data_stat

September 11, 2018

```
In [6]: #!/usr/env/bin python3
        import os
        import sys
        import numpy as np
        import json
        import matplotlib.pyplot as plt
        %matplotlib inline
        from matplotlib.font_manager import FontProperties
```

```
In [7]: def getChineseFont():
            return FontProperties(fname='/System/Library/Fonts/PingFang.ttc',size=15)
```

```
In [8]: infile = './dataset/maimai_sample_5w.txt'
```

```
In [9]: '''
        1. likes
        2.
        3. text
        4. keysdict_keys(['is_freeze', 'username', 'id', 'likes', 'text', 'search_qs', 'search_
        '''
```

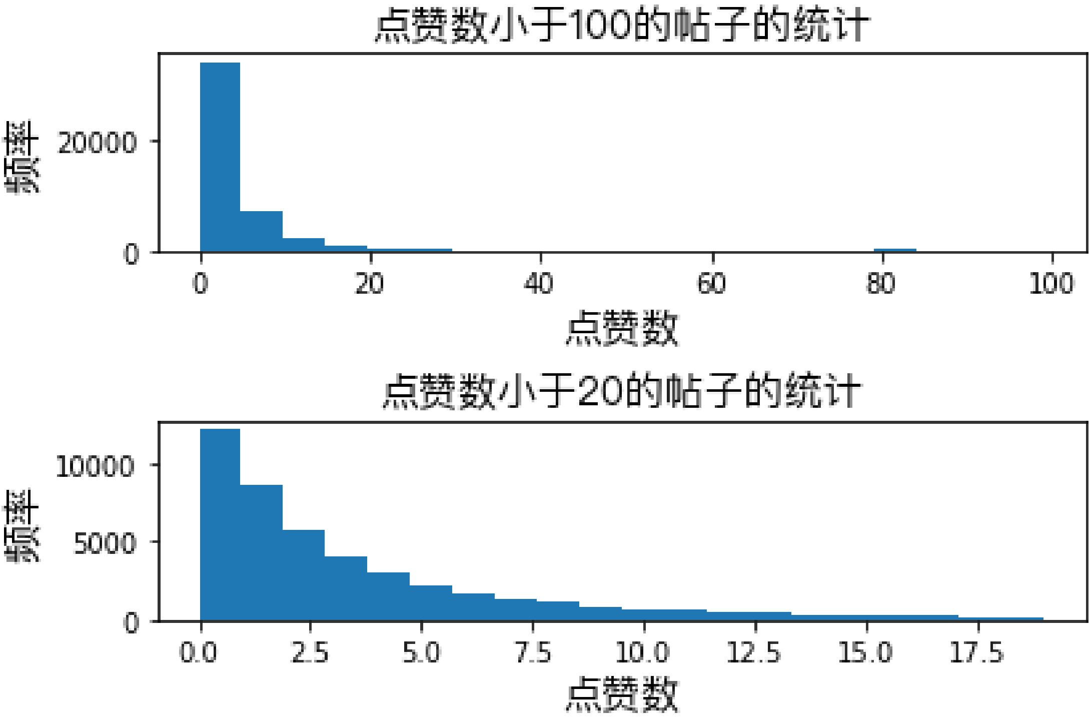
```
Out[9]: "\n1. likes\n2. \n3. text\n4. keysdict_keys(['is_freeze', 'username', 'id', 'likes', '
```

```
In [10]: gossip_set = []
         gossip_set_likes = []
         gossip_set_usernames = []
         gossip_set_texts = []
```

```
In [14]: with open(infile, 'r', encoding='utf8') as f:
             for line in f.readlines():
                 line = line.strip()
                 if line is None or line == '':
                     continue
                 gossip_item = json.loads(line)
                 gossip_set_likes.append(int(gossip_item['likes']))
                 gossip_set_usernames.append(gossip_item['username'])
                 gossip_set_texts.append(gossip_item['text'])
                 gossip_set.append(gossip_item)
```

```
In [51]: print('', len(gossip_set))
         # likes
         likes_max = np.max(np.array(gossip_set_likes))
         print('', likes_max)
         likes_min = np.min(np.array(gossip_set_likes))
         print('', likes_min)
         likes_exception = sum(int(i)>100 for i in gossip_set_likes)
         print('100: ', likes_exception)
         likes_zero = sum(int(i)==0 for i in gossip_set_likes)
         print('0: ', likes_zero)
         plt.figure()
         plt.subplot(211)
         plt.hist(np.array([i for i in gossip_set_likes if i < 100]), bins=20)
         plt.xlabel('', fontproperties=getChineseFont())
         plt.ylabel('', fontproperties=getChineseFont())
         plt.title('100', fontproperties=getChineseFont())
         plt.subplot(212)
         plt.hist(np.array([i for i in gossip_set_likes if i < 20]), bins=20)
         plt.xlabel('', fontproperties=getChineseFont())
         plt.ylabel('', fontproperties=getChineseFont())
         plt.title('20', fontproperties=getChineseFont())
         plt.tight_layout()
```

```
 50000
 4223
 0
100:  1460
0:  12222
```



```
In [41]: # 10
         gossip_set_likes_texts = zip(gossip_set_likes, gossip_set_texts)
         gossip_set_likes_texts = sorted(gossip_set_likes_texts, key=lambda x: x[0], reverse=T
         tmp = gossip_set_likes_texts[0:10]
         print('\n\n'.join(list(zip(*tmp))[1]))
```

```
sdnhr

:

...[][][]

HRofferP7P75P6...P9[]



2018/ToB8+211
```

p6□□□

58 T5 23k 500 2000 58T5

```
In [44]: #
         from collections import Counter
         username_counter = Counter(gossip_set_usernames)
         username_counter = sorted(username_counter.items(), key=lambda x: x[1], reverse=True)
         username_counter[:10]

Out[44]: [('****', 4654),
          ('', 980),
          ('', 356),
          ('', 331),
          ('', 330),
          ('', 329),
          ('', 317),
          ('', 311),
          ('', 308),
          ('', 305)]

In [73]: #
         gossip_set_usernames_texts = zip(gossip_set_usernames, gossip_set_likes, gossip_set_te
         gossip_set_bdyg_texts = [i for i in gossip_set_usernames_texts if i[0] == '']
         gossip_set_bdyg_texts = sorted(gossip_set_bdyg_texts, key=lambda x: x[1], reverse=True
         gossip_set_bdyg_texts[:100]

Out[73]: [('', 378, '□'),
          ('', 276, '~ '),
          ('',
           261,
           ''),
          ('', 223, 'feed'),
          ('', 201, '37000'),
          ('',
           198,
           ''),
          ('', 197, 'erp'),
          ('', 186, 'PMOBATPMPMO'),
          ('', 178, 'omg    '),
          ('', 160, ''),
          ('', 159, ''),
          ('', 153, '360 AI lab'),
          ('', 139, '16k'),
          ('', 136, ''),
          ('', 134, '□□□'),
          ('',
           131,
           ''),
          ('', 130, '          '),
          ('', 130, ''),
          ('', 126, '3.3  '),
          ('',
           124,
           'feed9001/2000'),
          ('', 121, '16'),
          ('', 120, 'transbot[][]'),
          ('',
           119,
           '[]TP~[]'),
          ('', 116, ''),
          ('', 116, ''),
          ('',
           114,
           '[]'),
          ('', 107, 'P7T65060Low\n'),
          ('',
           106,
           '2hr'),
          ('',
           104,
           '\n\nFeed \nAIIDLPm'),
          ('',
           100,
           '57[]U6[]'),
          ('', 100, ''),
          ('', 99, ''),
          ('', 96, ''),
          ('', 94, 'PPT[]'),
          ('', 93, '58'),
          ('',
           92,
           ''),
          ('', 88, ' 9'),
          ('', 88, '16 26kAI'),
          ('', 88, 'ppptb'),
          ('', 88, '10'),
          ('', 85, ''),
          ('', 85, ''),
          ('', 85, ''),
          ('', 83, 'leader'),
          ('', 81, '5?'),
          ('', 80, ' KPI [][][]'),
          ('', 80, '3~ --- '),
          ('', 80, '11t5t6'),
          ('', 79, ''),
          ('', 78, ''),
          ('',
           77,
           ''),
          ('', 75, ''),
          ('', 74, '34[]'),
          ('', 72, '242'),
          ('', 70, '\n\n\n\n'),
          ('', 65, '\n'),
          ('', 65, ''),
          ('', 65, ''),
          ('', 57, 'sng'),
          ('', 54, '\n\n  '),
          ('', 54, ''),
          ('', 54, '\n\n'),
          ('', 49, 'pm'),
          ('', 48, '18'),
          ('', 47, 'T3T4'),
          ('', 46, ''),
          ('', 45, ''),
          ('', 42, '17*16 ()\n\n     20*14\n\n'),
          ('', 41, ''),
          ('', 39, ''),
          ('', 37, '3'),
          ('', 35, 't6hr'),
          ('', 35, ''),
          ('', 34, '0.60.6*0.3+1*0.7*2=1.76[]'),
          ('', 33, 'feed120'),
          ('', 32, ''),
          ('', 30, '  '),
          ('', 29, 'offer'),
          ('',
           27,
           ':1.2.3.4.:'),
          ('', 26, ''),
          ('',
           24,
           'qiqibat\nRobin\n'),
          ('', 23, '21211t330t6t6'),
          ('', 23, 'hr'),
          ('', 23, ''),
          ('', 22, '[][][][]'),
          ('', 22, 'T3.1T?'),
          ('', 21, 'IT'),
          ('',
           21,
           'erp'),
          ('',
           21,
           '29(48)T420K[]'),
          ('', 19, ''),
          ('', 19, 'PM'),
          ('', 19, '2p415k'),
          ('', 19, ''),
          ('',
           18,
           '15[]510'),
          ('', 17, ' '),
          ('', 17, ' '),
          ('', 17, ''),
          ('',
           17,
           'offer\n1. java sp \n2.  java \n56\n '),
          ('', 17, '2012'),
          ('', 16, '')]
```

```
In [66]: # /100/0
         import jieba
         gossip_set_words = []
         for i in gossip_set_texts:
             seg_words = list(jieba.cut(i))
             gossip_set_words += seg_words
         words_freq_counter = Counter(gossip_set_words)
         words_freq_counter = sorted(words_freq_counter.items(), key=lambda x: x[1], reverse=T
```

```
In [70]: words_freq_counter = [i for i in words_freq_counter if len(i[0]) >= 2]
         words_freq_counter[:50]
```

```
Out[70]: [('', 12017),
          ('', 8912),
          ('', 6979),
          ('', 6676),
          ('', 5983),
          ('', 5902),
          ('', 5862),
          ('', 4629),
          ('', 4168),
          ('', 3788),
          ('', 3491),
          ('', 3415),
          ('', 3136),
          ('', 3032),
          ('', 2999),
          ('offer', 2944),
          ('', 2920),
          ('', 2916),
          ('', 2912),
          ('', 2806),
          ('', 2698),
          ('', 2680),
          ('', 2530),
          ('', 2492),
          ('', 2424),
          ('', 2289),
          ('', 2282),
          ('', 2168),
          ('', 2157),
          ('', 2057),
          ('', 2026),
          ('', 2002),
          ('', 1967),
          ('', 1889),
          ('360', 1864),
          ('', 1818),
          ('', 1814),
          ('', 1807),
          ('', 1783),
          ('', 1714),
          ('', 1657),
          ('', 1606),
          ('', 1599),
          ('', 1589),
          ('', 1579),
          ('', 1565),
          ('', 1542),
          ('', 1510),
          ('', 1508),
          ('', 1498)]
```

```
In [75]: # 0
         # todo
         gossip_set_likes_texts = zip(gossip_set_likes, gossip_set_texts)
         gossip_set_likes_texts_0 = [i for i in gossip_set_likes_texts if i[0] > 0]
         gossip_set_texts_0 = list(zip(*gossip_set_likes_texts_0))[1]
         gossip_set_texts_0
         import jieba
         gossip_set_words_0 = []
         for i in gossip_set_texts_0:
             seg_words = list(jieba.cut(i))
             gossip_set_words_0 += seg_words
         words_freq_counter_0 = Counter(gossip_set_words_0)
         words_freq_counter_0 = sorted(words_freq_counter_0.items(), key=lambda x: x[1], revers
```

```
In [76]: words_freq_counter_0 = [i for i in words_freq_counter_0 if len(i[0]) >= 2]
         words_freq_counter_0[:50]
```

```
Out[76]: [('', 9091),
          ('', 6711),
          ('', 5151),
          ('', 5020),
          ('', 4495),
          ('', 4451),
          ('', 4428),
          ('', 3443),
          ('', 3178),
          ('', 2931),
          ('', 2610),
          ('', 2587),
          ('', 2341),
          ('', 2234),
          ('', 2233),
          ('', 2230),
          ('', 2200),
          ('offer', 2190),
          ('', 2151),
          ('', 2122),
          ('', 2020),
          ('', 1992),
          ('', 1928),
          ('', 1854),
          ('', 1801),
          ('', 1721),
          ('', 1693),
          ('', 1646),
          ('', 1594),
          ('', 1554),
          ('', 1515),
          ('', 1496),
          ('', 1491),
          ('', 1452),
          ('', 1421),
          ('', 1390),
          ('', 1383),
          ('360', 1381),
          ('', 1341),
          ('', 1280),
          ('', 1250),
          ('', 1226),
          ('', 1207),
          ('', 1207),
          ('', 1190),
          ('', 1165),
          ('', 1164),
          ('', 1151),
         ('', 1134),
         ('', 1109)]
```

```
In [77]: # 100
         # todo
         gossip_set_likes_texts = zip(gossip_set_likes, gossip_set_texts)
         gossip_set_likes_texts_100 = [i for i in gossip_set_likes_texts if i[0] > 0]
         gossip_set_texts_100 = list(zip(*gossip_set_likes_texts_100))[1]
         gossip_set_texts_100
         import jieba
         gossip_set_words_100 = []
         for i in gossip_set_texts_100:
             seg_words = list(jieba.cut(i))
             gossip_set_words_100 += seg_words
         words_freq_counter_100 = Counter(gossip_set_words_100)
         words_freq_counter_100 = sorted(words_freq_counter_100.items(), key=lambda x: x[1], re
```

```
In [78]: words_freq_counter_100 = [i for i in words_freq_counter_100 if len(i[0]) >= 2]
         words_freq_counter_100[:50]
```

```
Out[78]: [('', 9091),
          ('', 6711),
          ('', 5151),
          ('', 5020),
          ('', 4495),
          ('', 4451),
          ('', 4428),
          ('', 3443),
          ('', 3178),
          ('', 2931),
          ('', 2610),
          ('', 2587),
          ('', 2341),
          ('', 2234),
          ('', 2233),
          ('', 2230),
          ('', 2200),
          ('offer', 2190),
          ('', 2151),
          ('', 2122),
          ('', 2020),
          ('', 1992),
          ('', 1928),
          ('', 1854),
          ('', 1801),
          ('', 1721),
          ('', 1693),
          ('', 1646),
          ('', 1594),
          ('', 1554),
          ('', 1515),
          ('', 1496),
          ('', 1491),
          ('', 1452),
          ('', 1421),
          ('', 1390),
          ('', 1383),
          ('360', 1381),
          ('', 1341),
          ('', 1280),
          ('', 1250),
          ('', 1226),
          ('', 1207),
          ('', 1207),
          ('', 1190),
          ('', 1165),
          ('', 1164),
          ('', 1151),
          ('', 1134),
          ('', 1109)]
```