StackOverflow Tag Predictor - Model Optimization

By Aziz Presswala

In [22]:

```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import sqlite3
import csv
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from wordcloud import WordCloud
import re
import os
from sqlalchemy import create_engine # database connection
import datetime as dt
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.multiclass import OneVsRestClassifier
from sklearn.linear_model import SGDClassifier
from sklearn import metrics
from sklearn.metrics import f1_score,precision_score,recall_score
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from datetime import datetime
from sklearn.model_selection import GridSearchCV
```

In [2]:

```
def create_connection(db_file):
    """ create a database connection to the SQLite database
        specified by db_file
    :param db_file: database file
    :return: Connection object or None
    """
    try:
        conn = sqlite3.connect(db_file)
        return conn
    except Error as e:
        print(e)

    return None
```

In [3]:

```
def tags_to_choose(n):
    t = multilabel_y.sum(axis=0).tolist()[0]
    sorted_tags_i = sorted(range(len(t)), key=lambda i: t[i], reverse=True)
    multilabel_yn=multilabel_y[:,sorted_tags_i[:n]]
    return multilabel_yn

def questions_explained_fn(n):
    multilabel_yn = tags_to_choose(n)
    x= multilabel_yn.sum(axis=1)
    return (np.count_nonzero(x==0))
```

In [4]:

```
write_db = 'Titlemoreweight.db'
if os.path.isfile(write_db):
    conn_r = create_connection(write_db)
    if conn_r is not None:
        preprocessed_data = pd.read_sql_query("""SELECT question, Tags FROM QuestionsProces
conn_r.commit()
conn_r.close()
```

In [5]:

```
preprocessed_data.head()
```

Out[5]:

	question	tags
0	dynam datagrid bind silverlight dynam datagrid...	c# silverlight data-binding
1	dynam datagrid bind silverlight dynam datagrid...	c# silverlight data-binding columns
2	java.lang.noclassdeffounderror javax servlet j...	jsp jstl
3	java.sql.sqlexcept microsoft odbc driver manag...	java jdbc
4	better way updat feed fb php sdk better way up...	facebook api facebook-php-sdk

In [6]:

```
print("number of data points in sample :", preprocessed_data.shape[0])
print("number of dimensions :", preprocessed_data.shape[1])
```

```
number of data points in sample : 500000
number of dimensions : 2
```

Converting string Tags to multilable output variables

In [11]:

```
vectorizer = CountVectorizer(tokenizer = lambda x: x.split(), binary='true')
multilabel_y = vectorizer.fit_transform(preprocessed_data['tags'])
```

Selecting 500 Tags

In [12]:

```
questions_explained = []
total_tags=multilabel_y.shape[1]
total_qs=preprocessed_data.shape[0]
for i in range(500, total_tags, 100):
    questions_explained.append(np.round(((total_qs-questions_explained_fn(i))/total_qs)*100
```

In [13]:

```
# we will be taking 500 tags
multilabel_yx = tags_to_choose(500)
print("number of questions that are not covered :", questions_explained_fn(500),"out of ",
```

```
number of questions that are not covered : 45221 out of  500000
```

In [14]:

```
x_train=preprocessed_data.head(400000)
x_test=preprocessed_data.tail(preprocessed_data.shape[0] - 400000)

y_train = multilabel_yx[0:400000,:]
y_test = multilabel_yx[400000:preprocessed_data.shape[0],:]
```

In [15]:

```
print("Number of data points in train data :", y_train.shape)
print("Number of data points in test data :", y_test.shape)
```

```
Number of data points in train data : (400000, 500)
Number of data points in test data : (100000, 500)
```

Part 1 - Logistic Regression with BoW

Featurizing data with BoW vectorizer - ngram_range=(1,4)

In [16]:

```
start = datetime.now()
vectorizer = CountVectorizer(min_df=0.00009, max_features=200000,
                             tokenizer = lambda x: x.split(), ngram_range=(1,4))
x_train_multilabel = vectorizer.fit_transform(x_train['question'])
x_test_multilabel = vectorizer.transform(x_test['question'])
print("Time taken to run this cell :", datetime.now() - start)
```

```
Time taken to run this cell : 3:41:09.156093
```

In [17]:

```
print("Dimensions of train data X:",x_train_multilabel.shape, "Y :",y_train.shape)
print("Dimensions of test data X:",x_test_multilabel.shape,"Y:",y_test.shape)
```

```
Dimensions of train data X: (400000, 95585) Y : (400000, 500)
Dimensions of test data X: (100000, 95585) Y: (100000, 500)
```

Applying Logistic Regression with OneVsRestClassifier

In [20]:

```
start = datetime.now()
classifier = OneVsRestClassifier(SGDClassifier(loss='log', alpha=0.00001, penalty='l1'))
classifier.fit(x_train_multilabel, y_train)
predictions = classifier.predict (x_test_multilabel)


print("Accuracy :",metrics.accuracy_score(y_test, predictions))
print("Hamming loss ",metrics.hamming_loss(y_test,predictions))


precision = precision_score(y_test, predictions, average='micro')
recall = recall_score(y_test, predictions, average='micro')
f1 = f1_score(y_test, predictions, average='micro')

print("Micro-average quality numbers")
print("Precision: {:.4f}, Recall: {:.4f}, F1-measure: {:.4f}".format(precision, recall, f1))

precision = precision_score(y_test, predictions, average='macro')
recall = recall_score(y_test, predictions, average='macro')
f1 = f1_score(y_test, predictions, average='macro')

print("Macro-average quality numbers")
print("Precision: {:.4f}, Recall: {:.4f}, F1-measure: {:.4f}".format(precision, recall, f1))

print (metrics.classification_report(y_test, predictions))
print("Time taken to run this cell :", datetime.now() - start)
```

```
Accuracy : 0.10774
Hamming loss  0.0060036
Micro-average quality numbers
Precision: 0.2847, Recall: 0.4805, F1-measure: 0.3575
Macro-average quality numbers
Precision: 0.2056, Recall: 0.4105, F1-measure: 0.2661
              precision    recall  f1-score   support

           0       0.73      0.80      0.76      5519
           1       0.44      0.44      0.44      8190
           2       0.52      0.52      0.52      6529
           3       0.49      0.60      0.54      3231
           4       0.54      0.53      0.53      6430
           5       0.42      0.53      0.47      2879
           6       0.58      0.62      0.60      5086
           7       0.58      0.68      0.63      4533
           8       0.23      0.22      0.22      3000
           9       0.54      0.63      0.59      2765
          10       0.32      0.34      0.33      3051
          11       0.46      0.52      0.49      3009
          12       0.38      0.44      0.41      2630
          13       0.35      0.45      0.39      1426
          14       0.59      0.67      0.63      2548
          15       0.36      0.40      0.38      2371
          16       0.26      0.34      0.30       873
          17       0.54      0.72      0.62      2151
          18       0.30      0.38      0.34      2204
          19       0.26      0.50      0.34       831
          20       0.51      0.58      0.54      1860
          21       0.18      0.29      0.22      2023
          22       0.29      0.39      0.33      1513
```

23	0.47	0.69	0.56	1207
24	0.22	0.43	0.29	506
25	0.22	0.44	0.29	425
26	0.29	0.52	0.37	793
27	0.36	0.48	0.41	1291
28	0.40	0.51	0.45	1208
29	0.12	0.26	0.17	406
30	0.16	0.35	0.22	504
31	0.12	0.23	0.16	732
32	0.20	0.43	0.28	441
33	0.34	0.46	0.39	1645
34	0.25	0.36	0.29	1058
35	0.47	0.62	0.53	946
36	0.23	0.39	0.29	644
37	0.24	0.76	0.37	136
38	0.31	0.54	0.39	570
39	0.29	0.45	0.35	766
40	0.32	0.44	0.37	1132
41	0.09	0.35	0.15	174
42	0.29	0.63	0.39	210
43	0.32	0.55	0.40	433
44	0.31	0.54	0.39	626
45	0.28	0.46	0.35	852
46	0.28	0.57	0.38	534
47	0.15	0.30	0.20	350
48	0.33	0.62	0.43	496
49	0.52	0.74	0.61	785
50	0.12	0.27	0.17	475
51	0.10	0.27	0.14	305
52	0.08	0.16	0.10	251
53	0.35	0.47	0.40	914
54	0.19	0.31	0.24	728
55	0.06	0.13	0.08	258
56	0.22	0.41	0.29	821
57	0.14	0.26	0.18	541
58	0.31	0.43	0.36	748
59	0.60	0.77	0.67	724
60	0.19	0.34	0.24	660
61	0.15	0.34	0.21	235
62	0.54	0.79	0.64	718
63	0.44	0.76	0.56	468
64	0.17	0.50	0.26	191
65	0.12	0.28	0.17	429
66	0.12	0.24	0.16	415
67	0.27	0.62	0.38	274
68	0.39	0.64	0.49	510
69	0.30	0.53	0.38	466
70	0.10	0.20	0.13	305
71	0.11	0.34	0.17	247
72	0.36	0.55	0.43	401
73	0.23	0.83	0.37	86
74	0.20	0.49	0.28	120
75	0.36	0.72	0.48	129
76	0.08	0.16	0.11	473
77	0.08	0.32	0.13	143
78	0.34	0.58	0.43	347
79	0.23	0.38	0.29	479
80	0.20	0.44	0.27	279
81	0.19	0.34	0.24	461
82	0.07	0.15	0.10	298
83	0.31	0.59	0.41	396

84	0.17	0.43	0.24	184
85	0.22	0.38	0.28	573
86	0.11	0.22	0.15	325
87	0.18	0.39	0.24	273
88	0.07	0.24	0.11	135
89	0.14	0.28	0.18	232
90	0.29	0.50	0.37	409
91	0.23	0.49	0.31	420
92	0.36	0.63	0.46	408
93	0.21	0.59	0.31	241
94	0.07	0.18	0.10	211
95	0.13	0.27	0.17	277
96	0.11	0.17	0.13	410
97	0.41	0.58	0.48	501
98	0.22	0.71	0.34	136
99	0.23	0.42	0.30	239
100	0.13	0.27	0.17	324
101	0.50	0.77	0.60	277
102	0.69	0.81	0.75	613
103	0.11	0.32	0.16	157
104	0.09	0.19	0.12	295
105	0.31	0.53	0.39	334
106	0.22	0.43	0.29	335
107	0.34	0.60	0.43	389
108	0.20	0.41	0.27	251
109	0.27	0.46	0.34	317
110	0.06	0.19	0.09	187
111	0.08	0.24	0.12	140
112	0.17	0.55	0.26	154
113	0.24	0.35	0.28	332
114	0.17	0.41	0.24	323
115	0.19	0.39	0.25	344
116	0.36	0.58	0.44	370
117	0.17	0.34	0.22	313
118	0.60	0.81	0.69	874
119	0.17	0.35	0.23	293
120	0.07	0.23	0.11	200
121	0.37	0.60	0.46	463
122	0.10	0.29	0.15	119
123	0.03	0.06	0.04	256
124	0.46	0.73	0.56	195
125	0.09	0.30	0.13	138
126	0.34	0.63	0.44	376
127	0.03	0.12	0.05	122
128	0.07	0.17	0.10	252
129	0.22	0.44	0.30	144
130	0.09	0.34	0.14	150
131	0.06	0.16	0.09	210
132	0.22	0.38	0.28	361
133	0.58	0.67	0.62	453
134	0.39	0.80	0.52	124
135	0.04	0.14	0.06	91
136	0.12	0.45	0.19	128
137	0.18	0.48	0.26	218
138	0.13	0.31	0.18	243
139	0.11	0.35	0.17	149
140	0.38	0.58	0.46	318
141	0.07	0.25	0.11	159
142	0.31	0.54	0.39	274
143	0.57	0.83	0.67	362
144	0.09	0.33	0.14	118

145	0.18	0.54	0.27	164
146	0.24	0.42	0.31	461
147	0.24	0.53	0.33	159
148	0.10	0.31	0.15	166
149	0.40	0.65	0.49	346
150	0.15	0.25	0.18	350
151	0.16	0.73	0.27	55
152	0.41	0.61	0.49	387
153	0.20	0.31	0.25	150
154	0.10	0.21	0.14	281
155	0.10	0.28	0.14	202
156	0.31	0.68	0.42	130
157	0.13	0.21	0.16	245
158	0.37	0.75	0.49	177
159	0.17	0.54	0.25	130
160	0.16	0.31	0.21	336
161	0.42	0.74	0.53	220
162	0.09	0.29	0.14	229
163	0.41	0.54	0.46	316
164	0.24	0.52	0.32	283
165	0.16	0.43	0.23	197
166	0.19	0.54	0.28	101
167	0.13	0.28	0.18	231
168	0.17	0.46	0.25	370
169	0.20	0.34	0.25	258
170	0.05	0.25	0.09	101
171	0.08	0.34	0.13	89
172	0.18	0.40	0.25	193
173	0.24	0.44	0.31	309
174	0.08	0.24	0.12	172
175	0.30	0.79	0.44	95
176	0.57	0.71	0.63	346
177	0.42	0.66	0.51	322
178	0.30	0.60	0.40	232
179	0.06	0.15	0.09	125
180	0.19	0.46	0.27	145
181	0.04	0.26	0.06	77
182	0.08	0.24	0.12	182
183	0.22	0.47	0.30	257
184	0.07	0.20	0.10	216
185	0.14	0.29	0.19	242
186	0.11	0.25	0.15	165
187	0.35	0.66	0.46	263
188	0.10	0.27	0.14	174
189	0.38	0.51	0.44	136
190	0.44	0.72	0.55	202
191	0.06	0.22	0.10	134
192	0.21	0.49	0.30	230
193	0.09	0.27	0.13	90
194	0.25	0.57	0.34	185
195	0.05	0.17	0.07	156
196	0.04	0.14	0.07	160
197	0.14	0.32	0.19	266
198	0.15	0.25	0.19	284
199	0.07	0.17	0.10	145
200	0.48	0.82	0.61	212
201	0.19	0.38	0.26	317
202	0.41	0.69	0.52	427
203	0.12	0.30	0.17	232
204	0.19	0.37	0.25	217
205	0.39	0.58	0.46	527

206	0.05	0.19	0.08	124
207	0.22	0.45	0.29	103
208	0.37	0.59	0.46	287
209	0.07	0.20	0.10	193
210	0.21	0.50	0.30	220
211	0.13	0.32	0.18	140
212	0.08	0.20	0.12	161
213	0.24	0.57	0.34	72
214	0.37	0.59	0.46	396
215	0.16	0.53	0.25	134
216	0.17	0.30	0.22	400
217	0.10	0.41	0.17	75
218	0.57	0.84	0.68	219
219	0.20	0.42	0.27	210
220	0.43	0.77	0.55	298
221	0.53	0.74	0.62	266
222	0.41	0.58	0.48	290
223	0.02	0.07	0.03	128
224	0.20	0.50	0.28	159
225	0.15	0.45	0.22	164
226	0.18	0.44	0.25	144
227	0.35	0.55	0.42	276
228	0.04	0.09	0.05	235
229	0.07	0.13	0.09	216
230	0.08	0.24	0.12	228
231	0.21	0.61	0.31	64
232	0.06	0.19	0.09	103
233	0.26	0.49	0.34	216
234	0.15	0.32	0.20	116
235	0.12	0.48	0.20	77
236	0.40	0.76	0.52	67
237	0.13	0.24	0.17	218
238	0.07	0.28	0.11	139
239	0.04	0.12	0.06	94
240	0.12	0.43	0.19	77
241	0.08	0.20	0.12	167
242	0.23	0.55	0.33	86
243	0.06	0.34	0.10	58
244	0.32	0.44	0.37	269
245	0.09	0.22	0.12	112
246	0.66	0.83	0.73	255
247	0.06	0.28	0.10	58
248	0.02	0.12	0.04	81
249	0.05	0.16	0.08	131
250	0.09	0.32	0.14	93
251	0.11	0.50	0.18	154
252	0.03	0.08	0.05	129
253	0.15	0.45	0.23	83
254	0.10	0.24	0.14	191
255	0.09	0.23	0.13	219
256	0.04	0.20	0.07	130
257	0.13	0.43	0.20	93
258	0.33	0.53	0.40	217
259	0.10	0.33	0.15	141
260	0.19	0.40	0.26	143
261	0.12	0.25	0.16	219
262	0.15	0.49	0.23	107
263	0.22	0.41	0.29	236
264	0.12	0.31	0.18	119
265	0.08	0.31	0.13	72
266	0.05	0.16	0.07	70

267	0.11	0.27	0.15	107
268	0.27	0.57	0.37	169
269	0.13	0.43	0.20	129
270	0.35	0.54	0.43	159
271	0.26	0.62	0.37	190
272	0.18	0.41	0.25	248
273	0.61	0.81	0.69	264
274	0.44	0.77	0.56	105
275	0.07	0.24	0.11	104
276	0.02	0.07	0.04	115
277	0.34	0.66	0.45	170
278	0.31	0.56	0.40	145
279	0.46	0.72	0.56	230
280	0.16	0.50	0.24	80
281	0.39	0.65	0.49	217
282	0.38	0.54	0.44	175
283	0.14	0.29	0.19	269
284	0.16	0.47	0.24	74
285	0.28	0.62	0.39	206
286	0.44	0.75	0.56	227
287	0.23	0.56	0.32	130
288	0.05	0.13	0.07	129
289	0.05	0.17	0.07	80
290	0.07	0.23	0.10	99
291	0.30	0.50	0.37	208
292	0.05	0.25	0.08	67
293	0.27	0.56	0.36	109
294	0.15	0.46	0.22	140
295	0.13	0.32	0.19	241
296	0.10	0.35	0.16	72
297	0.07	0.21	0.10	107
298	0.25	0.56	0.35	61
299	0.32	0.62	0.43	77
300	0.07	0.23	0.10	111
301	0.01	0.02	0.01	126
302	0.04	0.11	0.06	73
303	0.25	0.52	0.34	176
304	0.67	0.83	0.74	230
305	0.47	0.74	0.57	156
306	0.20	0.51	0.29	146
307	0.09	0.21	0.12	98
308	0.01	0.04	0.01	78
309	0.07	0.17	0.10	94
310	0.22	0.52	0.31	162
311	0.30	0.66	0.41	116
312	0.11	0.42	0.18	57
313	0.05	0.25	0.09	65
314	0.16	0.40	0.23	138
315	0.26	0.42	0.32	195
316	0.14	0.45	0.21	69
317	0.07	0.25	0.11	134
318	0.23	0.49	0.32	148
319	0.38	0.58	0.46	161
320	0.10	0.37	0.16	104
321	0.33	0.63	0.44	156
322	0.19	0.47	0.27	134
323	0.32	0.50	0.39	232
324	0.08	0.23	0.12	92
325	0.20	0.43	0.27	197
326	0.05	0.17	0.08	126
327	0.02	0.07	0.03	115

328	0.49	0.74	0.59	198
329	0.16	0.38	0.22	125
330	0.12	0.32	0.17	81
331	0.11	0.20	0.14	94
332	0.05	0.14	0.08	56
333	0.07	0.17	0.10	260
334	0.07	0.27	0.11	60
335	0.10	0.26	0.15	110
336	0.22	0.61	0.33	71
337	0.03	0.14	0.04	66
338	0.18	0.49	0.26	150
339	0.01	0.04	0.01	54
340	0.40	0.65	0.50	195
341	0.31	0.58	0.41	79
342	0.12	0.50	0.19	38
343	0.13	0.49	0.21	43
344	0.15	0.26	0.19	68
345	0.24	0.55	0.33	73
346	0.05	0.17	0.08	116
347	0.19	0.60	0.29	111
348	0.03	0.17	0.05	63
349	0.37	0.73	0.49	104
350	0.15	0.57	0.23	44
351	0.17	0.42	0.24	40
352	0.34	0.60	0.44	136
353	0.11	0.41	0.17	54
354	0.07	0.25	0.11	134
355	0.21	0.44	0.28	120
356	0.28	0.54	0.37	228
357	0.31	0.46	0.37	269
358	0.17	0.53	0.26	80
359	0.33	0.62	0.43	140
360	0.09	0.25	0.13	125
361	0.53	0.69	0.60	169
362	0.03	0.12	0.04	56
363	0.44	0.79	0.56	154
364	0.07	0.21	0.10	58
365	0.09	0.30	0.14	71
366	0.41	0.76	0.53	54
367	0.07	0.16	0.09	116
368	0.02	0.07	0.04	54
369	0.03	0.21	0.05	71
370	0.03	0.16	0.05	61
371	0.05	0.20	0.08	71
372	0.17	0.52	0.26	52
373	0.35	0.62	0.45	150
374	0.15	0.48	0.22	93
375	0.04	0.13	0.06	67
376	0.03	0.09	0.04	76
377	0.16	0.39	0.23	106
378	0.02	0.09	0.04	86
379	0.02	0.29	0.03	14
380	0.24	0.61	0.34	122
381	0.04	0.18	0.07	104
382	0.05	0.23	0.09	66
383	0.16	0.40	0.23	110
384	0.05	0.11	0.06	155
385	0.09	0.44	0.15	50
386	0.07	0.20	0.10	64
387	0.08	0.17	0.11	93
388	0.18	0.43	0.25	102

389	0.02	0.07	0.04	108
390	0.61	0.73	0.67	178
391	0.15	0.37	0.22	115
392	0.15	0.55	0.24	42
393	0.03	0.05	0.04	134
394	0.05	0.19	0.08	112
395	0.14	0.30	0.19	176
396	0.11	0.24	0.15	125
397	0.40	0.65	0.49	224
398	0.29	0.76	0.42	63
399	0.02	0.10	0.03	59
400	0.13	0.51	0.21	63
401	0.11	0.38	0.17	98
402	0.15	0.33	0.20	162
403	0.10	0.33	0.15	83
404	0.25	0.79	0.38	19
405	0.07	0.20	0.10	92
406	0.11	0.61	0.18	41
407	0.20	0.40	0.26	43
408	0.26	0.54	0.36	160
409	0.07	0.26	0.12	50
410	0.01	0.16	0.02	19
411	0.14	0.26	0.18	175
412	0.07	0.19	0.10	72
413	0.06	0.16	0.08	95
414	0.06	0.18	0.09	97
415	0.05	0.23	0.09	48
416	0.15	0.45	0.23	83
417	0.02	0.10	0.04	40
418	0.08	0.18	0.11	91
419	0.23	0.57	0.33	90
420	0.09	0.30	0.13	37
421	0.09	0.24	0.13	66
422	0.11	0.44	0.17	73
423	0.11	0.39	0.17	56
424	0.38	0.91	0.54	33
425	0.03	0.09	0.04	76
426	0.03	0.09	0.04	81
427	0.58	0.78	0.66	150
428	0.33	0.79	0.47	29
429	0.88	0.85	0.86	389
430	0.31	0.48	0.37	167
431	0.06	0.21	0.09	123
432	0.10	0.44	0.17	39
433	0.11	0.33	0.17	82
434	0.42	0.73	0.53	66
435	0.18	0.54	0.27	93
436	0.16	0.49	0.25	87
437	0.06	0.15	0.08	86
438	0.38	0.63	0.47	104
439	0.09	0.28	0.13	100
440	0.08	0.18	0.11	141
441	0.20	0.49	0.28	110
442	0.11	0.33	0.16	123
443	0.12	0.32	0.18	71
444	0.11	0.19	0.14	109
445	0.13	0.48	0.20	48
446	0.16	0.49	0.24	76
447	0.05	0.32	0.09	38
448	0.26	0.69	0.38	81
449	0.25	0.36	0.30	132

450	0.14	0.37	0.20	81
451	0.15	0.43	0.23	76
452	0.08	0.20	0.11	44
453	0.02	0.09	0.04	44
454	0.27	0.63	0.38	70
455	0.14	0.36	0.20	155
456	0.10	0.44	0.17	43
457	0.15	0.43	0.22	72
458	0.04	0.18	0.06	62
459	0.07	0.33	0.12	69
460	0.03	0.08	0.05	119
461	0.20	0.32	0.25	79
462	0.09	0.23	0.13	47
463	0.11	0.38	0.17	104
464	0.20	0.42	0.28	106
465	0.09	0.36	0.15	64
466	0.26	0.46	0.33	173
467	0.27	0.57	0.37	107
468	0.18	0.34	0.23	126
469	0.05	0.10	0.06	114
470	0.59	0.84	0.69	140
471	0.21	0.48	0.29	79
472	0.22	0.49	0.31	143
473	0.35	0.39	0.37	158
474	0.08	0.14	0.10	138
475	0.04	0.20	0.07	59
476	0.22	0.45	0.30	88
477	0.43	0.74	0.55	176
478	0.23	0.83	0.36	24
479	0.07	0.17	0.10	92
480	0.33	0.60	0.43	100
481	0.20	0.41	0.27	103
482	0.09	0.34	0.14	74
483	0.42	0.70	0.52	105
484	0.06	0.16	0.08	83
485	0.03	0.13	0.05	82
486	0.08	0.25	0.12	71
487	0.17	0.29	0.21	120
488	0.09	0.19	0.13	105
489	0.21	0.37	0.27	87
490	0.41	0.88	0.55	32
491	0.02	0.06	0.03	69
492	0.01	0.04	0.02	49
493	0.02	0.06	0.03	117
494	0.12	0.33	0.18	61
495	0.79	0.76	0.77	344
496	0.14	0.29	0.19	52
497	0.21	0.40	0.28	137
498	0.10	0.20	0.14	98
499	0.08	0.32	0.13	79
micro avg	0.28	0.48	0.36	173812
macro avg	0.21	0.41	0.27	173812
weighted avg	0.35	0.48	0.39	173812
samples avg	0.36	0.45	0.36	173812

```
Time taken to run this cell : 0:39:01.831393
```

```
C:\Users\Aziz\Anaconda3\lib\site-packages\sklearn\metrics\classification.p
y:1143: UndefinedMetricWarning: Precision and F-score are ill-defined and
being set to 0.0 in samples with no predicted labels.
```

```
  'precision', 'predicted', average, warn_for)
C:\Users\Aziz\Anaconda3\lib\site-packages\sklearn\metrics\classification.p
y:1145: UndefinedMetricWarning: Recall and F-score are ill-defined and bei
ng set to 0.0 in samples with no true labels.
  'recall', 'true', average, warn_for)
```

Part 2 - Hyperparamter tuning for Logistic Regression

In [25]:

```
start = datetime.now()

#setting alpha values that need to be tried on the classifier
params = {'estimator__alpha':[10**-3, 10**-2, 10**-1, 10**0, 10**1, 10**2, 10**3]}

classifier = OneVsRestClassifier(SGDClassifier(loss='log', penalty='l1'))
gscv = GridSearchCV(classifier, param_grid = params, scoring='f1_micro', return_train_score
gscv.fit(x_train_multilabel, y_train)
print(gscv.best_score_)
print(gscv.best_params_)
print("Time taken to run this cell :", datetime.now() - start)
```

```
0.4485524752691557
{'estimator__alpha': 0.001}
Time taken to run this cell : 4:37:02.728530
```

In [28]:

```
start = datetime.now()

#setting alpha values that need to be tried on the classifier
params = {'estimator__alpha':[10**-5, 10**-4]}

classifier = OneVsRestClassifier(SGDClassifier(loss='log', penalty='l1'))
gscv = GridSearchCV(classifier, param_grid = params, scoring='f1_micro', return_train_score
gscv.fit(x_train_multilabel, y_train)
print(gscv.best_score_)
print(gscv.best_params_)
print("Time taken to run this cell :", datetime.now() - start)
```

```
0.4426540793261227
{'estimator__alpha': 0.0001}
Time taken to run this cell : 1:42:20.344928
```

From the above 2 cells, we observe that the best value of alpha is *0.001*

In [27]:

```
# Training the model with the optimal value of alpha
start = datetime.now()

classifier = OneVsRestClassifier(SGDClassifier(loss='log', alpha=0.001, penalty='l1'))
classifier.fit(x_train_multilabel, y_train)
predictions = classifier.predict (x_test_multilabel)


print("Accuracy :",metrics.accuracy_score(y_test, predictions))
print("Hamming loss ",metrics.hamming_loss(y_test,predictions))


precision = precision_score(y_test, predictions, average='micro')
recall = recall_score(y_test, predictions, average='micro')
f1 = f1_score(y_test, predictions, average='micro')

print("Micro-average quality numbers")
print("Precision: {:.4f}, Recall: {:.4f}, F1-measure: {:.4f}".format(precision, recall, f1))

precision = precision_score(y_test, predictions, average='macro')
recall = recall_score(y_test, predictions, average='macro')
f1 = f1_score(y_test, predictions, average='macro')

print("Macro-average quality numbers")
print("Precision: {:.4f}, Recall: {:.4f}, F1-measure: {:.4f}".format(precision, recall, f1))

print (metrics.classification_report(y_test, predictions))
print("Time taken to run this cell :", datetime.now() - start)
```

```
Accuracy : 0.18564
Hamming loss  0.00322702
Micro-average quality numbers
Precision: 0.5618, Recall: 0.3258, F1-measure: 0.4125
Macro-average quality numbers
Precision: 0.4054, Recall: 0.2385, F1-measure: 0.2821
              precision    recall  f1-score   support

           0       0.80      0.68      0.74      5519
           1       0.53      0.23      0.32      8190
           2       0.64      0.42      0.51      6529
           3       0.60      0.48      0.53      3231
           4       0.81      0.36      0.50      6430
           5       0.65      0.42      0.51      2879
           6       0.76      0.56      0.64      5086
           7       0.84      0.57      0.68      4533
           8       0.53      0.14      0.22      3000
           9       0.76      0.48      0.59      2765
          10       0.58      0.14      0.22      3051
          11       0.57      0.38      0.45      3009
          12       0.63      0.24      0.35      2630
          13       0.53      0.15      0.23      1426
          14       0.79      0.64      0.70      2548
          15       0.55      0.14      0.22      2371
          16       0.57      0.27      0.37       873
          17       0.75      0.66      0.70      2151
          18       0.51      0.24      0.32      2204
          19       0.58      0.42      0.49       831
          20       0.63      0.52      0.57      1860
```

21	0.25	0.10	0.14	2023
22	0.36	0.26	0.30	1513
23	0.76	0.63	0.69	1207
24	0.38	0.36	0.37	506
25	0.46	0.34	0.39	425
26	0.49	0.36	0.42	793
27	0.48	0.36	0.41	1291
28	0.70	0.35	0.46	1208
29	0.30	0.14	0.19	406
30	0.42	0.27	0.33	504
31	0.25	0.18	0.21	732
32	0.53	0.26	0.35	441
33	0.39	0.10	0.15	1645
34	0.73	0.23	0.35	1058
35	0.68	0.67	0.68	946
36	0.66	0.18	0.28	644
37	0.39	0.80	0.53	136
38	0.56	0.39	0.46	570
39	0.57	0.36	0.44	766
40	0.52	0.26	0.34	1132
41	0.15	0.25	0.19	174
42	0.44	0.47	0.46	210
43	0.63	0.52	0.57	433
44	0.58	0.43	0.50	626
45	0.48	0.29	0.36	852
46	0.38	0.61	0.47	534
47	0.20	0.19	0.20	350
48	0.58	0.52	0.55	496
49	0.71	0.65	0.68	785
50	0.18	0.11	0.14	475
51	0.26	0.13	0.17	305
52	0.28	0.04	0.07	251
53	0.37	0.56	0.44	914
54	0.25	0.26	0.25	728
55	0.00	0.00	0.00	258
56	0.23	0.24	0.23	821
57	0.42	0.10	0.16	541
58	0.79	0.24	0.37	748
59	0.86	0.75	0.80	724
60	0.25	0.06	0.10	660
61	0.67	0.27	0.39	235
62	0.90	0.69	0.78	718
63	0.86	0.49	0.62	468
64	0.52	0.43	0.47	191
65	0.19	0.12	0.15	429
66	0.25	0.05	0.09	415
67	0.61	0.63	0.62	274
68	0.84	0.49	0.62	510
69	0.42	0.55	0.48	466
70	0.24	0.12	0.16	305
71	0.33	0.22	0.27	247
72	0.47	0.52	0.49	401
73	0.58	0.81	0.68	86
74	0.67	0.31	0.42	120
75	0.83	0.69	0.75	129
76	0.08	0.03	0.04	473
77	0.27	0.28	0.27	143
78	0.70	0.59	0.64	347
79	0.60	0.34	0.44	479
80	0.33	0.35	0.34	279
81	0.58	0.15	0.24	461

82	0.09	0.06	0.07	298
83	0.61	0.38	0.47	396
84	0.40	0.32	0.36	184
85	0.47	0.19	0.27	573
86	0.25	0.06	0.09	325
87	0.50	0.21	0.30	273
88	0.30	0.21	0.24	135
89	0.19	0.20	0.19	232
90	0.31	0.41	0.35	409
91	0.45	0.51	0.48	420
92	0.70	0.45	0.54	408
93	0.52	0.48	0.50	241
94	0.21	0.09	0.12	211
95	0.17	0.13	0.15	277
96	0.27	0.02	0.03	410
97	0.90	0.25	0.39	501
98	0.69	0.58	0.63	136
99	0.42	0.37	0.39	239
100	0.47	0.08	0.14	324
101	0.95	0.45	0.61	277
102	0.92	0.65	0.76	613
103	0.51	0.20	0.28	157
104	0.16	0.12	0.14	295
105	0.81	0.26	0.39	334
106	0.33	0.01	0.01	335
107	0.77	0.49	0.60	389
108	0.41	0.35	0.38	251
109	0.49	0.39	0.44	317
110	0.45	0.05	0.09	187
111	0.28	0.21	0.24	140
112	0.12	0.02	0.03	154
113	0.65	0.21	0.32	332
114	0.38	0.23	0.29	323
115	0.40	0.12	0.19	344
116	0.70	0.44	0.54	370
117	0.45	0.16	0.24	313
118	0.78	0.49	0.61	874
119	0.38	0.18	0.24	293
120	0.00	0.00	0.00	200
121	0.74	0.42	0.53	463
122	0.24	0.22	0.23	119
123	0.00	0.00	0.00	256
124	0.90	0.71	0.79	195
125	0.37	0.25	0.30	138
126	0.67	0.45	0.54	376
127	0.20	0.06	0.09	122
128	0.16	0.06	0.09	252
129	0.09	0.01	0.02	144
130	0.12	0.02	0.03	150
131	0.00	0.00	0.00	210
132	0.18	0.07	0.11	361
133	0.83	0.62	0.71	453
134	0.54	0.81	0.65	124
135	0.01	0.01	0.01	91
136	0.36	0.21	0.26	128
137	0.37	0.37	0.37	218
138	0.04	0.00	0.01	243
139	0.34	0.23	0.28	149
140	0.71	0.31	0.43	318
141	0.08	0.16	0.10	159
142	0.63	0.34	0.44	274

143	0.85	0.56	0.67	362
144	0.23	0.42	0.29	118
145	0.53	0.41	0.46	164
146	0.57	0.27	0.37	461
147	0.64	0.47	0.54	159
148	0.34	0.14	0.20	166
149	0.92	0.58	0.71	346
150	0.35	0.02	0.03	350
151	0.70	0.55	0.61	55
152	0.77	0.37	0.50	387
153	0.20	0.01	0.01	150
154	0.31	0.15	0.20	281
155	0.18	0.13	0.15	202
156	0.66	0.58	0.61	130
157	0.30	0.09	0.14	245
158	0.81	0.47	0.60	177
159	0.46	0.25	0.33	130
160	0.42	0.20	0.27	336
161	0.76	0.70	0.73	220
162	0.14	0.06	0.08	229
163	0.84	0.37	0.52	316
164	0.67	0.15	0.25	283
165	0.51	0.25	0.34	197
166	0.15	0.11	0.13	101
167	0.37	0.23	0.28	231
168	0.15	0.15	0.15	370
169	0.33	0.28	0.30	258
170	0.08	0.09	0.09	101
171	0.44	0.18	0.26	89
172	0.35	0.28	0.31	193
173	0.41	0.28	0.33	309
174	0.40	0.11	0.17	172
175	0.60	0.84	0.70	95
176	0.65	0.55	0.59	346
177	0.98	0.27	0.42	322
178	0.53	0.37	0.44	232
179	0.67	0.08	0.14	125
180	0.42	0.23	0.30	145
181	0.19	0.19	0.19	77
182	0.13	0.07	0.09	182
183	0.55	0.32	0.40	257
184	0.23	0.02	0.04	216
185	0.22	0.09	0.12	242
186	0.27	0.13	0.17	165
187	0.65	0.61	0.63	263
188	0.27	0.10	0.15	174
189	0.57	0.15	0.24	136
190	0.93	0.56	0.70	202
191	0.31	0.11	0.16	134
192	0.68	0.55	0.61	230
193	0.25	0.14	0.18	90
194	0.47	0.66	0.55	185
195	0.08	0.05	0.06	156
196	0.00	0.00	0.00	160
197	0.00	0.00	0.00	266
198	0.26	0.03	0.05	284
199	0.22	0.03	0.06	145
200	0.84	0.72	0.78	212
201	0.19	0.03	0.05	317
202	0.59	0.52	0.55	427
203	0.18	0.11	0.14	232

204	0.14	0.21	0.17	217
205	0.44	0.33	0.38	527
206	0.05	0.10	0.07	124
207	0.50	0.01	0.02	103
208	0.83	0.42	0.56	287
209	0.12	0.10	0.11	193
210	0.47	0.21	0.29	220
211	0.73	0.06	0.11	140
212	0.07	0.07	0.07	161
213	0.35	0.17	0.23	72
214	0.61	0.45	0.51	396
215	0.78	0.42	0.54	134
216	0.00	0.00	0.00	400
217	0.50	0.31	0.38	75
218	0.63	0.69	0.66	219
219	0.79	0.35	0.49	210
220	0.88	0.34	0.49	298
221	0.93	0.59	0.72	266
222	0.82	0.25	0.38	290
223	0.11	0.05	0.07	128
224	0.79	0.35	0.49	159
225	0.34	0.19	0.24	164
226	0.44	0.33	0.38	144
227	0.42	0.39	0.40	276
228	0.06	0.01	0.01	235
229	0.03	0.01	0.01	216
230	0.30	0.23	0.26	228
231	0.65	0.47	0.55	64
232	0.08	0.11	0.09	103
233	0.62	0.34	0.44	216
234	0.00	0.00	0.00	116
235	0.60	0.48	0.53	77
236	0.88	0.75	0.81	67
237	0.00	0.00	0.00	218
238	0.05	0.01	0.02	139
239	0.22	0.02	0.04	94
240	0.12	0.18	0.15	77
241	0.46	0.04	0.07	167
242	0.78	0.33	0.46	86
243	0.40	0.10	0.16	58
244	0.18	0.06	0.09	269
245	0.11	0.04	0.06	112
246	0.90	0.76	0.82	255
247	0.43	0.21	0.28	58
248	0.23	0.04	0.06	81
249	0.00	0.00	0.00	131
250	0.32	0.16	0.21	93
251	0.49	0.23	0.31	154
252	0.07	0.05	0.06	129
253	0.42	0.27	0.32	83
254	0.20	0.06	0.10	191
255	0.11	0.01	0.02	219
256	0.15	0.03	0.05	130
257	0.37	0.37	0.37	93
258	0.66	0.34	0.45	217
259	0.18	0.07	0.10	141
260	0.88	0.15	0.25	143
261	0.47	0.08	0.14	219
262	0.37	0.39	0.38	107
263	0.33	0.29	0.31	236
264	0.15	0.10	0.12	119

265	0.27	0.24	0.25	72
266	0.13	0.17	0.15	70
267	0.26	0.07	0.12	107
268	0.52	0.44	0.48	169
269	0.19	0.14	0.16	129
270	0.64	0.45	0.53	159
271	0.42	0.12	0.18	190
272	0.47	0.06	0.10	248
273	0.88	0.67	0.76	264
274	0.86	0.56	0.68	105
275	0.00	0.00	0.00	104
276	0.06	0.01	0.02	115
277	0.84	0.55	0.66	170
278	0.40	0.12	0.18	145
279	0.91	0.37	0.53	230
280	0.55	0.36	0.44	80
281	0.68	0.53	0.59	217
282	0.67	0.70	0.69	175
283	0.56	0.05	0.10	269
284	0.60	0.28	0.39	74
285	0.72	0.53	0.61	206
286	0.89	0.56	0.69	227
287	0.81	0.22	0.35	130
288	0.27	0.09	0.13	129
289	0.06	0.01	0.02	80
290	0.11	0.08	0.09	99
291	0.74	0.25	0.37	208
292	0.31	0.13	0.19	67
293	0.59	0.21	0.31	109
294	0.25	0.21	0.23	140
295	0.17	0.20	0.18	241
296	0.16	0.11	0.13	72
297	0.23	0.10	0.14	107
298	0.85	0.18	0.30	61
299	0.75	0.31	0.44	77
300	0.16	0.05	0.08	111
301	0.00	0.00	0.00	126
302	0.00	0.00	0.00	73
303	0.46	0.41	0.44	176
304	0.96	0.54	0.69	230
305	0.97	0.54	0.70	156
306	0.30	0.42	0.35	146
307	0.14	0.05	0.07	98
308	0.17	0.01	0.02	78
309	0.27	0.07	0.12	94
310	0.35	0.31	0.33	162
311	0.74	0.46	0.56	116
312	0.50	0.32	0.39	57
313	0.01	0.02	0.01	65
314	0.48	0.31	0.38	138
315	0.43	0.23	0.30	195
316	0.43	0.42	0.43	69
317	0.00	0.00	0.00	134
318	0.32	0.22	0.26	148
319	0.82	0.29	0.43	161
320	0.17	0.20	0.19	104
321	0.51	0.71	0.59	156
322	0.51	0.29	0.37	134
323	0.53	0.31	0.39	232
324	0.27	0.15	0.19	92
325	0.29	0.06	0.10	197

326	0.04	0.05	0.04	126
327	0.00	0.00	0.00	115
328	0.92	0.68	0.78	198
329	0.51	0.26	0.34	125
330	0.62	0.06	0.11	81
331	0.07	0.05	0.06	94
332	0.07	0.07	0.07	56
333	0.08	0.02	0.04	260
334	0.00	0.00	0.00	60
335	0.22	0.16	0.19	110
336	0.47	0.48	0.47	71
337	0.14	0.06	0.09	66
338	0.42	0.25	0.32	150
339	0.00	0.00	0.00	54
340	0.82	0.46	0.59	195
341	0.00	0.00	0.00	79
342	0.34	0.34	0.34	38
343	0.36	0.23	0.28	43
344	0.00	0.00	0.00	68
345	0.52	0.41	0.46	73
346	0.07	0.07	0.07	116
347	0.90	0.32	0.48	111
348	0.13	0.05	0.07	63
349	0.73	0.63	0.68	104
350	0.70	0.32	0.44	44
351	0.00	0.00	0.00	40
352	0.97	0.26	0.41	136
353	0.36	0.30	0.32	54
354	0.00	0.00	0.00	134
355	0.54	0.23	0.32	120
356	0.29	0.06	0.10	228
357	0.56	0.09	0.15	269
358	0.67	0.35	0.46	80
359	0.79	0.36	0.49	140
360	0.23	0.05	0.08	125
361	0.94	0.26	0.41	169
362	0.11	0.05	0.07	56
363	0.77	0.67	0.72	154
364	0.00	0.00	0.00	58
365	0.13	0.15	0.14	71
366	0.97	0.52	0.67	54
367	0.24	0.05	0.09	116
368	0.00	0.00	0.00	54
369	0.03	0.01	0.02	71
370	0.10	0.02	0.03	61
371	0.00	0.00	0.00	71
372	0.45	0.56	0.50	52
373	0.76	0.17	0.28	150
374	0.38	0.16	0.23	93
375	0.20	0.01	0.03	67
376	0.00	0.00	0.00	76
377	1.00	0.08	0.14	106
378	0.50	0.01	0.02	86
379	0.12	0.07	0.09	14
380	1.00	0.26	0.42	122
381	0.13	0.04	0.06	104
382	0.08	0.15	0.10	66
383	0.28	0.46	0.35	110
384	0.00	0.00	0.00	155
385	0.07	0.02	0.03	50
386	0.21	0.20	0.21	64

387	0.00	0.00	0.00	93
388	0.56	0.20	0.29	102
389	0.02	0.03	0.02	108
390	0.94	0.56	0.70	178
391	0.60	0.16	0.25	115
392	0.94	0.36	0.52	42
393	0.00	0.00	0.00	134
394	0.00	0.00	0.00	112
395	0.28	0.03	0.05	176
396	0.00	0.00	0.00	125
397	0.62	0.17	0.26	224
398	0.80	0.56	0.65	63
399	0.00	0.00	0.00	59
400	0.37	0.29	0.32	63
401	0.11	0.02	0.03	98
402	0.41	0.06	0.10	162
403	0.28	0.35	0.31	83
404	0.71	0.63	0.67	19
405	0.13	0.09	0.11	92
406	0.32	0.15	0.20	41
407	0.60	0.14	0.23	43
408	0.00	0.00	0.00	160
409	0.25	0.22	0.23	50
410	0.00	0.00	0.00	19
411	0.31	0.17	0.22	175
412	0.09	0.04	0.06	72
413	0.06	0.02	0.03	95
414	0.09	0.06	0.07	97
415	0.25	0.12	0.17	48
416	0.35	0.22	0.27	83
417	0.00	0.00	0.00	40
418	0.11	0.18	0.14	91
419	0.45	0.28	0.34	90
420	0.27	0.22	0.24	37
421	0.05	0.02	0.02	66
422	0.56	0.27	0.37	73
423	0.35	0.20	0.25	56
424	0.93	0.79	0.85	33
425	0.07	0.01	0.02	76
426	0.14	0.01	0.02	81
427	0.99	0.53	0.69	150
428	0.95	0.62	0.75	29
429	0.00	0.00	0.00	389
430	0.58	0.19	0.29	167
431	0.00	0.00	0.00	123
432	0.41	0.28	0.33	39
433	0.32	0.23	0.27	82
434	1.00	0.58	0.73	66
435	0.51	0.38	0.43	93
436	0.70	0.22	0.33	87
437	0.36	0.06	0.10	86
438	0.62	0.36	0.45	104
439	0.00	0.00	0.00	100
440	0.33	0.01	0.01	141
441	0.27	0.23	0.25	110
442	0.15	0.09	0.11	123
443	0.00	0.00	0.00	71
444	0.33	0.05	0.08	109
445	0.20	0.12	0.15	48
446	0.42	0.22	0.29	76
447	0.11	0.08	0.09	38

448	0.67	0.44	0.53	81
449	0.24	0.08	0.12	132
450	0.34	0.25	0.29	81
451	0.00	0.00	0.00	76
452	0.00	0.00	0.00	44
453	0.00	0.00	0.00	44
454	0.81	0.31	0.45	70
455	0.00	0.00	0.00	155
456	0.24	0.14	0.18	43
457	0.35	0.12	0.18	72
458	0.18	0.13	0.15	62
459	0.00	0.00	0.00	69
460	0.21	0.03	0.06	119
461	0.67	0.15	0.25	79
462	0.13	0.04	0.06	47
463	0.50	0.10	0.16	104
464	0.58	0.29	0.39	106
465	0.07	0.08	0.07	64
466	0.50	0.20	0.28	173
467	0.73	0.33	0.45	107
468	0.00	0.00	0.00	126
469	0.00	0.00	0.00	114
470	0.76	0.73	0.74	140
471	0.00	0.00	0.00	79
472	0.33	0.27	0.29	143
473	0.24	0.03	0.06	158
474	0.00	0.00	0.00	138
475	0.09	0.05	0.07	59
476	0.55	0.35	0.43	88
477	0.79	0.44	0.57	176
478	0.92	0.50	0.65	24
479	0.00	0.00	0.00	92
480	0.80	0.36	0.50	100
481	0.37	0.32	0.34	103
482	0.29	0.22	0.25	74
483	0.77	0.46	0.57	105
484	0.03	0.01	0.02	83
485	0.17	0.01	0.02	82
486	0.29	0.07	0.11	71
487	0.34	0.19	0.25	120
488	0.00	0.00	0.00	105
489	0.62	0.21	0.31	87
490	1.00	0.62	0.77	32
491	0.00	0.00	0.00	69
492	0.00	0.00	0.00	49
493	0.00	0.00	0.00	117
494	0.80	0.07	0.12	61
495	0.00	0.00	0.00	344
496	0.00	0.00	0.00	52
497	0.18	0.37	0.24	137
498	0.00	0.00	0.00	98
499	0.64	0.09	0.16	79
micro avg	0.56	0.33	0.41	173812
macro avg	0.41	0.24	0.28	173812
weighted avg	0.54	0.33	0.39	173812
samples avg	0.38	0.31	0.32	173812

```
Time taken to run this cell : 0:22:22.256690
```

Part 3 - Linear SVM with OneVsRestClassifier

In [24]:

```
start = datetime.now()
classifier = OneVsRestClassifier(SGDClassifier(loss='hinge', alpha=0.00001, penalty='l1'))
classifier.fit(x_train_multilabel, y_train)
predictions = classifier.predict (x_test_multilabel)


print("Accuracy :",metrics.accuracy_score(y_test, predictions))
print("Hamming loss ",metrics.hamming_loss(y_test,predictions))


precision = precision_score(y_test, predictions, average='micro')
recall = recall_score(y_test, predictions, average='micro')
f1 = f1_score(y_test, predictions, average='micro')

print("Micro-average quality numbers")
print("Precision: {:.4f}, Recall: {:.4f}, F1-measure: {:.4f}".format(precision, recall, f1))

precision = precision_score(y_test, predictions, average='macro')
recall = recall_score(y_test, predictions, average='macro')
f1 = f1_score(y_test, predictions, average='macro')

print("Macro-average quality numbers")
print("Precision: {:.4f}, Recall: {:.4f}, F1-measure: {:.4f}".format(precision, recall, f1))

print (metrics.classification_report(y_test, predictions))
print("Time taken to run this cell :", datetime.now() - start)
```

```
Accuracy : 0.10822
Hamming loss  0.00591506
Micro-average quality numbers
Precision: 0.2886, Recall: 0.4787, F1-measure: 0.3601
Macro-average quality numbers
Precision: 0.2087, Recall: 0.4072, F1-measure: 0.2683
              precision    recall  f1-score   support

           0       0.71      0.81      0.75      5519
           1       0.45      0.47      0.46      8190
           2       0.49      0.53      0.51      6529
           3       0.54      0.57      0.55      3231
           4       0.53      0.54      0.53      6430
           5       0.40      0.50      0.44      2879
           6       0.59      0.62      0.60      5086
           7       0.60      0.67      0.64      4533
           8       0.22      0.24      0.23      3000
           9       0.57      0.64      0.60      2765
          10       0.32      0.34      0.33      3051
          11       0.45      0.53      0.48      3009
          12       0.37      0.39      0.38      2630
          13       0.35      0.43      0.39      1426
          14       0.60      0.68      0.63      2548
          15       0.35      0.38      0.36      2371
          16       0.26      0.39      0.31       873
          17       0.55      0.70      0.62      2151
          18       0.30      0.36      0.33      2204
          19       0.27      0.49      0.35       831
          20       0.51      0.57      0.54      1860
          21       0.17      0.22      0.20      2023
          22       0.29      0.41      0.34      1513
```

23	0.41	0.68	0.51	1207
24	0.23	0.43	0.30	506
25	0.21	0.44	0.29	425
26	0.33	0.57	0.42	793
27	0.34	0.48	0.40	1291
28	0.46	0.49	0.47	1208
29	0.10	0.22	0.14	406
30	0.19	0.37	0.25	504
31	0.14	0.24	0.18	732
32	0.22	0.39	0.29	441
33	0.37	0.50	0.42	1645
34	0.25	0.34	0.29	1058
35	0.42	0.65	0.51	946
36	0.23	0.42	0.30	644
37	0.29	0.76	0.42	136
38	0.31	0.53	0.39	570
39	0.27	0.46	0.34	766
40	0.32	0.49	0.39	1132
41	0.10	0.32	0.15	174
42	0.32	0.62	0.42	210
43	0.35	0.50	0.41	433
44	0.34	0.56	0.42	626
45	0.28	0.47	0.35	852
46	0.33	0.53	0.41	534
47	0.13	0.36	0.19	350
48	0.33	0.61	0.43	496
49	0.55	0.70	0.61	785
50	0.10	0.20	0.14	475
51	0.08	0.22	0.12	305
52	0.08	0.19	0.11	251
53	0.33	0.52	0.40	914
54	0.20	0.32	0.25	728
55	0.07	0.14	0.09	258
56	0.20	0.37	0.26	821
57	0.12	0.26	0.16	541
58	0.28	0.44	0.34	748
59	0.60	0.75	0.67	724
60	0.20	0.30	0.24	660
61	0.18	0.33	0.24	235
62	0.55	0.80	0.65	718
63	0.48	0.73	0.58	468
64	0.17	0.46	0.24	191
65	0.12	0.21	0.15	429
66	0.12	0.21	0.15	415
67	0.27	0.58	0.37	274
68	0.42	0.60	0.50	510
69	0.34	0.53	0.41	466
70	0.10	0.23	0.14	305
71	0.10	0.24	0.14	247
72	0.37	0.59	0.46	401
73	0.28	0.81	0.41	86
74	0.20	0.57	0.30	120
75	0.36	0.70	0.48	129
76	0.08	0.11	0.10	473
77	0.09	0.42	0.15	143
78	0.40	0.61	0.48	347
79	0.21	0.34	0.26	479
80	0.19	0.49	0.27	279
81	0.19	0.35	0.25	461
82	0.08	0.16	0.11	298
83	0.34	0.63	0.44	396

84	0.20	0.46	0.28	184
85	0.27	0.34	0.30	573
86	0.10	0.22	0.14	325
87	0.20	0.41	0.26	273
88	0.08	0.25	0.12	135
89	0.13	0.26	0.17	232
90	0.31	0.49	0.38	409
91	0.22	0.44	0.29	420
92	0.43	0.64	0.51	408
93	0.21	0.56	0.30	241
94	0.08	0.18	0.11	211
95	0.11	0.22	0.15	277
96	0.12	0.19	0.15	410
97	0.49	0.59	0.54	501
98	0.22	0.65	0.33	136
99	0.21	0.42	0.28	239
100	0.13	0.26	0.18	324
101	0.46	0.76	0.57	277
102	0.67	0.78	0.72	613
103	0.09	0.33	0.14	157
104	0.10	0.22	0.14	295
105	0.31	0.53	0.39	334
106	0.22	0.42	0.29	335
107	0.32	0.60	0.42	389
108	0.20	0.45	0.28	251
109	0.27	0.43	0.33	317
110	0.08	0.20	0.11	187
111	0.09	0.29	0.14	140
112	0.20	0.56	0.30	154
113	0.21	0.37	0.27	332
114	0.22	0.35	0.27	323
115	0.16	0.41	0.23	344
116	0.37	0.58	0.45	370
117	0.18	0.39	0.25	313
118	0.60	0.76	0.67	874
119	0.15	0.33	0.20	293
120	0.03	0.07	0.04	200
121	0.40	0.61	0.48	463
122	0.10	0.26	0.15	119
123	0.03	0.04	0.03	256
124	0.47	0.74	0.57	195
125	0.10	0.30	0.15	138
126	0.35	0.64	0.46	376
127	0.03	0.12	0.04	122
128	0.10	0.19	0.13	252
129	0.22	0.37	0.28	144
130	0.10	0.27	0.14	150
131	0.06	0.16	0.09	210
132	0.22	0.42	0.29	361
133	0.52	0.68	0.59	453
134	0.47	0.83	0.60	124
135	0.04	0.18	0.06	91
136	0.09	0.45	0.15	128
137	0.18	0.50	0.26	218
138	0.13	0.33	0.19	243
139	0.10	0.28	0.15	149
140	0.36	0.55	0.44	318
141	0.09	0.23	0.13	159
142	0.34	0.57	0.43	274
143	0.64	0.85	0.73	362
144	0.08	0.33	0.13	118

145	0.18	0.49	0.27	164
146	0.23	0.43	0.30	461
147	0.22	0.55	0.32	159
148	0.10	0.29	0.15	166
149	0.42	0.65	0.51	346
150	0.16	0.27	0.20	350
151	0.24	0.73	0.36	55
152	0.36	0.59	0.45	387
153	0.18	0.25	0.21	150
154	0.13	0.23	0.17	281
155	0.11	0.30	0.17	202
156	0.33	0.69	0.45	130
157	0.14	0.20	0.17	245
158	0.36	0.70	0.48	177
159	0.15	0.47	0.23	130
160	0.19	0.31	0.24	336
161	0.37	0.70	0.48	220
162	0.10	0.25	0.15	229
163	0.36	0.56	0.44	316
164	0.26	0.54	0.35	283
165	0.18	0.42	0.25	197
166	0.23	0.57	0.33	101
167	0.15	0.24	0.18	231
168	0.22	0.43	0.29	370
169	0.22	0.46	0.30	258
170	0.05	0.23	0.09	101
171	0.08	0.35	0.14	89
172	0.16	0.41	0.23	193
173	0.25	0.46	0.33	309
174	0.08	0.24	0.12	172
175	0.25	0.83	0.38	95
176	0.57	0.71	0.63	346
177	0.44	0.66	0.53	322
178	0.29	0.53	0.38	232
179	0.06	0.16	0.09	125
180	0.20	0.45	0.27	145
181	0.03	0.22	0.06	77
182	0.07	0.24	0.11	182
183	0.24	0.48	0.32	257
184	0.07	0.16	0.10	216
185	0.15	0.28	0.19	242
186	0.12	0.30	0.17	165
187	0.38	0.60	0.47	263
188	0.09	0.25	0.14	174
189	0.28	0.48	0.35	136
190	0.41	0.69	0.51	202
191	0.10	0.29	0.15	134
192	0.23	0.52	0.32	230
193	0.07	0.27	0.11	90
194	0.26	0.62	0.37	185
195	0.04	0.17	0.06	156
196	0.05	0.18	0.08	160
197	0.15	0.30	0.20	266
198	0.14	0.25	0.18	284
199	0.07	0.19	0.10	145
200	0.53	0.78	0.63	212
201	0.19	0.40	0.26	317
202	0.46	0.63	0.54	427
203	0.11	0.18	0.14	232
204	0.19	0.46	0.27	217
205	0.37	0.58	0.45	527

206	0.03	0.10	0.04	124
207	0.19	0.42	0.27	103
208	0.37	0.59	0.46	287
209	0.08	0.17	0.11	193
210	0.20	0.47	0.28	220
211	0.12	0.33	0.17	140
212	0.05	0.14	0.08	161
213	0.22	0.53	0.31	72
214	0.43	0.59	0.50	396
215	0.20	0.48	0.28	134
216	0.21	0.29	0.24	400
217	0.13	0.36	0.19	75
218	0.59	0.79	0.67	219
219	0.27	0.51	0.36	210
220	0.57	0.72	0.64	298
221	0.53	0.74	0.62	266
222	0.39	0.58	0.47	290
223	0.03	0.12	0.05	128
224	0.21	0.44	0.28	159
225	0.18	0.48	0.26	164
226	0.21	0.47	0.29	144
227	0.33	0.59	0.42	276
228	0.05	0.14	0.07	235
229	0.05	0.13	0.07	216
230	0.11	0.32	0.16	228
231	0.20	0.64	0.30	64
232	0.06	0.23	0.10	103
233	0.24	0.47	0.32	216
234	0.15	0.34	0.21	116
235	0.14	0.40	0.20	77
236	0.39	0.78	0.52	67
237	0.11	0.26	0.16	218
238	0.10	0.27	0.15	139
239	0.03	0.07	0.05	94
240	0.13	0.44	0.20	77
241	0.07	0.18	0.10	167
242	0.24	0.53	0.33	86
243	0.06	0.31	0.09	58
244	0.31	0.46	0.37	269
245	0.06	0.18	0.09	112
246	0.66	0.86	0.75	255
247	0.07	0.29	0.11	58
248	0.03	0.16	0.05	81
249	0.04	0.13	0.06	131
250	0.10	0.39	0.16	93
251	0.16	0.45	0.24	154
252	0.05	0.12	0.07	129
253	0.12	0.45	0.19	83
254	0.09	0.20	0.12	191
255	0.07	0.20	0.11	219
256	0.06	0.22	0.09	130
257	0.14	0.44	0.21	93
258	0.35	0.60	0.44	217
259	0.10	0.27	0.14	141
260	0.16	0.36	0.22	143
261	0.14	0.28	0.19	219
262	0.14	0.45	0.21	107
263	0.24	0.39	0.30	236
264	0.10	0.39	0.16	119
265	0.12	0.43	0.19	72
266	0.06	0.19	0.09	70

267	0.12	0.35	0.17	107
268	0.22	0.55	0.31	169
269	0.12	0.26	0.16	129
270	0.31	0.65	0.42	159
271	0.23	0.57	0.33	190
272	0.19	0.39	0.26	248
273	0.57	0.80	0.66	264
274	0.43	0.79	0.56	105
275	0.06	0.23	0.10	104
276	0.03	0.12	0.05	115
277	0.42	0.63	0.50	170
278	0.30	0.58	0.40	145
279	0.47	0.77	0.59	230
280	0.13	0.41	0.19	80
281	0.45	0.64	0.53	217
282	0.37	0.62	0.46	175
283	0.16	0.34	0.22	269
284	0.13	0.46	0.20	74
285	0.36	0.58	0.44	206
286	0.48	0.74	0.58	227
287	0.20	0.55	0.29	130
288	0.06	0.16	0.09	129
289	0.04	0.24	0.07	80
290	0.06	0.23	0.10	99
291	0.25	0.49	0.33	208
292	0.04	0.21	0.07	67
293	0.23	0.52	0.32	109
294	0.14	0.38	0.21	140
295	0.09	0.25	0.14	241
296	0.07	0.17	0.10	72
297	0.08	0.27	0.13	107
298	0.28	0.62	0.38	61
299	0.30	0.64	0.41	77
300	0.07	0.21	0.10	111
301	0.00	0.01	0.00	126
302	0.04	0.11	0.06	73
303	0.25	0.53	0.34	176
304	0.61	0.84	0.70	230
305	0.48	0.78	0.60	156
306	0.25	0.49	0.34	146
307	0.07	0.28	0.12	98
308	0.02	0.08	0.03	78
309	0.09	0.19	0.12	94
310	0.28	0.47	0.35	162
311	0.32	0.63	0.43	116
312	0.10	0.35	0.16	57
313	0.05	0.15	0.07	65
314	0.17	0.38	0.24	138
315	0.23	0.39	0.29	195
316	0.12	0.41	0.19	69
317	0.09	0.26	0.14	134
318	0.22	0.49	0.30	148
319	0.35	0.52	0.42	161
320	0.08	0.31	0.13	104
321	0.36	0.62	0.46	156
322	0.18	0.54	0.27	134
323	0.28	0.43	0.34	232
324	0.09	0.28	0.14	92
325	0.17	0.28	0.21	197
326	0.05	0.16	0.07	126
327	0.02	0.05	0.03	115

328	0.51	0.75	0.61	198
329	0.18	0.41	0.25	125
330	0.13	0.36	0.20	81
331	0.07	0.14	0.10	94
332	0.08	0.23	0.12	56
333	0.07	0.15	0.09	260
334	0.05	0.17	0.08	60
335	0.10	0.24	0.14	110
336	0.22	0.58	0.32	71
337	0.05	0.24	0.09	66
338	0.19	0.45	0.26	150
339	0.01	0.04	0.01	54
340	0.46	0.68	0.55	195
341	0.26	0.57	0.36	79
342	0.10	0.39	0.16	38
343	0.16	0.49	0.24	43
344	0.20	0.35	0.26	68
345	0.20	0.49	0.29	73
346	0.05	0.18	0.08	116
347	0.15	0.59	0.24	111
348	0.03	0.13	0.05	63
349	0.42	0.75	0.54	104
350	0.17	0.57	0.26	44
351	0.12	0.33	0.18	40
352	0.31	0.65	0.42	136
353	0.09	0.39	0.15	54
354	0.08	0.20	0.12	134
355	0.15	0.41	0.21	120
356	0.25	0.45	0.32	228
357	0.33	0.49	0.40	269
358	0.17	0.46	0.25	80
359	0.33	0.61	0.43	140
360	0.11	0.34	0.17	125
361	0.48	0.76	0.59	169
362	0.02	0.12	0.04	56
363	0.54	0.75	0.63	154
364	0.12	0.33	0.18	58
365	0.07	0.24	0.10	71
366	0.37	0.72	0.49	54
367	0.04	0.14	0.06	116
368	0.04	0.15	0.07	54
369	0.03	0.18	0.05	71
370	0.02	0.10	0.03	61
371	0.07	0.23	0.10	71
372	0.21	0.58	0.31	52
373	0.39	0.60	0.47	150
374	0.11	0.47	0.18	93
375	0.02	0.06	0.03	67
376	0.02	0.08	0.04	76
377	0.22	0.38	0.27	106
378	0.03	0.09	0.04	86
379	0.02	0.29	0.03	14
380	0.17	0.61	0.26	122
381	0.03	0.12	0.05	104
382	0.06	0.26	0.10	66
383	0.14	0.43	0.21	110
384	0.04	0.08	0.05	155
385	0.10	0.54	0.16	50
386	0.04	0.16	0.07	64
387	0.07	0.20	0.11	93
388	0.20	0.44	0.27	102

389	0.03	0.09	0.04	108
390	0.67	0.77	0.72	178
391	0.12	0.29	0.17	115
392	0.29	0.52	0.37	42
393	0.04	0.07	0.05	134
394	0.05	0.14	0.08	112
395	0.16	0.33	0.21	176
396	0.07	0.21	0.10	125
397	0.39	0.61	0.47	224
398	0.31	0.71	0.43	63
399	0.02	0.08	0.03	59
400	0.12	0.38	0.18	63
401	0.09	0.35	0.14	98
402	0.14	0.28	0.19	162
403	0.08	0.30	0.13	83
404	0.29	0.84	0.43	19
405	0.07	0.26	0.11	92
406	0.08	0.51	0.14	41
407	0.21	0.40	0.27	43
408	0.29	0.52	0.37	160
409	0.07	0.30	0.11	50
410	0.02	0.21	0.04	19
411	0.16	0.31	0.21	175
412	0.08	0.31	0.12	72
413	0.06	0.14	0.08	95
414	0.09	0.24	0.13	97
415	0.03	0.10	0.05	48
416	0.21	0.48	0.30	83
417	0.05	0.15	0.07	40
418	0.09	0.27	0.13	91
419	0.15	0.42	0.22	90
420	0.07	0.35	0.12	37
421	0.08	0.20	0.11	66
422	0.12	0.49	0.20	73
423	0.09	0.29	0.14	56
424	0.42	0.91	0.58	33
425	0.03	0.09	0.04	76
426	0.04	0.14	0.06	81
427	0.56	0.77	0.65	150
428	0.37	0.76	0.50	29
429	0.93	0.94	0.94	389
430	0.25	0.51	0.34	167
431	0.06	0.18	0.09	123
432	0.11	0.36	0.17	39
433	0.12	0.43	0.19	82
434	0.36	0.71	0.48	66
435	0.24	0.47	0.32	93
436	0.20	0.55	0.29	87
437	0.07	0.21	0.11	86
438	0.37	0.60	0.46	104
439	0.08	0.28	0.13	100
440	0.05	0.11	0.06	141
441	0.18	0.49	0.27	110
442	0.09	0.23	0.13	123
443	0.07	0.18	0.10	71
444	0.12	0.29	0.17	109
445	0.07	0.35	0.12	48
446	0.15	0.45	0.23	76
447	0.03	0.26	0.05	38
448	0.30	0.63	0.41	81
449	0.23	0.37	0.28	132

450	0.16	0.38	0.23	81
451	0.17	0.42	0.24	76
452	0.06	0.16	0.09	44
453	0.03	0.14	0.05	44
454	0.27	0.63	0.38	70
455	0.14	0.37	0.20	155
456	0.11	0.35	0.17	43
457	0.13	0.43	0.20	72
458	0.06	0.24	0.10	62
459	0.11	0.39	0.18	69
460	0.02	0.04	0.03	119
461	0.29	0.42	0.35	79
462	0.06	0.17	0.09	47
463	0.11	0.34	0.17	104
464	0.27	0.50	0.35	106
465	0.12	0.41	0.18	64
466	0.28	0.51	0.36	173
467	0.23	0.54	0.33	107
468	0.15	0.34	0.21	126
469	0.03	0.07	0.04	114
470	0.61	0.85	0.71	140
471	0.20	0.43	0.28	79
472	0.27	0.45	0.34	143
473	0.30	0.47	0.36	158
474	0.09	0.21	0.13	138
475	0.05	0.20	0.08	59
476	0.22	0.42	0.29	88
477	0.45	0.70	0.55	176
478	0.24	0.88	0.37	24
479	0.09	0.21	0.13	92
480	0.35	0.61	0.44	100
481	0.20	0.46	0.27	103
482	0.09	0.34	0.15	74
483	0.39	0.70	0.50	105
484	0.07	0.25	0.11	83
485	0.04	0.18	0.07	82
486	0.09	0.27	0.14	71
487	0.14	0.27	0.18	120
488	0.06	0.16	0.09	105
489	0.24	0.49	0.32	87
490	0.42	0.84	0.56	32
491	0.03	0.09	0.04	69
492	0.01	0.04	0.02	49
493	0.05	0.14	0.07	117
494	0.13	0.36	0.19	61
495	0.73	0.85	0.79	344
496	0.12	0.29	0.17	52
497	0.21	0.35	0.27	137
498	0.11	0.28	0.16	98
499	0.09	0.35	0.14	79
micro avg	0.29	0.48	0.36	173812
macro avg	0.21	0.41	0.27	173812
weighted avg	0.35	0.48	0.40	173812
samples avg	0.37	0.45	0.36	173812

```
Time taken to run this cell : 0:26:24.039445
```

Conclusion:-

In this assignment we improved the performance of the model by applying different techniques such as:-

1. Bag of Words with ngram_range=(1,4)
2. Hyperparameter tuning for alpha of Logistic Regression
3. Training the model using Linear SVM

After performing the above mentioned steps, the best performing model was **Linear Regression** with **alpha=0.001** trained using **Bag of Words** Vectorizer with a **ngram_range=(1,4)**.