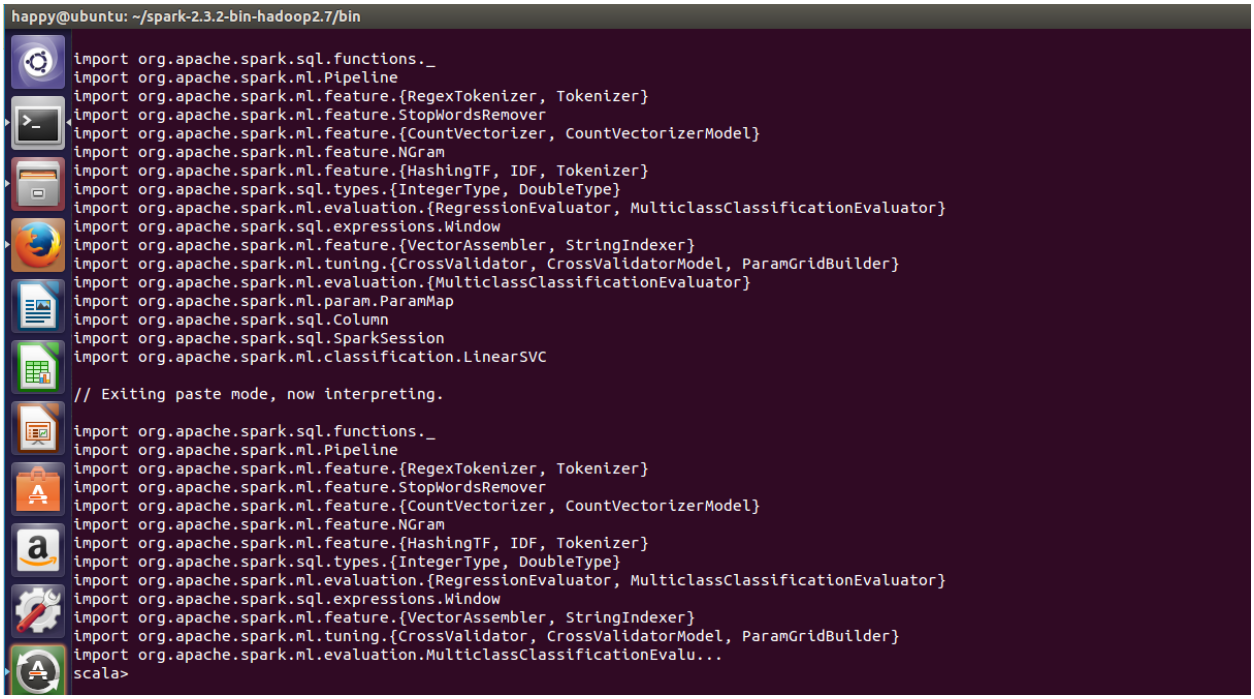
```
import org.apache.spark.sql.functions._
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.feature.{RegexTokenizer, Tokenizer}
import org.apache.spark.ml.feature.StopWordsRemover
import org.apache.spark.ml.feature.{CountVectorizer, CountVectorizerModel}
import org.apache.spark.ml.feature.NGram
import org.apache.spark.ml.feature.{HashingTF, IDF, Tokenizer}
import org.apache.spark.sql.types.{IntegerType, DoubleType}
import org.apache.spark.ml.evaluation.{RegressionEvaluator, MulticlassClassificationEvaluator}
import org.apache.spark.sql.expressions.Window
import org.apache.spark.ml.feature.{VectorAssembler, StringIndexer}
import org.apache.spark.ml.tuning.{CrossValidator, CrossValidatorModel, ParamGridBuilder}
import org.apache.spark.ml.evaluation.{MulticlassClassificationEvaluator}
import org.apache.spark.ml.param.ParamMap
import org.apache.spark.sql.Column
import org.apache.spark.sql.SparkSession
import org.apache.spark.ml.classification.LinearSVC
```



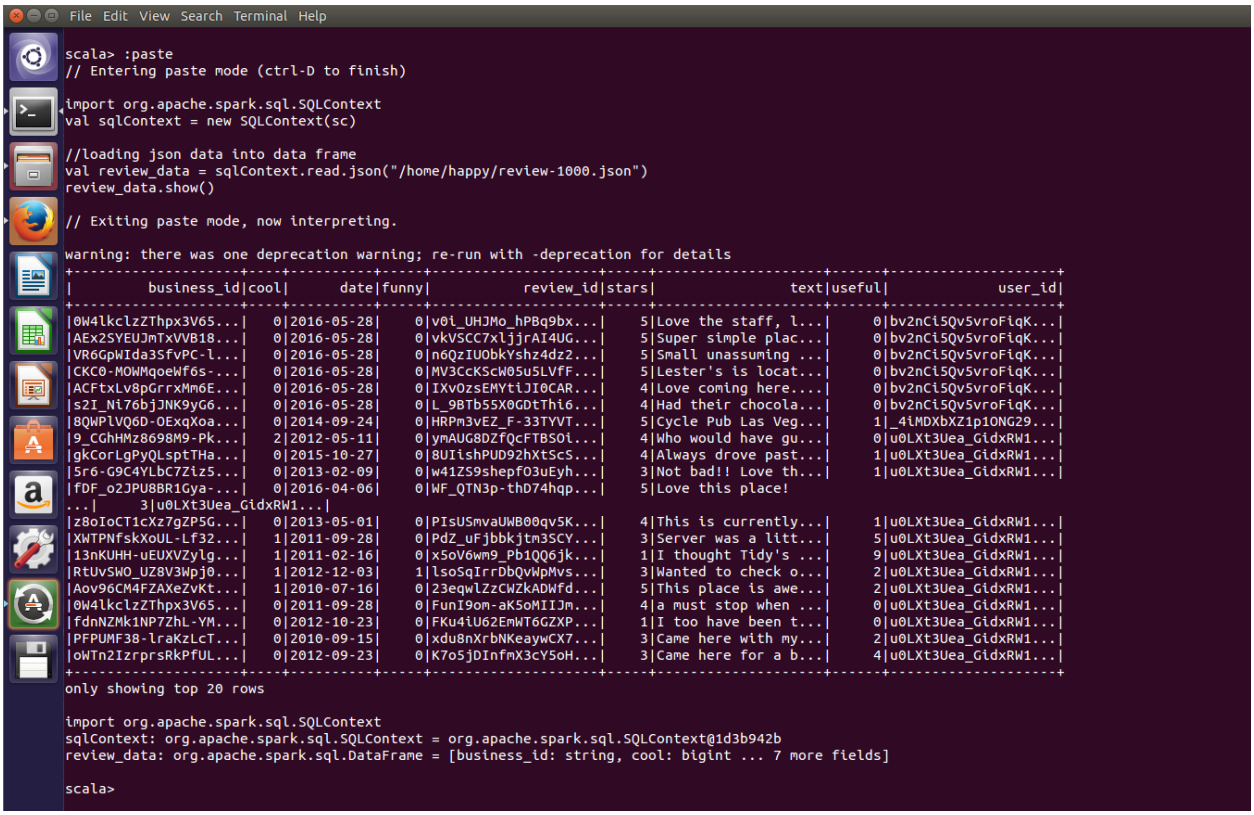
import org.apache.spark.sql.SQLContext

val sqlContext = new SQLContext(sc)

//loading json data into data frame

val review_data = sqlContext.read.json("/home/happy/review-1000.json")

review_data.show()



```
File  Edit  View  Search  Terminal  Help

scala> :paste
// Entering paste mode (ctrl-D to finish)

import org.apache.spark.sql.SQLContext
val sqlContext = new SQLContext(sc)

//loading json data into data frame
val review_data = sqlContext.read.json("/home/happy/review-1000.json")
review_data.show()

// Exiting paste mode, now interpreting.

warning: there was one deprecation warning; re-run with -deprecation for details
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+
|         business_id|cool|      date|funny|           review_id|stars|                text|useful|             user_id|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+
|0W4lkclzZThpx3V65...|   0|2016-05-28|    0|v0i_UHJMo_hPBq9bx...|    5|Love the staff, l...|     0|bv2nCi5Qv5vroFiqK...|
|AEx2SYEUJmTxVVB18...|   0|2016-05-28|    0|vkVSCC7xljjrAI4UG...|    5|Super simple plac...|     0|bv2nCi5Qv5vroFiqK...|
|VR6GpWIda3SfvPC-l...|   0|2016-05-28|    0|n6QzIUObkYshz4dz2...|    5|Small unassuming ...|     0|bv2nCi5Qv5vroFiqK...|
|CKC0-MOWMqoeWf6s-...|   0|2016-05-28|    0|MV3CcKScW05u5LVfF...|    5|Lester's is locat...|     0|bv2nCi5Qv5vroFiqK...|
|ACFtxLv8pGrrxMm6E...|   0|2016-05-28|    0|IXvOzsEMYtiJI0CAR...|    4|Love coming here....|     0|bv2nCi5Qv5vroFiqK...|
|s2I_Ni76bjJNK9yG6...|   0|2016-05-28|    0|L_9BTb55X0GDtThi6...|    4|Had their chocola...|     0|bv2nCi5Qv5vroFiqK...|
|8QWPlVQ6D-OExqXoa...|   0|2014-09-24|    0|HRPm3vEZ_F-33TYVT...|    5|Cycle Pub Las Veg...|     1|_4iMDXbXZ1p1ONG29...|
|9_CGhHMz8698M9-Pk...|   2|2012-05-11|    0|ymAUG8DZfQcFTBSOi...|    4|Who would have gu...|     0|u0LXt3Uea_GidxRW1...|
|gkCorLgPyQLsptTHa...|   0|2015-10-27|    0|8UIishPUD92hXtScS...|    4|Always drove past...|     1|u0LXt3Uea_GidxRW1...|
|5r6-G9C4YLbC7Ziz5...|   0|2013-02-09|    0|w41ZS9shepfO3uEyh...|    3|Not bad!! Love th...|     1|u0LXt3Uea_GidxRW1...|
|fDF_o2JPU8BR1Gya-...|   0|2016-04-06|    0|WF_QTN3p-thD74hqp...|    5|Love this place!
...|     3|u0LXt3Uea_GidxRW1...|
|z8oIoCT1cXz7gZP5G...|   0|2013-05-01|    0|PIsUSmvaUWB00qv5K...|    4|This is currently...|     1|u0LXt3Uea_GidxRW1...|
|XWTPNfskXoUL-Lf32...|   1|2011-09-28|    0|PdZ_uFjbbkjtm3SCY...|    3|Server was a litt...|     5|u0LXt3Uea_GidxRW1...|
|13nKUHH-uEUXVZylg...|   1|2011-02-16|    0|x5oV6wm9_Pb1QQ6jk...|    1|I thought Tidy's ...|     9|u0LXt3Uea_GidxRW1...|
|RtUvSWO_UZ8V3Wpj0...|   1|2012-12-03|    1|lsoSqIrrDbQvWpMvs...|    3|Wanted to check o...|     2|u0LXt3Uea_GidxRW1...|
|Aov96CM4FZAXeZvKt...|   1|2010-07-16|    0|23eqwlZzCWZkADWfd...|    5|This place is awe...|     2|u0LXt3Uea_GidxRW1...|
|0W4lkclzZThpx3V65...|   0|2011-09-28|    0|FunI9om-aK5oMIIJm...|    4|a must stop when ...|     0|u0LXt3Uea_GidxRW1...|
|fdnNZMk1NP7ZhL-YM...|   0|2012-10-23|    0|FKu4iU62EmWT6GZXP...|    1|I too have been t...|     0|u0LXt3Uea_GidxRW1...|
|PFPUMF38-lraKzLcT...|   0|2010-09-15|    0|xdu8nXrbNKeaywCX7...|    3|Came here with my...|     2|u0LXt3Uea_GidxRW1...|
|oWTn2IzrprsRkPfUL...|   0|2012-09-23|    0|K7o5jDInfmX3cY5oH...|    3|Came here for a b...|     4|u0LXt3Uea_GidxRW1...|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+
only showing top 20 rows

import org.apache.spark.sql.SQLContext
sqlContext: org.apache.spark.sql.SQLContext = org.apache.spark.sql.SQLContext@1d3b942b
review_data: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 7 more fields]

scala>
```

//removing all review with 3 stars

var review_filter_3star=review_data.filter(col("stars") !=="3")

review_filter_3star.show()

```
File  Edit  View  Search  Terminal  Help
scala> :paste
// Entering paste mode (ctrl-D to finish)

//removing all review with 3 stars
var review_filter_3star=review_data.filter(col("stars") !=="3")
review_filter_3star.show()

// Exiting paste mode, now interpreting.

warning: there was one deprecation warning; re-run with -deprecation for details
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+
|         business_id|cool|      date|funny|           review_id|stars|                text|useful|             user_id|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+
|0W4lkclzZThpx3V65...|   0|2016-05-28|    0|v0i_UHJMo_hPBq9bx...|    5|Love the staff, l...|     0|bv2nCi5Qv5vroFiqK...|
|AEx2SYEUJmTxVVB18...|   0|2016-05-28|    0|vkVSCC7xljjrAI4UG...|    5|Super simple plac...|     0|bv2nCi5Qv5vroFiqK...|
|VR6GpWIda3SfvPC-l...|   0|2016-05-28|    0|n6QzIUObkYshz4dz2...|    5|Small unassuming ...|     0|bv2nCi5Qv5vroFiqK...|
|CKC0-MOWMqoeWf6s-...|   0|2016-05-28|    0|MV3CcKScW05u5LVfF...|    5|Lester's is locat...|     0|bv2nCi5Qv5vroFiqK...|
|ACFtxLv8pGrrxMm6E...|   0|2016-05-28|    0|IXvOzsEMYtiJI0CAR...|    4|Love coming here....|     0|bv2nCi5Qv5vroFiqK...|
|s2I_Ni76bjJNK9yG6...|   0|2016-05-28|    0|L_9BTb55X0GDtThi6...|    4|Had their chocola...|     0|bv2nCi5Qv5vroFiqK...|
|8QWPlVQ6D-OExqXoa...|   0|2014-09-24|    0|HRPm3vEZ_F-33TYVT...|    5|Cycle Pub Las Veg...|     1|_4iMDXbXZ1p1ONG29...|
|9_CGhHMz8698M9-Pk...|   2|2012-05-11|    0|ymAUG8DZfQcFTBSOi...|    4|Who would have gu...|     0|u0LXt3Uea_GidxRW1...|
|gkCorLgPyQLsptTHa...|   0|2015-10-27|    0|8UIishPUD92hXtScS...|    4|Always drove past...|     1|u0LXt3Uea_GidxRW1...|
|fDF_o2JPU8BR1Gya-...|   0|2016-04-06|    0|WF_QTN3p-thD74hqp...|    5|Love this place!
...|     3|u0LXt3Uea_GidxRW1...|
|z8oIoCT1cXz7gZP5G...|   0|2013-05-01|    0|PIsUSmvaUWB00qv5K...|    4|This is currently...|     1|u0LXt3Uea_GidxRW1...|
|13nKUHH-uEUXVZylg...|   1|2011-02-16|    0|x5oV6wm9_Pb1QQ6jk...|    1|I thought Tidy's ...|     9|u0LXt3Uea_GidxRW1...|
|Aov96CM4FZAXeZvKt...|   1|2010-07-16|    0|23eqwlZzCWZkADWfd...|    5|This place is awe...|     2|u0LXt3Uea_GidxRW1...|
|0W4lkclzZThpx3V65...|   0|2011-09-28|    0|FunI9om-aK5oMIIJm...|    4|a must stop when ...|     0|u0LXt3Uea_GidxRW1...|
|fdnNZMk1NP7ZhL-YM...|   0|2012-10-23|    0|FKu4iU62EmWT6GZXP...|    1|I too have been t...|     0|u0LXt3Uea_GidxRW1...|
|zgQHtqX0gqMw1nlBZ...|   1|2012-10-30|    2|WYDFJOBOl7cycd7gN...|    1|really excited to...|     9|u0LXt3Uea_GidxRW1...|
|RWGI8u00x5GghYCEz...|   1|2011-10-14|    0|Kki2nwtP8U2qmWwRv...|    4|This place remind...|     0|u0LXt3Uea_GidxRW1...|
|hjk3ox7w1akbEuOgT...|   0|2012-05-10|    2|ypjtMQLKdAwKGRS-K...|    1|Food is very blan...|     4|u0LXt3Uea_GidxRW1...|
|zxJlg4XCHNoFy78WZ...|   0|2011-09-28|    0|y21Fom8a_SdAyC6I0...|    2|a few years ago, ...|     0|u0LXt3Uea_GidxRW1...|
|tOhRQqiupLyJdBJVQ...|   1|2012-09-23|    1|_ku1sDwkmQo2wIgWA...|    5|OMG - Definitely ...|     1|u0LXt3Uea_GidxRW1...|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+
only showing top 20 rows

review_filter_3star: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [business_id: string, cool: bigint ... 7 more fields]

scala>
```

//combining 1 & 2 stars as BAD & 4 & 5 stars as GOOD(giving 1&2 star rating 0 and 4&5 Good)

var review_category = review_filter_3star.withColumn("Rating_labels",when(col("Stars") === "4" || col("stars") === "5", "Good").otherwise(0))

review_category.show()

```
File  Edit  View  Search  Terminal  Help
scala> :paste
// Entering paste mode (ctrl-D to finish)

//combining 1 & 2 stars as BAD & 4 & 5 stars as GOOD(giving 1&2 star rating 0 and 4&5 Good)
var review_category = review_filter_3star.withColumn("Rating_labels",when(col("Stars") === "4" || col("stars") === "5", "Good").otherwise(0))
review_category.show()

// Exiting paste mode, now interpreting.

+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+-------------+
|         business_id|cool|      date|funny|           review_id|stars|                text|useful|             user_id|Rating_labels|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+-------------+
|0W4lkclzZThpx3V65...|   0|2016-05-28|    0|v0i_UHJMo_hPBq9bx...|    5|Love the staff, l...|     0|bv2nCi5Qv5vroFiqK...|         Good|
|AEx2SYEUJmTxVVB18...|   0|2016-05-28|    0|vkVSCC7xljjrAI4UG...|    5|Super simple plac...|     0|bv2nCi5Qv5vroFiqK...|         Good|
|VR6GpWIda3SfvPC-l...|   0|2016-05-28|    0|n6QzIUObkYshz4dz2...|    5|Small unassuming ...|     0|bv2nCi5Qv5vroFiqK...|         Good|
|CKC0-MOWMqoeWf6s-...|   0|2016-05-28|    0|MV3CcKScW05u5LVfF...|    5|Lester's is locat...|     0|bv2nCi5Qv5vroFiqK...|         Good|
|ACFtxLv8pGrrxMm6E...|   0|2016-05-28|    0|IXvOzsEMYtiJI0CAR...|    4|Love coming here....|     0|bv2nCi5Qv5vroFiqK...|         Good|
|s2I_Ni76bjJNK9yG6...|   0|2016-05-28|    0|L_9BTb55X0GDtThi6...|    4|Had their chocola...|     0|bv2nCi5Qv5vroFiqK...|         Good|
|8QWPlVQ6D-OExqXoa...|   0|2014-09-24|    0|HRPm3vEZ_F-33TYVT...|    5|Cycle Pub Las Veg...|     1|_4iMDXbXZ1p1ONG29...|         Good|
|9_CGhHMz8698M9-Pk...|   2|2012-05-11|    0|ymAUG8DZfQcFTBSOi...|    4|Who would have gu...|     0|u0LXt3Uea_GidxRW1...|         Good|
|gkCorLgPyQLsptTHa...|   0|2015-10-27|    0|8UIishPUD92hXtScS...|    4|Always drove past...|     1|u0LXt3Uea_GidxRW1...|         Good|
|fDF_o2JPU8BR1Gya-...|   0|2016-04-06|    0|WF_QTN3p-thD74hqp...|    5|Love this place!
...|     3|u0LXt3Uea_GidxRW1...|         Good|
|z8oIoCT1cXz7gZP5G...|   0|2013-05-01|    0|PIsUSmvaUWB00qv5K...|    4|This is currently...|     1|u0LXt3Uea_GidxRW1...|         Good|
|13nKUHH-uEUXVZylg...|   1|2011-02-16|    0|x5oV6wm9_Pb1QQ6jk...|    1|I thought Tidy's ...|     9|u0LXt3Uea_GidxRW1...|            0|
|Aov96CM4FZAXeZvKt...|   1|2010-07-16|    0|23eqwlZzCWZkADWfd...|    5|This place is awe...|     2|u0LXt3Uea_GidxRW1...|         Good|
|0W4lkclzZThpx3V65...|   0|2011-09-28|    0|FunI9om-aK5oMIIJm...|    4|a must stop when ...|     0|u0LXt3Uea_GidxRW1...|         Good|
|fdnNZMk1NP7ZhL-YM...|   0|2012-10-23|    0|FKu4iU62EmWT6GZXP...|    1|I too have been t...|     0|u0LXt3Uea_GidxRW1...|            0|
|zgQHtqX0gqMw1nlBZ...|   1|2012-10-30|    2|WYDFJOBOl7cycd7gN...|    1|really excited to...|     9|u0LXt3Uea_GidxRW1...|            0|
|RWGI8u00x5GghYCEz...|   1|2011-10-14|    0|Kki2nwtP8U2qmWwRv...|    4|This place remind...|     0|u0LXt3Uea_GidxRW1...|         Good|
|hjk3ox7w1akbEuOgT...|   0|2012-05-10|    2|ypjtMQLKdAwKGRS-K...|    1|Food is very blan...|     4|u0LXt3Uea_GidxRW1...|            0|
|zxJlg4XCHNoFy78WZ...|   0|2011-09-28|    0|y21Fom8a_SdAyC6I0...|    2|a few years ago, ...|     0|u0LXt3Uea_GidxRW1...|            0|
|tOhRQqiupLyJdBJVQ...|   1|2012-09-23|    1|_ku1sDwkmQo2wIgWA...|    5|OMG - Definitely ...|     1|u0LXt3Uea_GidxRW1...|         Good|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+-------------+
only showing top 20 rows

review_category: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 8 more fields]

scala>
```

//Giving 1 rating to Good

val review_rating = review_category.withColumn("number_Rating", when((col("Rating_labels") === "Good"),1).otherwise(0))

review_rating.show()

//creating list of words based on text field

val review_tokenizer = new Tokenizer().setInputCol("text").setOutputCol("tokenized_review")

val review_tokenized_regexp = new RegexTokenizer().setInputCol("text").setOutputCol("tokenized_review").setPattern("\\W") // alternatively .setPattern("\\w+").setGaps(false)

val review_token_counts = udf { (tokenized_review: Seq[String]) => tokenized_review.length }

val review_tokenized = review_tokenizer.transform(review_category)

review_tokenized.select("text", "tokenized_review").withColumn("tokens", review_token_counts(col("tokenized_review"))).show(false)

val review_regexp_tokenized = review_tokenized_regexp.transform(review_category)

review_regexp_tokenized.select("text", "tokenized_review").withColumn("tokens", review_token_counts(col("tokenized_review"))).show(false)

//removing all stop words from list

val stopword_remover = new StopWordsRemover()

.setInputCol("tokenized_review")

.setOutputCol("review_without_stopword")

val stopword_removed_review = stopword_remover.transform(review_regexp_tokenized)

stopword_removed_review.show()

```
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+-------------+--------------------+----------------------+
|         business_id|cool|      date|funny|           review_id|stars|                text|useful|             user_id|Rating_labels|    tokenized_review|review_without_stopword|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+-------------+--------------------+----------------------+
|0W4lkclzZThpx3V65...|   0|2016-05-28|    0|v0i_UHJMo_hPBq9bx...|    5|Love the staff, l...|     0|bv2nCi5Qv5vroFiqK...|         Good|[love, the, staff...|   [love, staff, lov...|
|AEx2SYEUJmTxVVB18...|   0|2016-05-28|    0|vkvSCC7xljjrAI4UG...|    5|Super simple plac...|     0|bv2nCi5Qv5vroFiqK...|         Good|[super, simple, p...|   [super, simple, p...|
|VR6GpWIda3SfvPC-l...|   0|2016-05-28|    0|n6QzIUObkYshz4dz2...|    5|Small unassuming ...|     0|bv2nCi5Qv5vroFiqK...|         Good|[small, unassumin...|   [small, unassumin...|
|CKC0-MOWMqoeWf6s-...|   0|2016-05-28|    0|MV3CcKScW05u5LVfF...|    5|Lester's is locat...|     0|bv2nCi5Qv5vroFiqK...|         Good|[lester, s, is, l...|   [lester, located,...|
|ACFtxLv8pGrrxMNvGE...|   0|2016-05-28|    0|IXvOzsEMYtiJI0CAR...|    4|Love coming here....|     0|bv2nCi5Qv5vroFiqK...|         Good|[love, coming, he...|   [love, coming, ye...|
|sZI_NL76bjJNK9yG6...|   0|2016-05-28|    0|L_9BTb5SX0GDtThl6...|    4|Had their chocola...|     0|bv2nCi5Qv5vroFiqK...|         Good|[had, their, choc...|   [chocolate, almon...|
|8QWPlVQ6D-OExqXoa...|   0|2014-09-24|    0|HRPm3vEZ_F-33TYVT...|    5|Cycle Pub Las Veg...|     1|_4iMDXbXZ1p1ONGZ9...|         Good|[cycle, pub, las,...|   [cycle, pub, las,...|
|9_CGhHMz8698M9-Pk...|   2|2012-05-11|    0|ymAUGBDZfQcFTBSOl...|    4|Who would have gu...|     0|u0LXt3Uea_GldxRW1...|         Good|[who, would, have...|   [guess, able, get...|
|gkCorLgPyQLsptTHa...|   0|2015-10-27|    0|8UIishPUD92hXtScS...|    4|Always drove past...|     1|u0LXt3Uea_GldxRW1...|         Good|[always, drove, p...|   [always, drove, p...|
|fDF_o2JPU8BR1Gya-...|   0|2016-04-06|    0|WF_QTN3p-thD74hqp...|    5|Love this place!
...|     3|u0LXt3Uea_GldxRW1...|         Good|[love, this, plac...|  [love, place, peg...|
|z8oIoCT1cXz7gZP5G...|   0|2013-05-01|    0|PISUSmvaUWB00qv5K...|    4|This is currently...|     1|u0LXt3Uea_GldxRW1...|         Good|[this, is, curren...|   [currently, paren...|
|13nKUHH-uEUXVZylg...|   1|2011-02-16|    0|x5oV6wm9_Pb1QQ6jk...|    1|I thought Tidy's ...|     9|u0LXt3Uea_GldxRW1...|            0|[i, thought, tidy...|   [thought, tidy, f...|
|Aov96CM4FZAXeZvKt...|   1|2010-07-16|    0|23eqwlZzCWZkADWfd...|    5|This place is awe...|     2|u0LXt3Uea_GldxRW1...|         Good|[this, place, is,...|   [place, awesome, ...|
|0W4lkclzZThpx3V65...|   0|2011-09-28|    0|Fun19om-aK5oMIIJm...|    4|a must stop when ...|     0|u0LXt3Uea_GldxRW1...|         Good|[a, must, stop, w...|   [must, stop, mont...|
|fdnNZMk1NP7ZhL-YM...|   0|2012-10-23|    0|FKu4lU62EnWT6GZXP...|    1|I too have been t...|     0|u0LXt3Uea_GldxRW1...|            0|[i, too, have, be...|   [trying, book, ap...|
|zgQHtqX0gqMw1nlBZ...|   1|2012-10-30|    2|WYDFJOBOl7cycd7gN...|    1|really excited to...|     9|u0LXt3Uea_GldxRW1...|            0|[really, excited,...|   [really, excited,...|
|RWGI8u00x5GghYCEz...|   1|2011-10-14|    0|Kki2nwtP8U2qmWwRv...|    4|This place remind...|     0|u0LXt3Uea_GldxRW1...|         Good|[this, place, rem...|   [place, reminds, ...|
|hjk3ox7w1akbEu0gT...|   0|2012-05-10|    2|ypjtMQLKdAwKGR5-K...|    1|Food is very blan...|     4|u0LXt3Uea_GldxRW1...|            0|[food, is, very, ...|   [food, bland, aut...|
|zxJlg4XCHNoFy78WZ...|   0|2011-09-28|    0|y21Fom8a_SdAyC6I0...|    2|a few years ago, ...|     0|u0LXt3Uea_GldxRW1...|            0|[a, few, years, a...|   [years, ago, used...|
|tOhRQqiupLyJdBJVQ...|   1|2012-09-23|    1|_ku1sDwkmQo2wIgWA...|    5|OMG - Definitely ...|     1|u0LXt3Uea_GldxRW1...|         Good|[omg, definitely,...|   [omg, definitely,...|
+--------------------+----+----------+-----+--------------------+-----+--------------------+------+--------------------+-------------+--------------------+----------------------+
only showing top 20 rows

review_tokenizer: org.apache.spark.ml.feature.Tokenizer = tok_fe740aa199d0
review_tokenized_regexp: org.apache.spark.ml.feature.RegexTokenizer = regexTok_46f52620631c
review_token_counts: org.apache.spark.sql.expressions.UserDefinedFunction = UserDefinedFunction(<function1>,IntegerType,Some(List(ArrayType(StringType,true))))
review_tokenized: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 9 more fields]
review_regexp_tokenized: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 9 more fields]
stopword_remover: org.apache.spark.ml.feature.StopWordsRemover = stopWords_d215352f3823
stopword_removed_review: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 10 more fields]

scala>
```

//making pair of words from the list

val bigram = new NGram().setN(2).setInputCol("review_without_stopword").setOutputCol("bigram_review")

val review_bigram = bigram.transform(stopword_removed_review)

review_bigram.select("bigram_review").show()

val review_hashingTF = new HashingTF().setInputCol("review_without_stopword").setOutputCol("RawReviewFeature").setNumFeatures(25)

val featurized_review = review_hashingTF.transform(review_bigram)

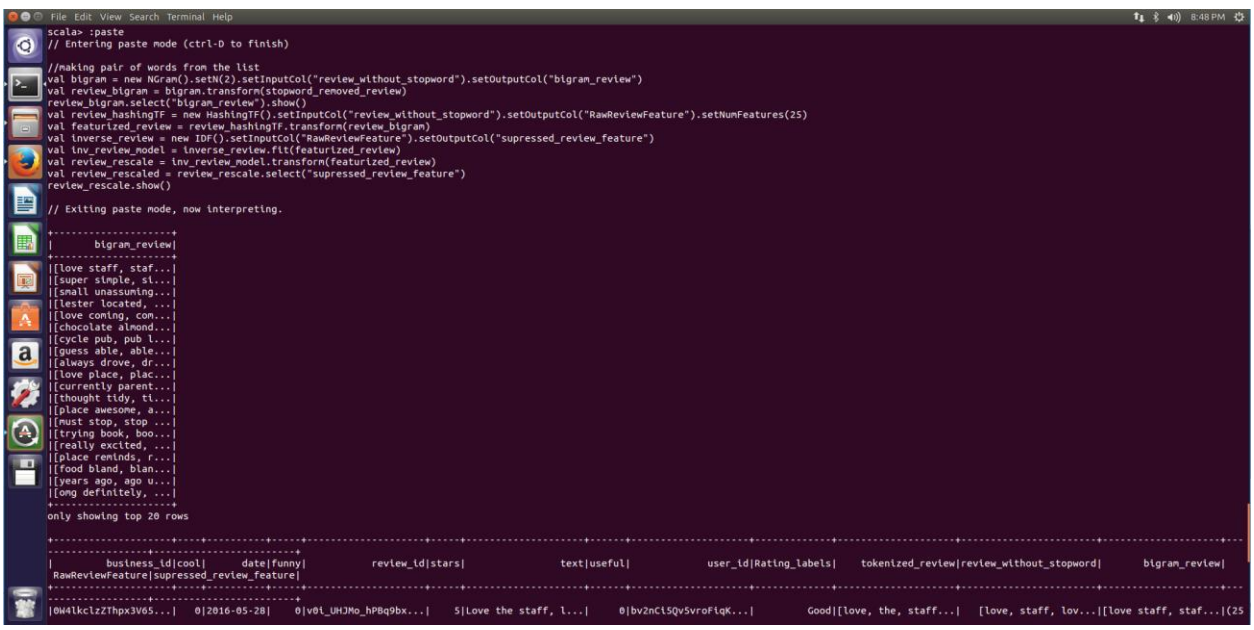
val inverse_review = new IDF().setInputCol("RawReviewFeature").setOutputCol("supressed_review_feature")

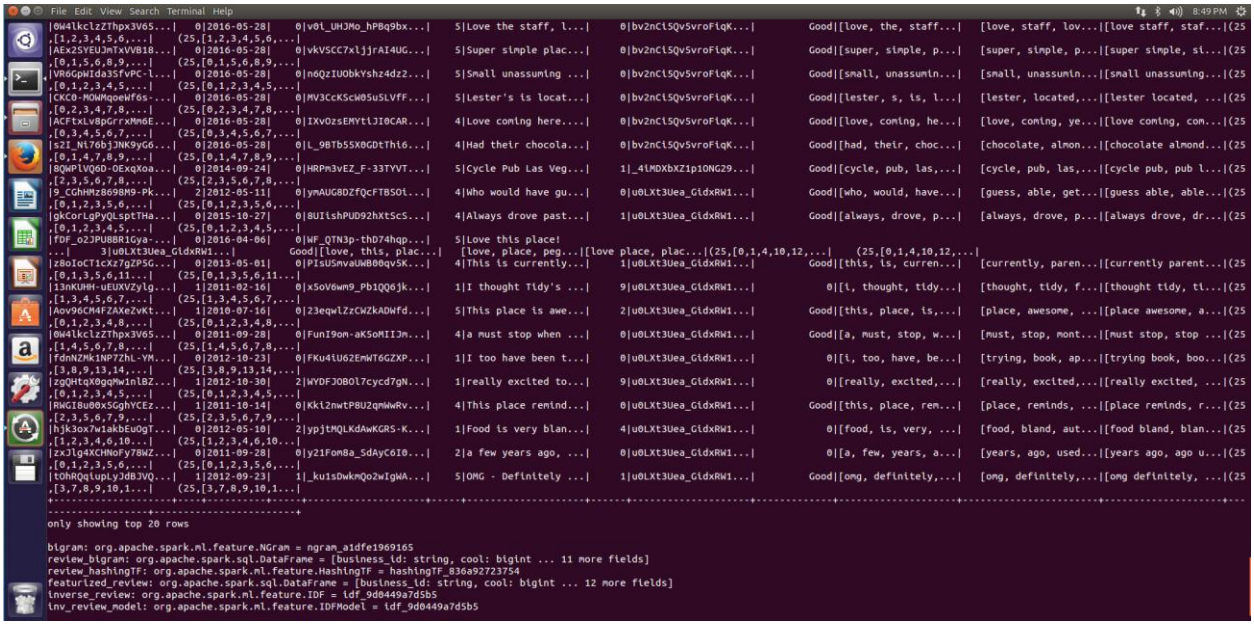
val inv_review_model = inverse_review.fit(featurized_review)

val review_rescale = inv_review_model.transform(featurized_review)

val review_rescaled = review_rescale.select("supressed_review_feature")

review_rescale.show()





```
only showing top 20 rows

bigram: org.apache.spark.ml.feature.NGram = ngram_a1dfe1969165
review_bigram: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 11 more fields]
review_hashingTF: org.apache.spark.ml.feature.HashingTF = hashingTF_836a92723754
featurized_review: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 12 more fields]
inverse_review: org.apache.spark.ml.feature.IDF = idf_9d0449a7d5b5
inv_review_model: org.apache.spark.ml.feature.IDFModel = idf_9d0449a7d5b5
```

val review_lsvc = new LinearSVC().setFeaturesCol("supressed_review_feature").setLabelCol("number_Rating")

val review_pipeline = new Pipeline().setStages(Array(review_tokenized_regexp,stopword_remover,bigram,review_hashingTF,inverse_review,review_lsvc))

val review_evaluator = new MulticlassClassificationEvaluator()

.setLabelCol("number_Rating")

.setPredictionCol("prediction")

.setMetricName("accuracy")

val review_cross_validator = new CrossValidator()

.setEstimator(review_pipeline)

.setEvaluator(review_evaluator)

.setEstimatorParamMaps(new ParamGridBuilder().build)

.setNumFolds(4)

val Array(review_trainingData, review_testData) = review_rating.randomSplit(Array(0.85, 0.15),3296)

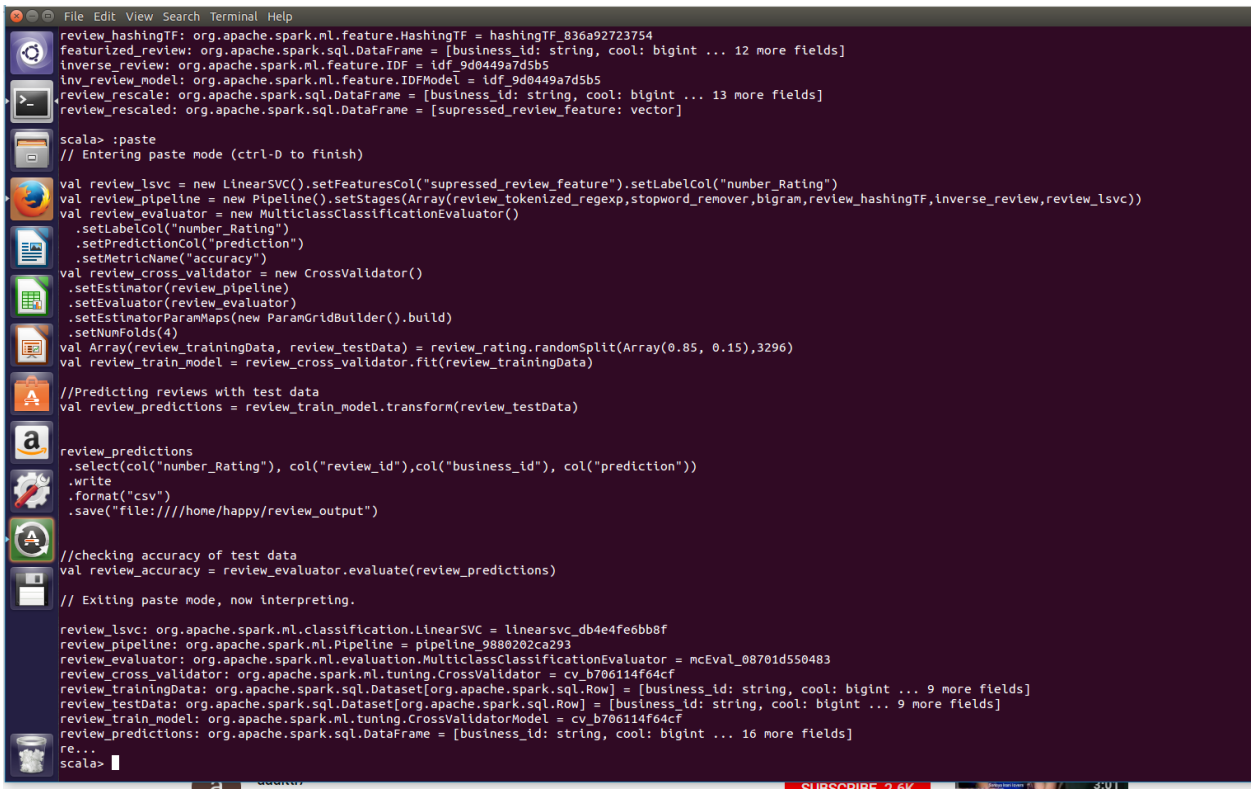
val review_train_model = review_cross_validator.fit(review_trainingData)

//Predicting reviews with test data

val review_predictions = review_train_model.transform(review_testData)

```
review_predictions
 .select(col("number_Rating"), col("review_id"),col("business_id"), col("prediction"))
 .write
 .format("csv")
 .save("file:////home/happy/review_output")
```

```
//checking accuracy of test data
val review_accuracy = review_evaluator.evaluate(review_predictions)
```



```
File  Edit  View  Search  Terminal  Help
review_hashingTF: org.apache.spark.ml.feature.HashingTF = hashingTF_836a92723754
featurized_review: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 12 more fields]
inverse_review: org.apache.spark.ml.feature.IDF = idf_9d0449a7d5b5
inv_review_model: org.apache.spark.ml.feature.IDFModel = idf_9d0449a7d5b5
review_rescale: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 13 more fields]
review_rescaled: org.apache.spark.sql.DataFrame = [supressed_review_feature: vector]

scala> :paste
// Entering paste mode (ctrl-D to finish)

val review_lsvc = new LinearSVC().setFeaturesCol("supressed_review_feature").setLabelCol("number_Rating")
val review_pipeline = new Pipeline().setStages(Array(review_tokenized_regexp,stopword_remover,bigram,review_hashingTF,inverse_review,review_lsvc))
val review_evaluator = new MulticlassClassificationEvaluator()
  .setLabelCol("number_Rating")
  .setPredictionCol("prediction")
  .setMetricName("accuracy")
val review_cross_validator = new CrossValidator()
 .setEstimator(review_pipeline)
 .setEvaluator(review_evaluator)
 .setEstimatorParamMaps(new ParamGridBuilder().build)
 .setNumFolds(4)
val Array(review_trainingData, review_testData) = review_rating.randomSplit(Array(0.85, 0.15),3296)
val review_train_model = review_cross_validator.fit(review_trainingData)

//Predicting reviews with test data
val review_predictions = review_train_model.transform(review_testData)


review_predictions
 .select(col("number_Rating"), col("review_id"),col("business_id"), col("prediction"))
 .write
 .format("csv")
 .save("file:////home/happy/review_output")


//checking accuracy of test data
val review_accuracy = review_evaluator.evaluate(review_predictions)

// Exiting paste mode, now interpreting.

review_lsvc: org.apache.spark.ml.classification.LinearSVC = linearsvc_db4e4fe6bb8f
review_pipeline: org.apache.spark.ml.Pipeline = pipeline_9880202ca293
review_evaluator: org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator = mcEval_08701d550483
review_cross_validator: org.apache.spark.ml.tuning.CrossValidator = cv_b706114f64cf
review_trainingData: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [business_id: string, cool: bigint ... 9 more fields]
review_testData: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [business_id: string, cool: bigint ... 9 more fields]
review_train_model: org.apache.spark.ml.tuning.CrossValidatorModel = cv_b706114f64cf
review_predictions: org.apache.spark.sql.DataFrame = [business_id: string, cool: bigint ... 16 more fields]
re...
scala>
```