```
1 E:\python.exe C:/Users/ROG/Desktop/□□/JD_NLP-
      text classfication Project1/JD NLP1-
      text_classfication_assignment1/model.py
2 ['---\n', '000\n', '0÷000\n', '"0\n', '00\n', '00\n
      ', ':\n', '→\n', '□\n', '&\n', '*\n', '□□\n', '~~~\
n', ''\n', '.\n', '□\n', '.□\n', './\n', '--\n', '□\n
              '0"\n', '0\n', '000\n', '00\n', '0000\n', '0000\n
      ', '00\n', '000\n', '0\n', '//\n', '0\n', '0\n', '000
      0\n', '0000\n', '00\n', '}\n', '00\n', ''\n', '0\n
      ', '0000\n', '0000\n', '0000\n', '0000\n
              '0000\n', '00\n', '00\n', '00\n', '0000\n', ''\n
'***\n', '0000\n', '000\n', '0000\n', '000
      0\n', '0000\n', '0000\n', '0000\n', '0000\n', '0000\n
      ', '0000\n', '0000\n', '0000\n', '0000\n
               '0000\n', '0000\n', '0000\n', '0000\n', '0000\n
              '0000\n', '0000\n', '0000\n', '0000\n', '0.\n
            '0000\n', '.0\n', '00\n', '0000\n', '0/\n', '0000\
              , '0000\n', '0000\n', '0000\n', '0000\n', '0000\n
'0000\n', '0000\n', '0000\n', '0000\n', '0000\n
              '0000\n', '00\n', '00\n', '://\n', ''0\n', '000\n'
'0000\n', '000\n', '...\n
               '000000\n', '0000\n', '0\n', '0000\n', '00'00\n
              '000\n', ''0\n', '00\n', '00\n', '0\n', '0000\n
               '@@@Q\n', '@@@@\n', '@\n', '@@@@\n', '↑\n', '.@\n
              '0000\n', '00\n', '0000\n', '0000\n', '0000\n
               '0000\n', '0000\n', '0000\n', '0000\n
               '0000\n', '0000\n', '000\n', '000\n', '000\n', '000
      0\n', '0000\n', '0000\n', '0000\n', '0000\n', '0000\n'
      ', '0\n', '0\n', '000\n', '0000\n', '0000\n', '0000\n
               '000\n', '::\n', '000\n', '000\n', '000\n', '000\n
              '000\n', '000\n', '000\n', '.....\n', '--\n
               '?\n', '\|\n', '\|\n', \"\n', \"\n', \\|\n\
                                                       '@\n', '@\n', '@\n', '@\n', ',\n
               '□\n', '□\n',
               '-\n', '<\n', '>\n', '\n', '\n\n', '\n\n'
                                    ']\n',
               '[\n',
                                                        '(\n', ')\n', '-\n',
                                                                                                                        '+\n', '∐\n
                                                       '/\n',
               '×\n',
                                   '∐\n',
                                                                            '@\n', '@\n', '@\n',
               B\n', '''\n', ';\n', '\#\n', '@\n', '@\n', '\p\n', '\
      ', '@@\n', '× \n', '@\n', '@\n', 'sub\n', 'exp
        \n', 'sup\n', 'sub\n', 'Lex \n', '\sub\n', '
         , '@\n', '@\n', '@@\n', '@@\n', '@\n', '@\n
               'Δ±\n', 'ΔΔ\n', 'ΔΔ\n', 'ΔΔ\n', 'ΔΔ\n', '=\n', 'Δ\
```

2 n', 'ΔΔ\n', 'ΔΔ\n', 'Δ\n', 'Δ\n', 'Δ\n', 'Δ±\n ', '00\n', '0000\n', '0000\n', '≈ \n', '000 0\n', '0000\n', '00\n', '00\n', '00\n', '....\n ', '0\n', '0000\n', '00\n', '00\n', '00\n', '00\n', 00\n', '00\n', '0\n', '000\n', '0000\n', '0##0\n', '0 \n', '@\n', '@\n', '@@\n', '@@\n', '@@\ n', ' \square \n', ' \square \n '0\n', '00\n', '0\n', '00\n', '00\n', '00\n', '00\n', '0\n', '0\n', '00\n', '00\n', '0\n', '00\n ', '00\n', '00\n', '00\n', '00\n', '00\n', '00\n', ' $00\n'$, $00\n'$ ', '00\n', '00\n', '00\n', '00\n', '00\n', '0\n', '00 \n', '@\n', '@@\n', '@\n', '@\n', '@@@@\n', ' $00\n'$, $00\n'$ ', '@\n', '@\n', '@\n', '@\n', '@\n', '@\n '00\n', '0\n', '00\n', '0\n', '00\n', '00\n ', '00\n', '00 '00\n', '00\n', '0\n', '00\n', '0\n', '0\n', '00\n', '00\n', '00\n', '00\n', '00\n', '0\n', ' '\n', '\o\n', '\o\n', '\o\n', '\o\n', '\o\n', \o\n', \o\n' '00\n', '0\n', '0\n', '00\n', '00\n', '0\n '00\n', '00\n', '0000\n', '000\n', '0\n', '00\n '00\n', '00\n', '0\n', '00\n', '00\n', '0\n', '00\ $\square\square\backslash n'$, $\square\square\backslash n'$, $\square\square\backslash n'$, $\square\backslash n'$, $\square\square\backslash n'$, $\square\square\backslash n'$, $\square\square\backslash n'$, 00\n', '00\n', '00\n', '00\n', '00\n', '00\n' ', '@\n', '@@\n', '@\n', '@@\n', '@@@\n', '@\n', ' '00\n', '00\n', '0\n', '0\n', '0\n', '00\n', 00\n', '00\n', '0\n', '00\n', '00\n', '00\n', '000\n ', '00\n', '00\n', '000\n', '000\n', '00\n' '00\n', '0\n', '00\n', '0\n', '0\n', '0\n', '00\n', '0\n', '00\n', '00\n', '0\n', '0\n', '00\n', '0\n', '00\n', '

```
2 \n', '00\n', '00\n', '00\n', '00000\n', '00\n', '00\n
                    '0\n', '00\n', '00\n', '00\n', '0\n', '00\n', '0\n
'00\n', '00\n', '0\n', '00\n', '00\n', '00\n', '00
           \n', '00\n', '0000\n', '0\n', '00\n', '00\n', '0\n
           ', '00\n', '00\n', '00\n', '00\n', '00\n', '
           00\n', 0
           ', '00\n', '00\n', '0\n', '00\n', '00\n', '00\
                                                                                                            '00\n', '00\n', '0\n', '00\n',
          n', '\\n', '\\\\\\n',
                              '00\n', '0\n',
                                                                                                             '00\n', '00\n', '0\n', '00\n', '0\
           n', '00\n', '0\n', '00\n', '00\n', '00\n', '0
          00\n', '00\n', '00\n', '0\n', '0\n', '0000\n', '0\n
          ', '0\n', '00\n', '0\n', '0\n', '0\n', '0\n', '00\n
                   . '@\n', '@@\n', '@@@\n', '@@@\n', '@@\n', '
           0\n', '00\n', '00\n', '00\n', '00\n', '00\n', '00\n', '000\n'
          ', '00\n', '00\n', '00\n', '00\n', '0\n', '0\\n', '0\\n', '0\\n', '0\\n', '00\\n', '0\\n', '00\\n', '0
           00\n', 00\n', 00\n', 00\n', 00\n', 00\n', 00\n', 00\n', 00\n'
          ', '0\n', '0\n', '0000\n', '00\n', '0\n', '00\n', '00\n', '00\n', '00\n', '00\n', '00\n', '00\n', '0\n', '00\n', '0\n', '00\n', '0\n', '00\n', '0\n', '0\n', '00\n', '0\n', '0\n'
          ', '00\n', '0\n', '00\n', '0\n', '0\n', '0\n', '0\n'
           ', '00\n', '000\n', '000\n', '00\n', '0\n', '
          ', '0000\n', '00\n', '000\n', '0000\n', '0000\n', '000\\n', '00\\n', '00\\n
           \n', '000\n', '00\n', '00\n', '00\n', '00\n', '00\n'
                       '0\n', '00\n', '00\n', '0\n', '00\n', '0\n', '00\n
                         '000\n', '000\n', '00\n', '00\n', '00\n', '000\n
           ', '0000\n', '0000\n', '0000\n', '0000\n', '0
          \n', '00\n', '00\n', '00\n', '00\n', '0\n',
          0\n', '0\n', '0\n', '0\n', '0\n', '0\n', '0\n', '0\n'
           ', '\\n', '\\\n']
                           label
                                                                                                                                                                                                                                                                   text
4 0
                                                    ПППП
                                                                                                                                                               000 0 00 0
                                              0000 0 00 11 0 11 00 11 00 NUM 0 11 00 00
5 1
          00 0 0 ...
6 2
                                                               ΠП
                                                                                  7 3
                                                                                  00 00 0 00 00 000 0 || 0000 00 NUM 0
                                                               00
```

```
7 NUM 🛮 || ...
       ООО
          9 ...
10 28368
                          00 00 11
       00
 11 28369
     00000
                        12 28370 other
  00 00 || 0 00 0 0 000 || 0 00 00 00000
13 28371
      0000
  ( 000 )...
14 28372
       000
           000 0 00 || 00 00 0 || 00 00 0 00
  15
16 [28373 rows x 2 columns]
               17 0
  00 0
       0 0 00 || 0 || 00 || 00 NUM 0 || 00 00 00 0
18 1
 Ø ...
19 2
       20 3
       00 00 0 00 00 000 0 || 0000 00 NUM 0 NUM 0
  || ...
21 4
       Ø ...
22
                       23 28368
 00 00
24 28369
                     00 0 0 0 00 0 00 00
  25 28370
       || ...
26 28371
       )...
       000 0 00 || 00 00 0 || 00 00 0 00 00 || 00
27 28372
  0 00
28 Name: text, Length: 28373, dtype: object
29 0
                  иии иии
       0 00 || 0 || 00 || 00 NUM 0 || 00 00 00 0
30 1
 00...
```

```
31 2
   0...
         00 00 00 00 000 0 || 0000 00 NUM NUM 0 || 00
32 3
   0...
33 4
         0...
34
                          . . .
35 28368
                              00 00
36 28369
                              || 00
         37 28370
  00...
         38 28371
  00...
39 28372
            0 00
40 Name: text, Length: 28373, dtype: object
41
             ... tfidf22709
       label
42 0
        0.0
43 1
        0.0
44 2
                     0.0
          \square
45 3
                     0.0
          \square
46 4
         0.0
47 ...
         . . .
                     . . .
48 28368
          \square
                     0.0
                     0.0
49 28369
       00000
50 28370
                     0.0
       other
51 28371
        0.0
52 28372
         000
                     0.0
53
54 [28373 rows x 22712 columns]
55 transform w2v
56
       label
             ... w2v_4_max_299
57 0
        0.054239
58 1
        0.366395
59 2
          00
                   -0.033314
60 3
          \square
                    0.040619
61 4
         ППП
                    0.097385
62 ...
         . . .
63 28368
          00
                   -0.161887
64 28369
       00000
                   -0.085961
             . . .
65 28370
                    0.060511
       other
```

File - model

```
66 28371
           0.446367
 67 28372
            ИИИ
                         -0.370565
 68
 69 [28373 rows x 25712 columns]
 70
          label
                                     lda_10
 71 0
           [0.0]
 72 1
           [0.15926872193813324]
 73 2
             00
                                      [0.0]
 74 3
             ИИ
                                      [0.0]
 75 4
                                      [0.0]
            000
 76 ...
            . . .
 77 28368
                      [0.41494041681289673]
             00
 78 28369
                      [0.18126417696475983]
          00000
 79 28370
                                      [0.0]
          other
                                      [0.0]
 80 28371
           0000
 81 28372
                                      [0.0]
            ИИИ
 82
 83 [28373 rows x 25723 columns]
 84 Building prefix dict from the default dictionary ...
85 Loading model from cache C:\Users\ROG\AppData\Local\
   Temp\jieba.cache
 86 Loading model cost 0.591 seconds.
 87 Prefix dict has been built successfully.
 88
          label
                 ... mean_word_len
 89 0
           2.000000
 90 1
           1.647727
 91 2
             00
                          1.633333
 92 3
             00
                          1.904762
 93 4
            000
                          1.617647
 94 ...
             . . .
                               . . .
 95 28368
             00
                          2.285714
 96 28369
                          1.600000
          00000
                          1.653846
 97 28370
          other
 98 28371
                          1.757576
           0000
 99 28372
                          1.823529
            000
100
101 [28373 rows x 25740 columns]
102
         label
                                                 text
103 0
                пппп
    00 00 ...
104 1
          пппп
                00 0 00 NUM 0 || 0000 00 0 00 0 NUM 0
    || 00 NU...
105 2
          0000
```

```
0 00 0...
    107 4
   other
              108 ...
109 1575
    \square
      110 1576
    111 1577
    ППП
 00 || 00 0 B
112 1578
    0 0 000 00
113 1579
       0000
 00 00 0 0
114
115 [1580 rows x 2 columns]
    00 ...
117 1
    00 0 00 NUM 0 || 0000 00 0 00 0 NUM 0 || 00
 NU...
    118 2
 Ø...
119 3
    Ø ...
120 4
            0 00
121
122 1575
    1...
    123 1576
 Ø...
               124 1577
 0 00
125 1578
              126 1579
     127 Name: text, Length: 1580, dtype: object
    128 0
 Ø...
```

```
00 0 00 NUM 0 || 0000 00 0 00 NUM 0 || 00
129 1
  NUM-...
        130 2
  00 ...
       131 3
  132 4
                            0 00
133
134 1575
        Ø...
        135 1576
  000...
                              136 1577
   0 00
                           137 1578
  000 00
138 1579
                   00 0
139 Name: text, Length: 1580, dtype: object
140
       label ... tfidf22709
141 0
       0.0
142 1
                    0.0
       0000
143 2
       0.0
144 3
        0.0
       other
145 4
                    0.0
146 ...
        . . .
                    . . .
147 1575
         0.0
148 1576
       0.0
149 1577
       000
                    0.0
150 1578
                    0.0
        000
151 1579
                    0.0
       0000
152
153 [1580 rows x 22712 columns]
154 transform w2v
155
       label ... w2v_4_max_299
156 0
       0000
                  -0.016342
157 1
       0000
                  -0.047609
158 2
                  0.395582
       0000
159 3
                  0.180093
        160 4
       other
                  0.113059
            . . .
161 ...
        . . .
            . . .
162 1575
                  -0.282726
         00
            . . .
```

```
0000
                           0.074011
163 1576
164 1577
            ППП
                          -0.127310
165 1578
            000
                          0.263881
166 1579
           0000
                          -0.318620
167
168 [1580 rows x 25712 columns]
169
          label
                                      lda_10
170 0
           0000
                        [0.2440698742866516]
171 1
                                       [0.0]
           172 2
                        [0.5561922788619995]
           0000
173 3
            000
                                       [0.0]
174 4
                                       [0.0]
          other
175 ...
            . . .
176 1575
                                       [0.0]
             \square
177 1576
           [0.12865708768367767]
178 1577
                        [0.2714894413948059]
            000
179 1578
                                       [0.0]
            000
                                       [0.0]
180 1579
           181
182 [1580 rows x 25723 columns]
183
                 ... mean_word_len
          label
184 0
           0000
                           1.473684
185 1
                           2.194444
           0000
186 2
           1.588889
187 3
            000
                          1.566038
188 4
          other
                          1.416667
189 ...
            . . .
190 1575
             \square
                          1.500000
191 1576
           1.848485
192 1577
            2.142857
193 1578
            000
                          1.666667
194 1579
           1.666667
195
196 [1580 rows x 25740 columns]
197 X_train: (28373, 25738) y_train:
                                        (28373,)
198 E:\Lib\site-packages\lightgbm\engine.py:148:
    UserWarning: Found `num_trees` in params. Will use
    it instead of argument
      _log_warning("Found `{}` in params. Will use it
199
    instead of argument".format(alias))
200 [LightGBM] [Warning] Auto-choosing col-wise multi-
    threading, the overhead of testing was 1.463411
    seconds.
201 You can set `force_col_wise=true` to remove the
```

```
201 overhead.
202 [LightGBM] [Info] Total Bins 832855
203 [LightGBM] [Info] Number of data points in the train
     set: 28373, number of used features: 5152
204 [LightGBM] [Info] Start training from score 4.425581
205 [1] valid_0's rmse: 2.83823
206 [2] valid_0's rmse: 2.76574
207 [3] valid_0's rmse: 2.71759
208 [4] valid_0's rmse: 2.66237
209 [5] valid_0's rmse: 2.60817
210 [6] valid_0's rmse: 2.56308
211 [7] valid_0's rmse: 2.53058
212 [8] valid_0's rmse: 2.50946
213 [9] valid_0's rmse: 2.46862
214 [10]
            valid_0's rmse: 2.44128
215 [11]
            valid_0's rmse: 2.41514
216 [12]
            valid_0's rmse: 2.37837
217 [13]
            valid_0's rmse: 2.35329
218 [14]
            valid_0's rmse: 2.32861
219 [15]
            valid_0's rmse: 2.30758
220 [16]
            valid_0's rmse: 2.29475
221 [17]
            valid_0's rmse: 2.27373
222 [18]
            valid_0's rmse: 2.2581
223 [19]
            valid_0's rmse: 2.24231
224 [20]
            valid_0's rmse: 2.23075
225 [21]
            valid_0's rmse: 2.21192
226 [22]
            valid_0's rmse: 2.19205
227 [23]
            valid_0's rmse: 2.17985
228 [24]
            valid_0's rmse: 2.16711
229 [25]
            valid_0's rmse: 2.15825
230 [26]
            valid_0's rmse: 2.14875
231 [27]
            valid_0's rmse: 2.13851
232 [28]
            valid_0's rmse: 2.13092
233 [29]
            valid_0's rmse: 2.11863
234 [30]
            valid_0's rmse: 2.11206
235 [31]
            valid_0's rmse: 2.10853
236 [32]
            valid_0's rmse: 2.09986
237 [33]
            valid 0's rmse: 2.0915
238 [34]
            valid_0's rmse: 2.08591
            valid_0's rmse: 2.07995
239 [35]
240 [36]
            valid_0's rmse: 2.07387
241 [37]
            valid 0's rmse: 2.07053
242 [38]
            valid_0's rmse: 2.06498
243 [39]
            valid_0's rmse: 2.06091
```

File - model			
244 [40]	valid_0's	rmse:	2.05531
245 [41]	valid_0's	rmse:	2.05146
246 [42]	valid_0's	rmse:	2.04737
247 [43]	valid_0's	rmse:	2.04372
248 [44]	valid_0's	rmse:	2.03764
249 [45]	valid_0's	rmse:	2.03362
250 [46]	valid_0's	rmse:	2.03249
251 [47]	valid_0's	rmse:	2.02904
252 [48]	valid_0's	rmse:	2.02711
253 [49]	valid_0's	rmse:	2.02535
254 [50]	valid_0's	rmse:	2.02224
255 [51]	valid_0's	rmse:	2.0189
256 [52]	valid_0's	rmse:	2.01538
257 [53]	valid_0's	rmse:	2.01516
258 [54]	valid_0's	rmse:	2.01233
259 [55]	valid_0's	rmse:	2.01118
260 [56]	valid_0's	rmse:	2.00848
261 [57]	valid_0's	rmse:	2.00754
262 [58]	valid_0's	rmse:	2.00452
263 [59]	valid_0's	rmse:	2.0013
264 [60]	valid_0's	rmse:	1.99907
265 [61]	valid_0's	rmse:	1.99506
266 [62]	valid_0's	rmse:	1.99334
267 [63]	valid_0's	rmse:	1.99214
268 [64]	valid_0's	rmse:	1.99096
269 [65]	valid_0's	rmse:	1.98934
	valid_0's	rmse:	1.98643
271 [67]	valid_0's	rmse:	1.98466
272 [68]	valid_0's	rmse:	1.98257
273 [69]	valid_0's		
274 [70]	valid_0's	rmse:	1.97952
275 [71]	valid_0's	rmse:	1.97803
276 [72]	valid_0's		
277 [73]	valid_0's	rmse:	1.97558
278 [74]	valid_0's	rmse:	1.97354
279 [75]	valid_0's		
280 [76]	valid_0's		
281 [77]	valid_0's		
282 [78]	valid_0's		
283 [79]	valid_0's		
284 [80]	valid_0's		
285 [81]	valid_0's		
286 [82]	valid_0's		
287 [83]	valid_0's	rmse:	1.96513

```
File - model
288 [84]
            valid_0's rmse: 1.96347
289 [85]
            valid_0's rmse: 1.96121
290 [86]
            valid_0's rmse: 1.95968
291 [87]
            valid_0's rmse: 1.95881
292 [88]
            valid_0's rmse: 1.95823
293 [89]
            valid 0's rmse: 1.95686
294 [90]
            valid_0's rmse: 1.95618
295 [91]
            valid_0's rmse: 1.95569
296 [92]
            valid_0's rmse: 1.95489
297 [93]
            valid_0's rmse: 1.95358
298 [94]
            valid_0's rmse: 1.95331
299 [95]
            valid_0's rmse: 1.95066
300 [96]
            valid_0's rmse: 1.94991
301 [97]
            valid_0's rmse: 1.94969
302 [98]
            valid_0's rmse: 1.94835
303 [99]
            valid_0's rmse: 1.94694
304 [100]
           valid_0's rmse: 1.94564
305 [3.41703677 6.37474461 5.23204656 ... 3.35050522 2.
    6044324 4.781351991
306 0.36835443037974686
307
308 Process finished with exit code 0
309
```