   [{59100, 103569}]|
|      0|-8950445860725111188|   [{98166, 163177}]|
|      0|-8847112719701112263|   [{97503, 162224}]|
|      0|-8840263500075329849|   [{76388, 130589}]|
|      0|-8837973820842852960|    [{28236, 52378}]|
|      0|-8820194762220627155|[{33423, 61408}, ...|
|      0|-8767193704393368565|   [{83793, 141777}]|
|      0|-8733549606059257863|[{11752, 22682}, ...|
|      0|-8724898006211925323|   [{87139, 146804}]|
|      0|-8723675710942797727|    [{12071, 23284}]|
|      0|-8709671219524383574|   [{84195, 142372}]|
|      0|-8701223129789196813|    [{29053, 53796}]|
|      0|-8686888307652572036|    [{11143, 21544}]|
|      0|-8671445867209440873|   [{73614, 126310}]|
+-------+--------------------+--------------------+
only showing top 20 rows
```

Precision, recall and confusion matrix :

1. TN / True Negative: case was negative and predicted negative
2. TP / True Positive: case was positive and predicted positive
3. FN / False Negative: case was positive but predicted negative
4. FP / False Positive: case was negative but predicted positive

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives} \qquad precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Learnings from the assignment :

- We implemented the LSH theory which was taught in class.
- We became familiar with Pyspark programming and tried our hands on basic Pyspark functionalities and libraries available in it.

- We learned about the various preprocessing libraries and techniques that have to be applied to the data in order to get a better picture of the real content that we wish to study as there are lots of symbols, URLs and other punctuations which make it difficult to get the real content from the data.
- We initially got some errors while getting the confusion matrix
- We tried our best to improve the accuracy and get the is_duplicate value similar to the results as achieved by us from our own implementation of LSH which helped us explore various ways of improving the efficiency
- Initially we had difficulty in setting up the connection but we learned it about how to do it and then we uploaded our data to MongoDB and then read that data using spark into a dataframe and then we worked on this obtained dataframe.