

# 10月留言分析（品牌、行业）

- 直接从pysql读取
  - 以留言动作为主体
  - 以用户各项属性为主体
- 分析留言分布
  - 区分推荐和自然
- 分析留言文本
- 分析用户生命周期
  - 第几天用户会有再咨询需求

## chapter 1 品牌和行业综述

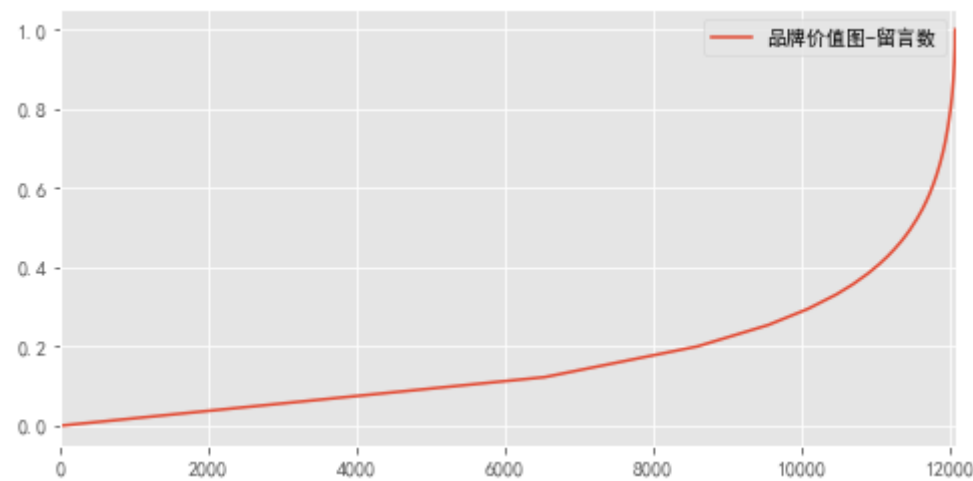
Out[56]:

	id	一级行业	二级行业	品牌	留言数
0	1	建材	其他	定荣家	1312
1	2	服装	其他	行业工具栏400电话	1079
2	3	餐饮	小吃	廖排骨	363
3	4	餐饮	饮品	一点点奶茶饮品	285
4	5	餐饮	其他	粒上皇	259

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12078 entries, 0 to 12077
Data columns (total 5 columns):
id          12078 non-null int64
一级行业    12049 non-null object
二级行业    11968 non-null object
品牌        12078 non-null object
留言数      12078 non-null int64
dtypes: int64(2), object(3)
memory usage: 471.9+ KB
```

Out[58]:

<matplotlib.legend.Legend at 0xa4b8198>



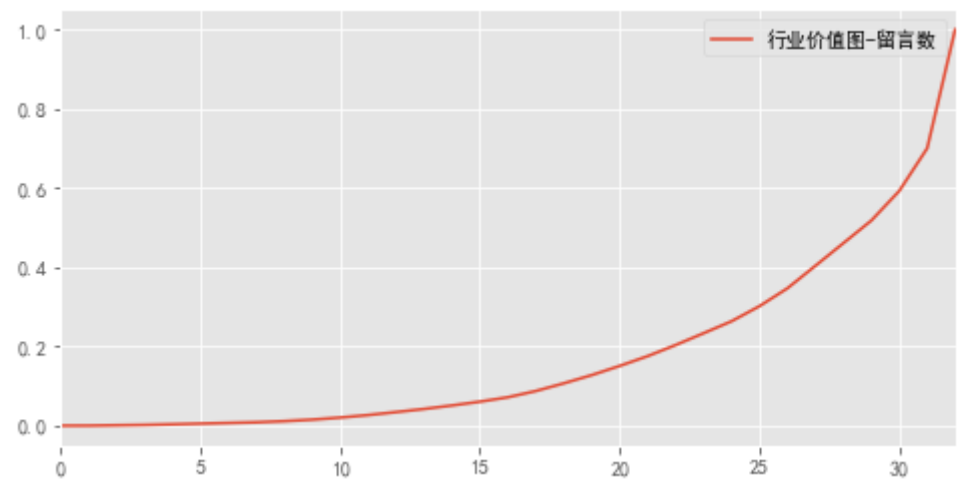
83%的品牌贡献了30%的留言

Out[63]:

	id	行业	留言数
0	1	餐饮	10453
1	2	服装	3759
2	3	教育	2607
3	4	美容	1991
4	5	幼儿	1991

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 33 entries, 0 to 32  
Data columns (total 3 columns):  
id 33 non-null int64  
行业 32 non-null object  
留言数 33 non-null int64  
dtypes: int64(2), object(1)  
memory usage: 872.0+ bytes

Out[65]:  
<matplotlib.legend.Legend at 0x9ac0390>



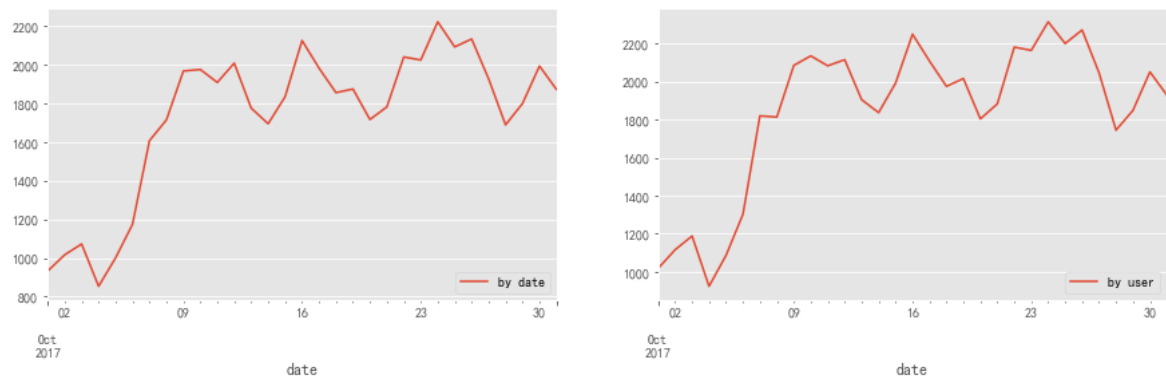
94%的行业产生了60%的留言

## chapter 2 数据清洗——个体异常值

Out[71]:

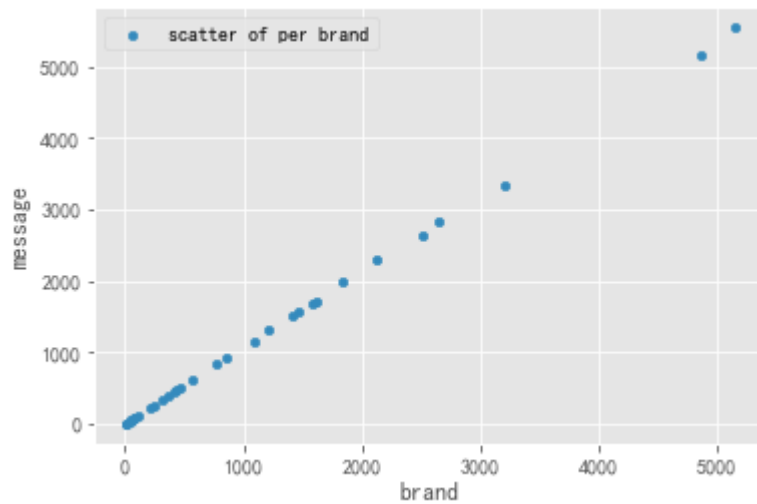
	user_id	raw_date	industry	brand	message	date
0	291312	20171001	餐饮	玛努卡甜品	1	2017-10-01
1	171578	20171001	建材	三棵树涂料	1	2017-10-01
2	291424	20171001	酒水	稻花香	1	2017-10-01
3	290440	20171001	餐饮	勾魂凉皮	1	2017-10-01
4	291336	20171001	教育	小荷作文	1	2017-10-01

Out[93]:  
<matplotlib.legend.Legend at 0x1452acc0>



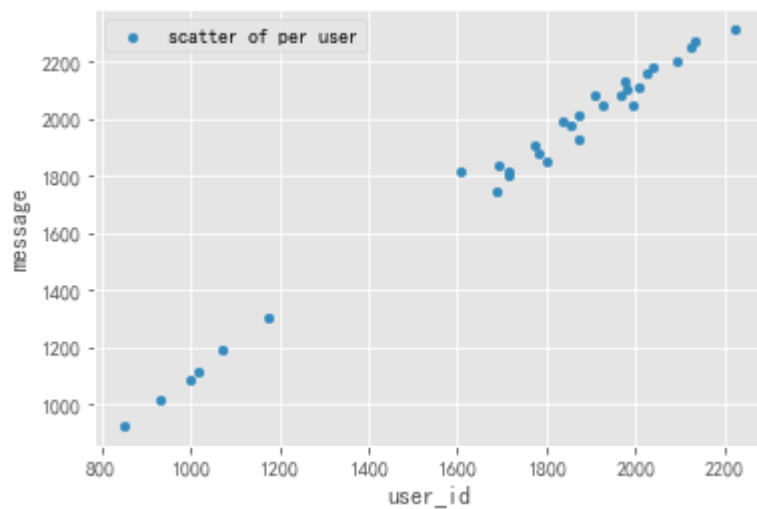
Out[95]:

<matplotlib.legend.Legend at 0x1452aef0>



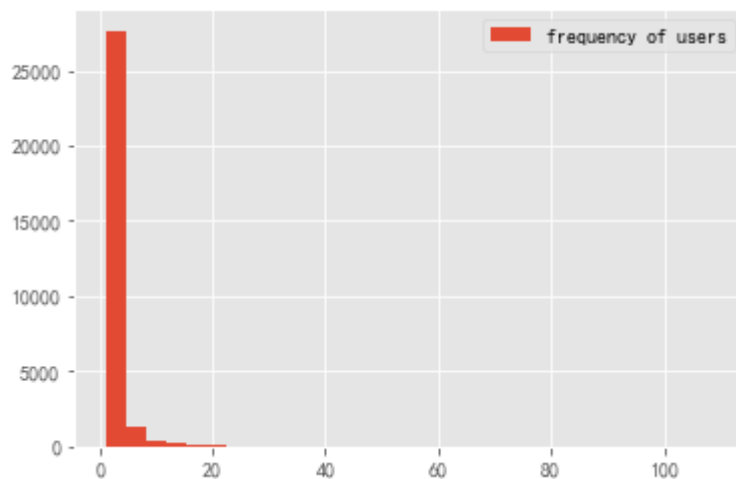
Out[96]:

<matplotlib.legend.Legend at 0x1174e160>



Out[102]:

<matplotlib.legend.Legend at 0x1196f438>



## chapter 3 头部品牌

- 品牌分析
  - 构成
  - 发源地
  - 口碑

Out[106]:

	id	一级行业	二级行业	品牌	留言数
2	3	餐饮	小吃	廖排骨	363
3	4	餐饮	饮品	一点点奶茶饮品	285
4	5	餐饮	其他	粒上皇	259
5	6	餐饮	快餐	贝克汉堡	259
6	7	精品	其他	爱刮刮自助彩票机	240

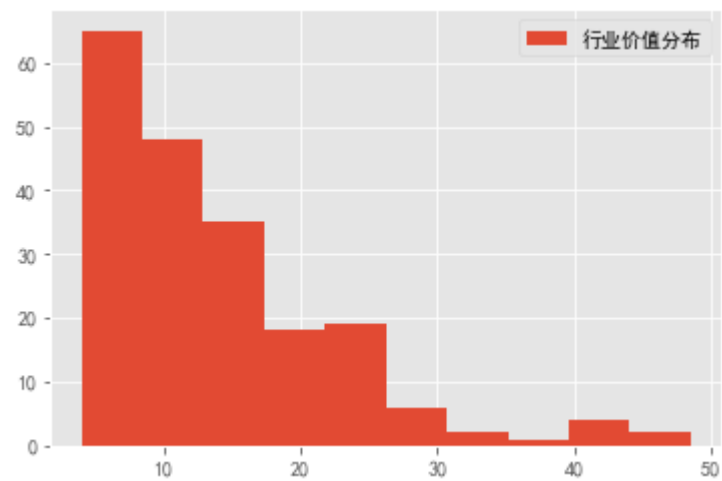
Out[110]:

		留言数	品牌数	平均留言数
一级行业	二级行业			
	小吃	2946	137	21.503650
餐饮	快餐	2900	132	21.969697
	饮品	1895	71	26.690141
教育	英语	1397	65	21.492308
餐饮	火锅	1288	76	16.947368

行业价值 = 平均带来留言数

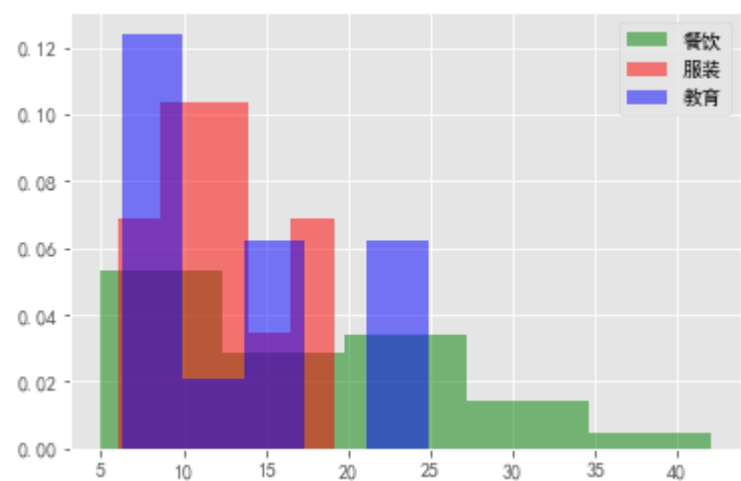
Out[112]:

<matplotlib.legend.Legend at 0x163b07b8>

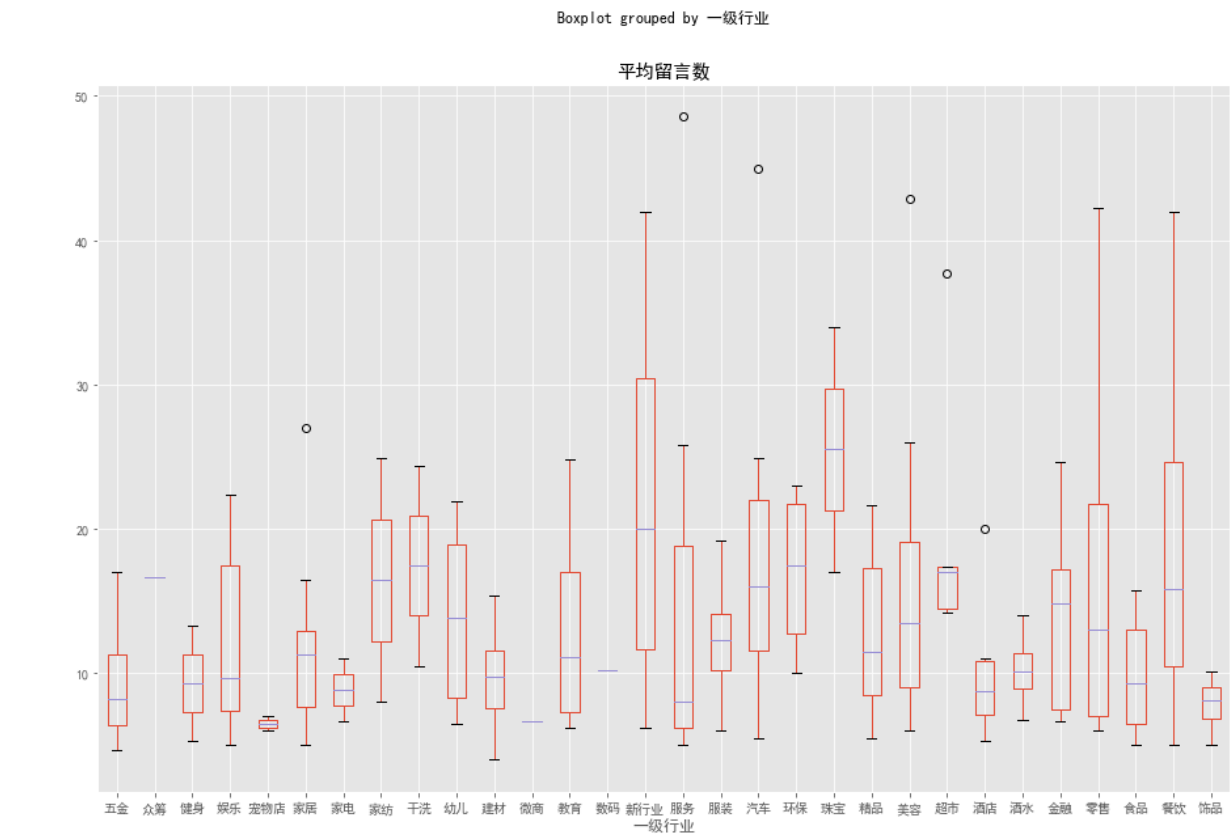


Out[116]:

<matplotlib.legend.Legend at 0x166f9400>



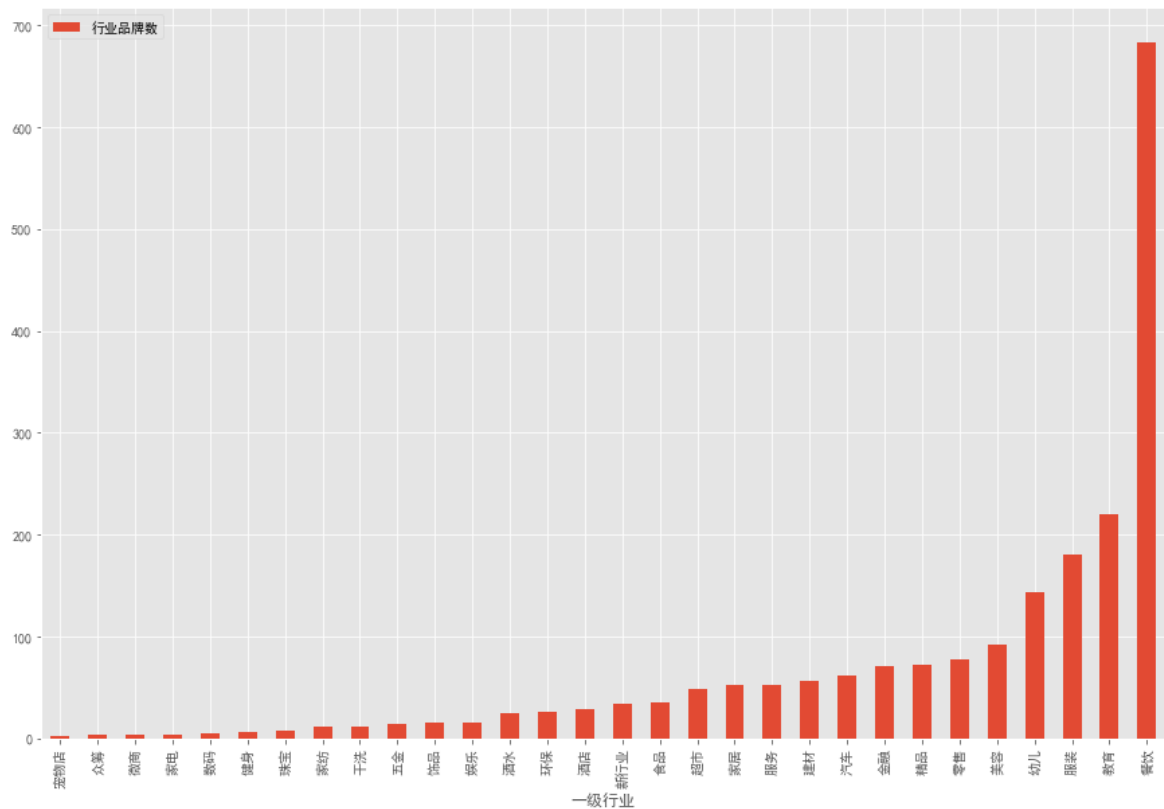
```
Out[133]:
<matplotlib.axes._subplots.AxesSubplot at 0x1c066898>
```



- 朝阳行业：新行业、珠宝、干洗、环保
- 衰退行业：餐饮、教育、零售

Out[132]:

```
<matplotlib.legend.Legend at 0x1c2f82e8>
```



传统优势行业（项目数前五名/占比64%的行业）性价比降低，新兴行业崛起明显

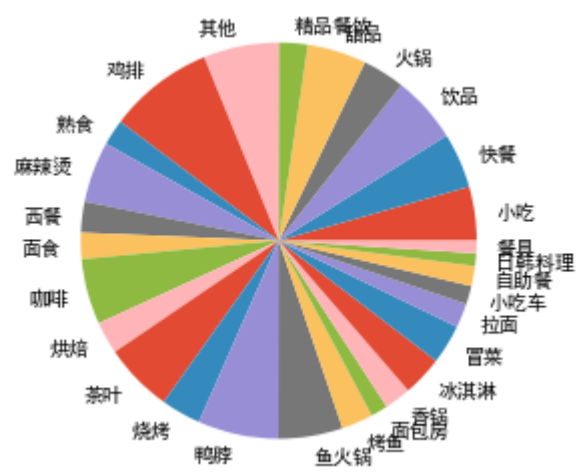
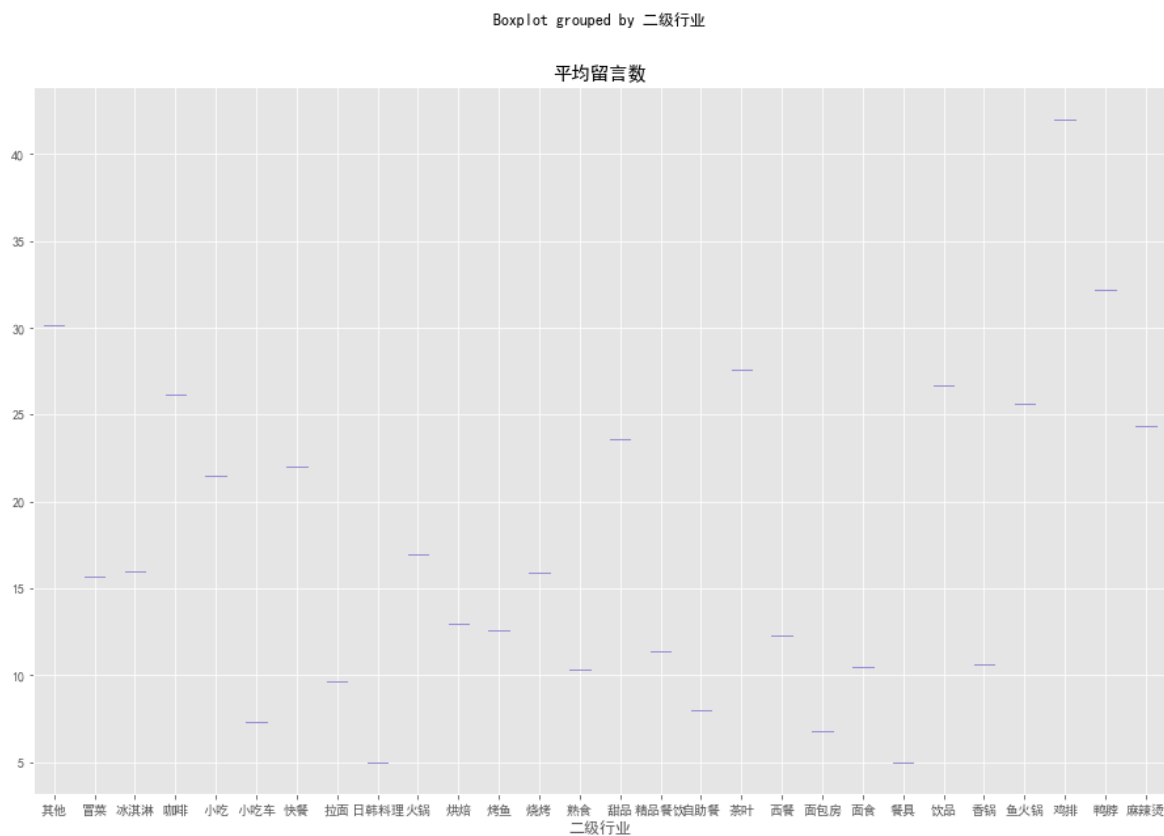
## chapter 4 按需供给——快餐

Out[141]:

	一级行业	二级行业	留言数	品牌数	平均留言数
0	餐饮	小吃	2946	137	21.503650
1	餐饮	快餐	2900	132	21.969697
2	餐饮	饮品	1895	71	26.690141
4	餐饮	火锅	1288	76	16.947368
10	餐饮	甜品	660	28	23.571429
21	餐饮	精品餐饮	410	36	11.388889
22	餐饮	其他	392	13	30.153846
23	餐饮	鸡排	378	9	42.000000
24	餐饮	熟食	361	35	10.314286
27	餐饮	麻辣烫	316	13	24.307692



Out[143]:  
<matplotlib.axes.\_subplots.AxesSubplot at 0x1827f8d0>



进行词频统计

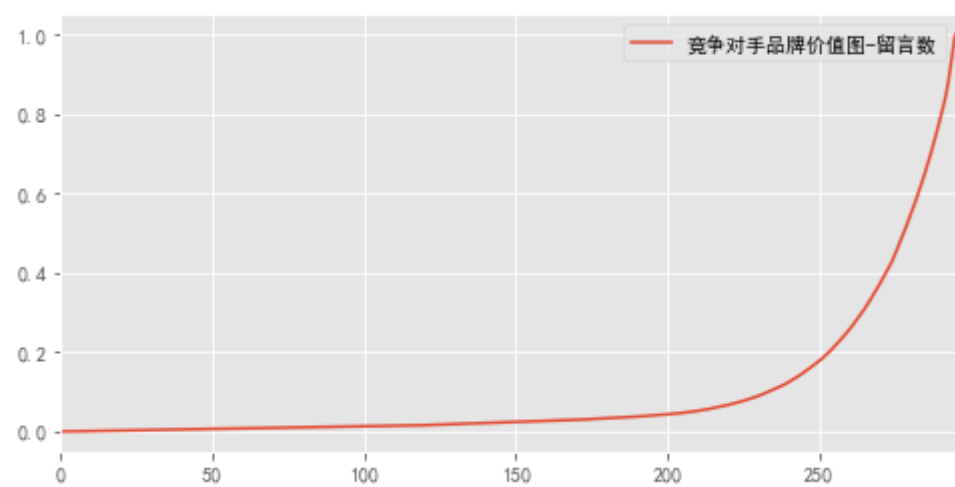
chapter 5 竞争对手

Out[184]:

	id	brand_name	count_msg	publish_date	date_gap	per_month
0	1	姥娘包子	1659	2017-07-01	4	415.0
1	2	将就中式快餐	1229	2017-08-01	3	410.0
2	3	锅先森卤肉饭	1368	2017-07-01	4	342.0
3	4	江湖纸包鱼	505	2017-09-01	2	252.0
4	5	瓦罐香沸快餐	950	2017-07-01	4	238.0

Out[186]:

<matplotlib.legend.Legend at 0x20758e10>



84%的品牌带来了20%的留言（反证留言取值正确的可能性）

进行词频统计