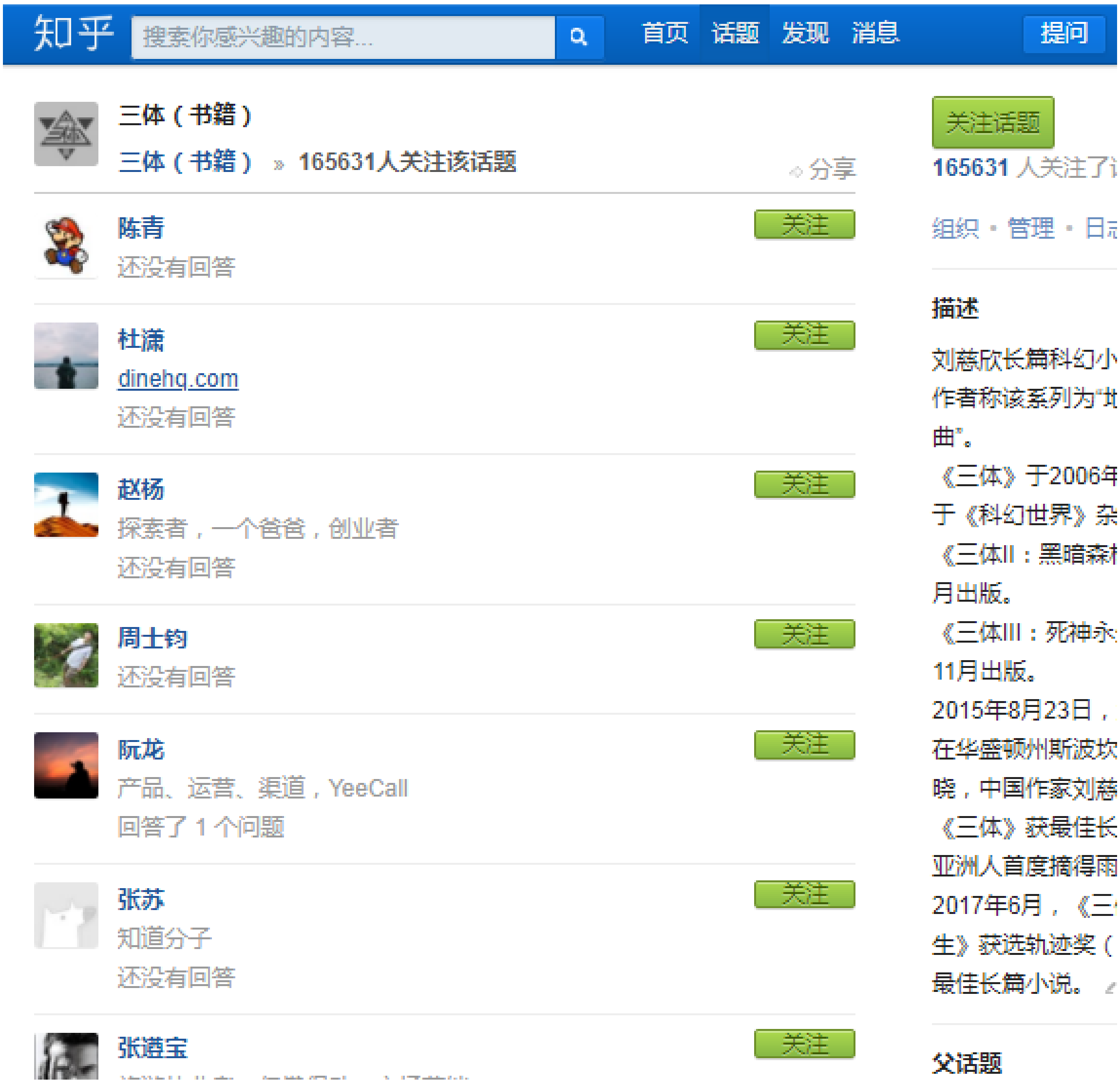
Descriptive Analysis of Userinformation

1. Data obtainment：

Start url: https://www.zhihu.com/topic/19559995/followers

Crawl basic information of users following the topic ‘三体’.



三体（书籍）

三体（书籍） » 165631人关注该话题

陈青
还没有回答

杜潇
dinehq.com
还没有回答

赵杨
探索者，一个爸爸，创业者
还没有回答

周士钧
还没有回答

阮龙
产品、运营、渠道，YeeCall
回答了 1 个问题

张苏
知道分子
还没有回答

张道宝

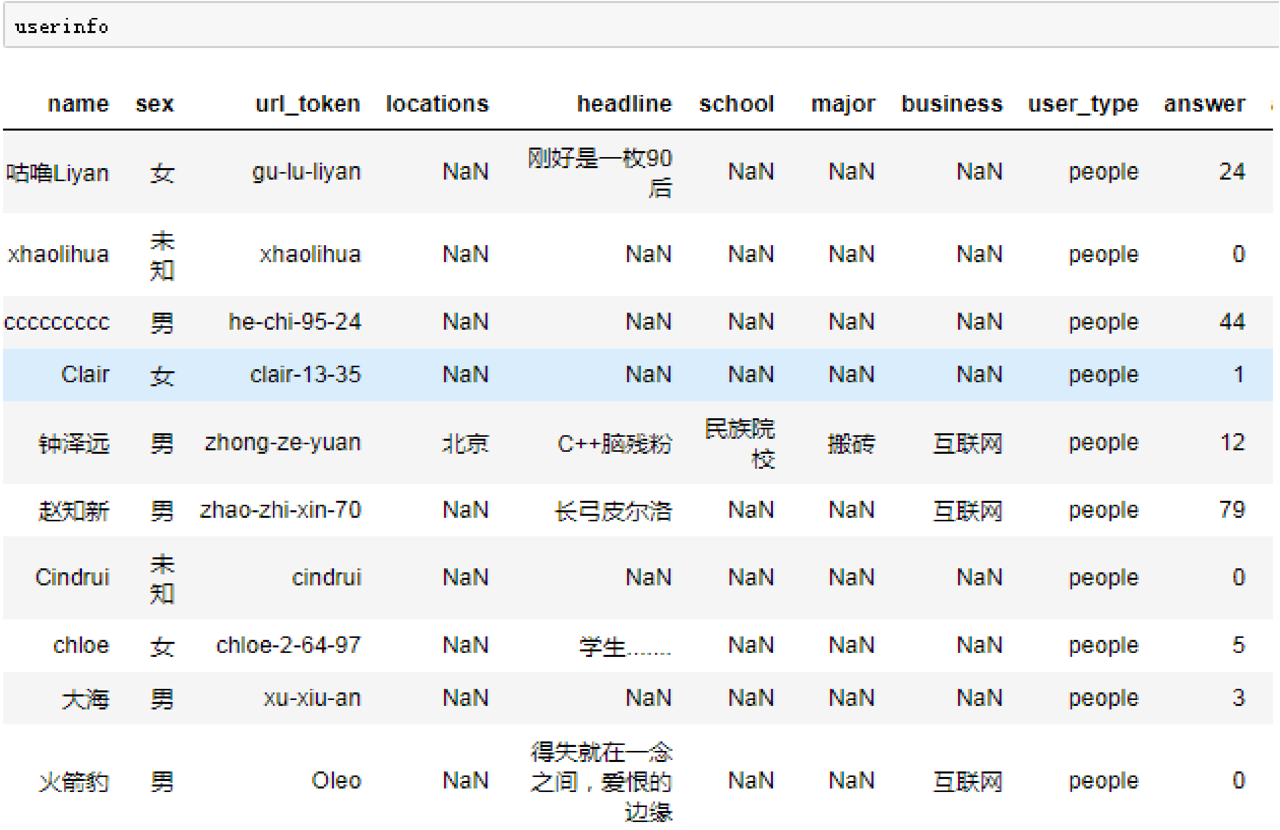
165631 人关注了

刘慈欣长篇科幻小
作者称该系列为"地
曲"。
《三体》于2006年
于《科幻世界》杂
《三体II：黑暗森
月出版。
《三体III：死神永
11月出版。
2015年8月23日，
在华盛顿州斯波坎
晓，中国作家刘慈
《三体》获最佳长
亚洲人首度摘得雨
2017年6月，《三
生》获选轨迹奖（
最佳长篇小说。

2. Data processing：

Transform raw data with the same meaning but different words to a unified form. For example, in ‘locations’, replace ‘广州市’, ‘广东广州’,‘广州市越秀区’with ‘广州’,

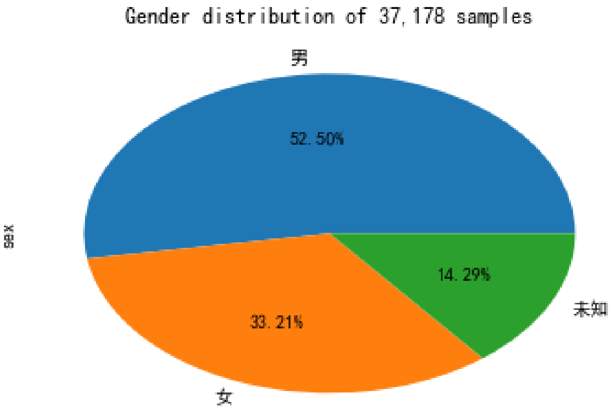
replace ‘纽约’,’加州’,’洛杉矶’ with ‘美国’; in schools, replace‘上交’,‘上海交大’,’上海交通大学医学院’ with ‘上海交通大学’; in majors, replace ‘计算机技术’，‘计算机科学技术’,‘计算机科学’with ‘计算机’. Raw data:



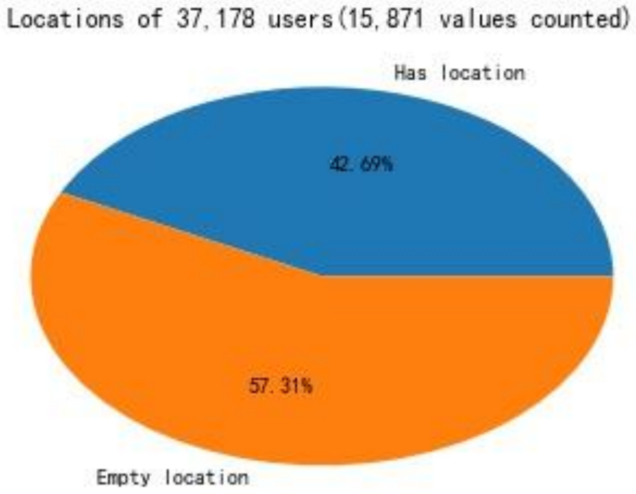
```
userinfo
```

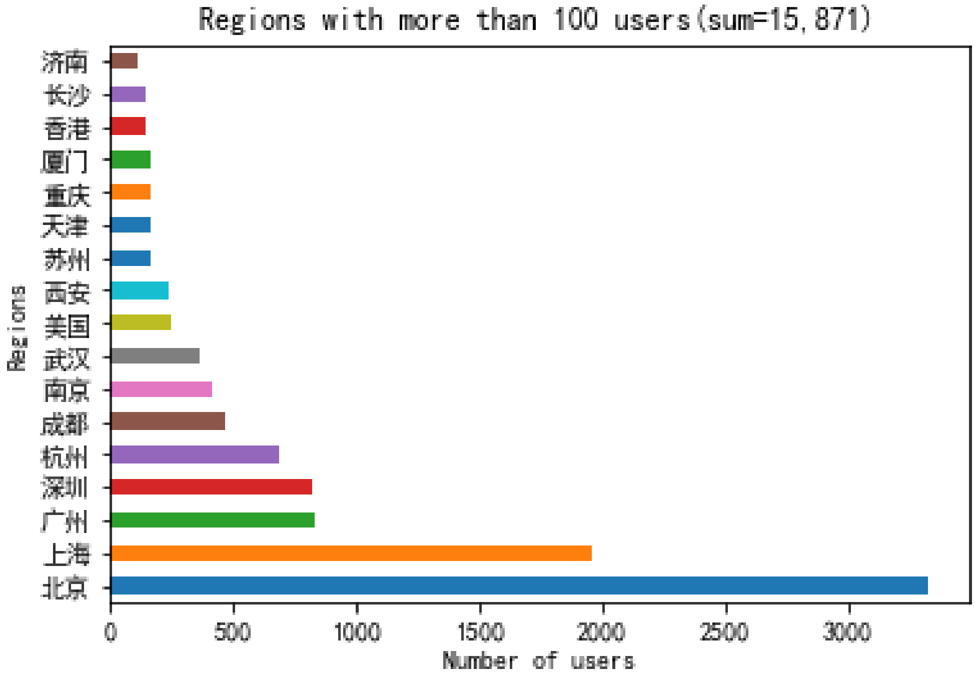
name	sex	url_token	locations	headline	school	major	business	user_type	answer
咕噜Liyan	女	gu-lu-liyan	NaN	刚好是一枚90后	NaN	NaN	NaN	people	24
xhaolihua	未知	xhaolihua	NaN	NaN	NaN	NaN	NaN	people	0
ccccccccc	男	he-chi-95-24	NaN	NaN	NaN	NaN	NaN	people	44
Clair	女	clair-13-35	NaN	NaN	NaN	NaN	NaN	people	1
钟泽远	男	zhong-ze-yuan	北京	C++脑残粉	民族院校	搬砖	互联网	people	12
赵知新	男	zhao-zhi-xin-70	NaN	长弓皮尔洛	NaN	NaN	互联网	people	79
Cindrui	未知	cindrui	NaN	NaN	NaN	NaN	NaN	people	0
chloe	女	chloe-2-64-97	NaN	学生......	NaN	NaN	NaN	people	5
大海	男	xu-xiu-an	NaN	NaN	NaN	NaN	NaN	people	3
火箭豹	男	Oleo	NaN	得失就在一念之间，爱恨的边缘	NaN	NaN	互联网	people	0

3. Gender distribution : Males occupy more than half.



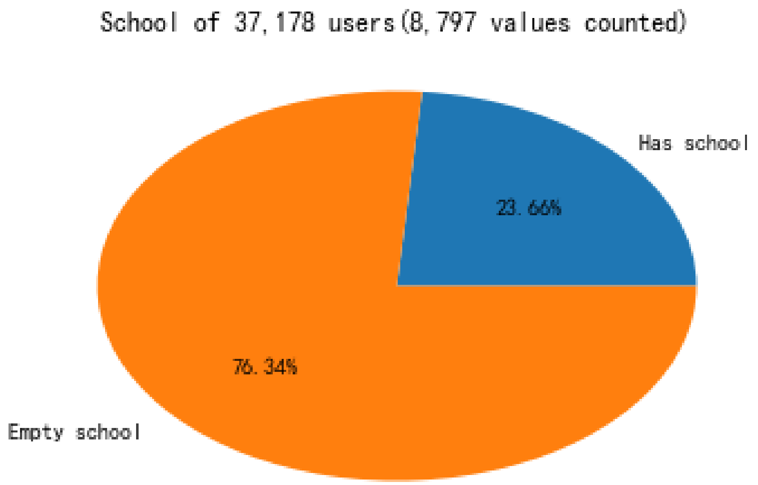
4. Locations:

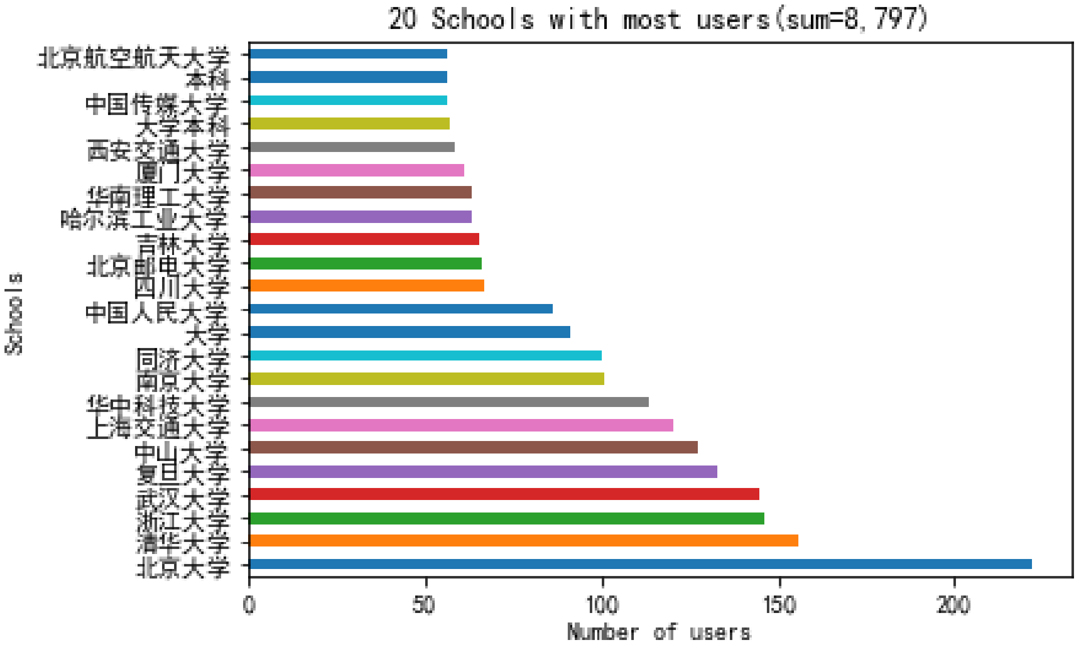




Numbers of users are positively correlated with the levels of cities. No too many overseas users.

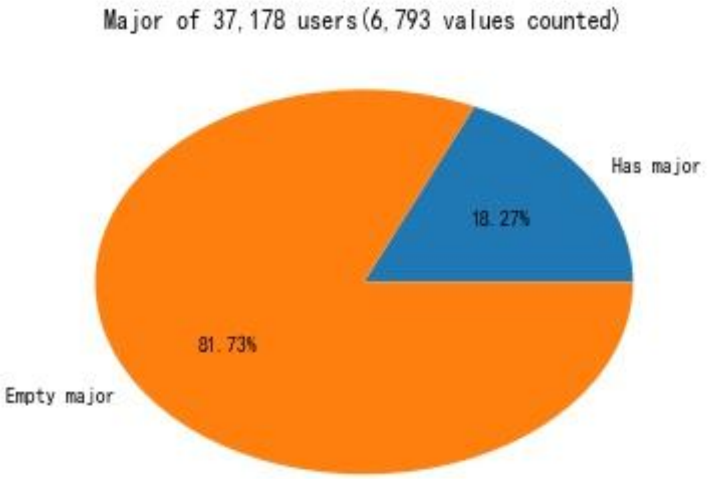
5. Schools

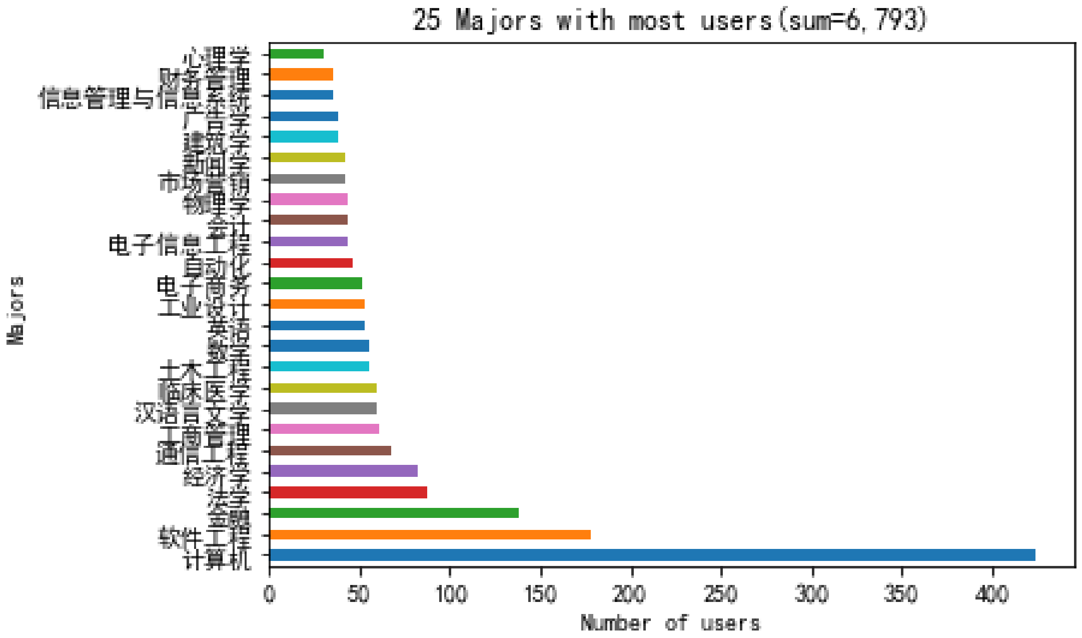




985&211 universities occupy the majority of users.

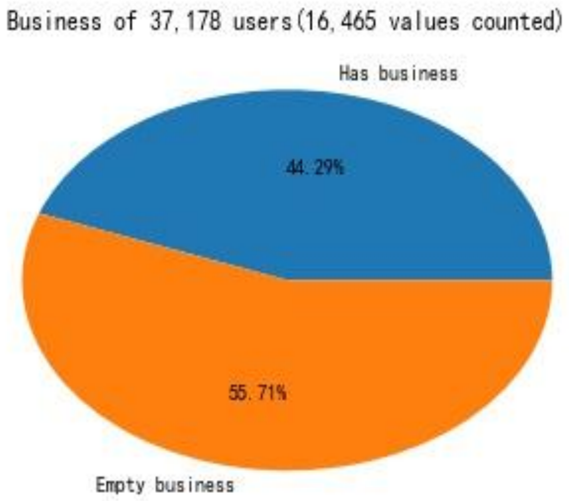
6. Majors

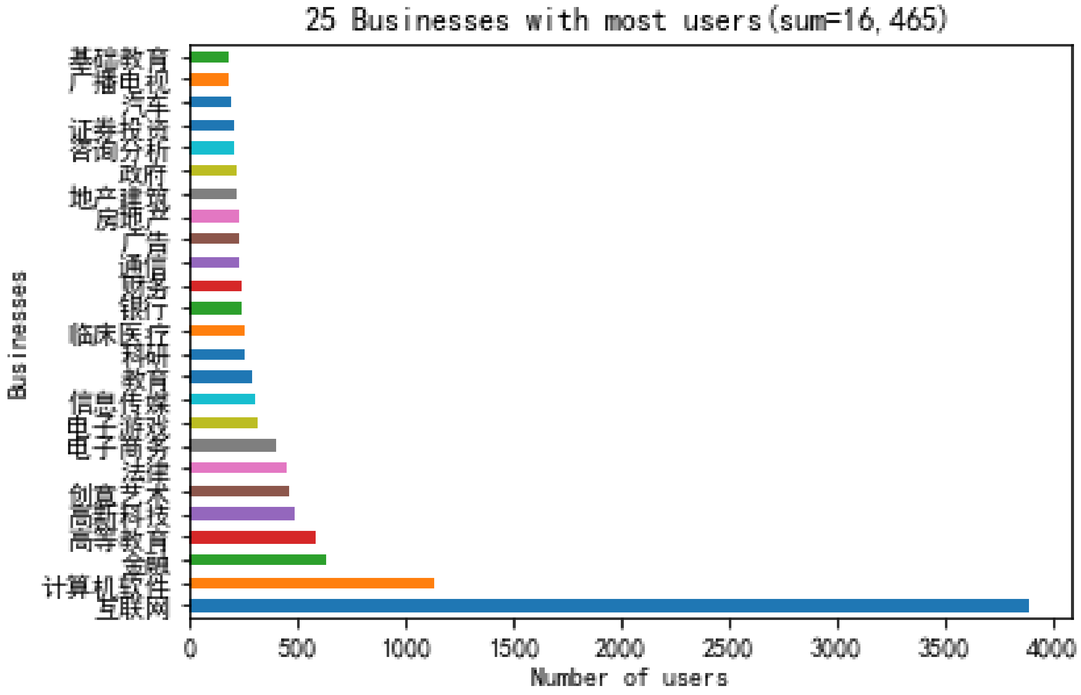




Computer-related majors, such as computer and software engineering occupy the largest propotion of users. Business-related majors like finance, business management, and e-commerce are the second largest part. Other parts vary from social sciences to natural sciences.

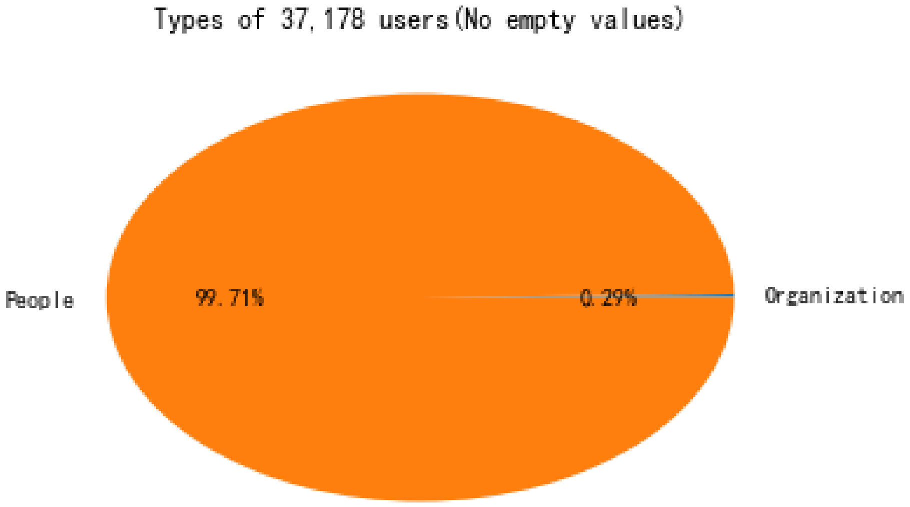
7. Business



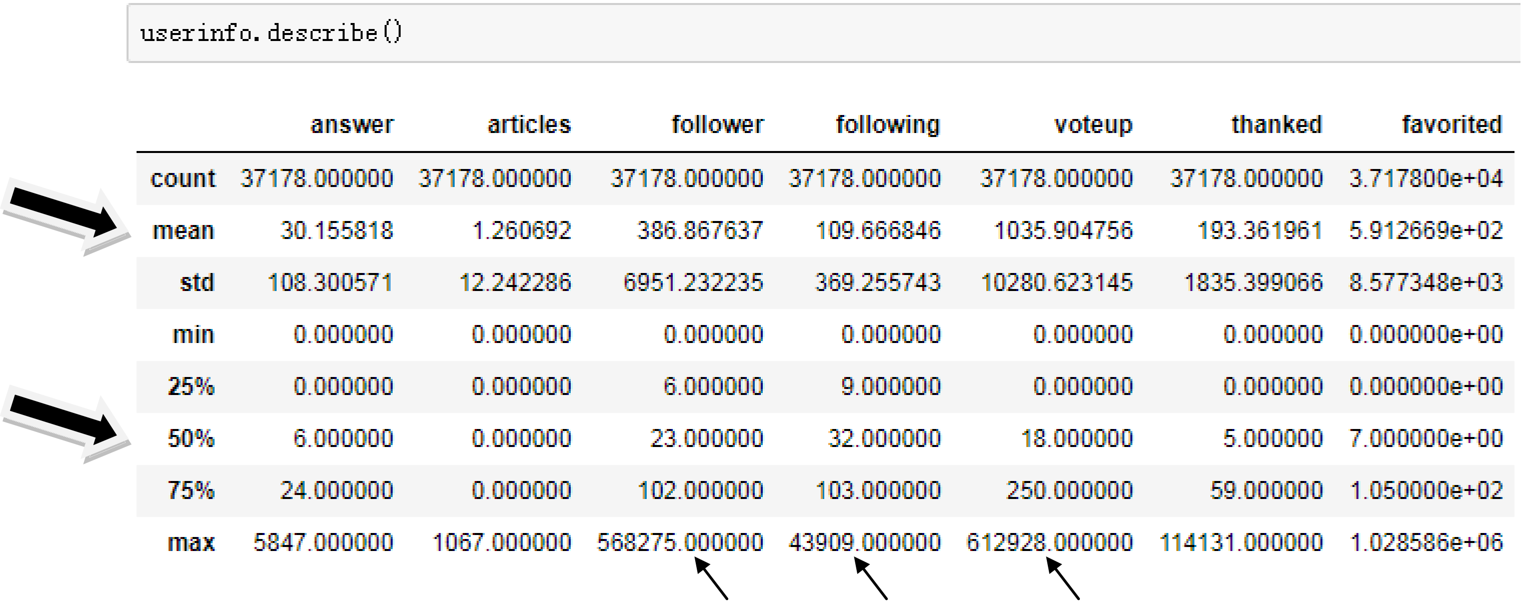


Internet users occupies the largest propotion, around 23.6%. Then follows computer software industry, around 6.9% and financial industry, around 3.9%. Most industries have relatively same proportion.

8. Usertypes:



9. Descriptive statistics of interval data



```
userinfo.describe()
```

	answer	articles	follower	following	voteup	thanked	favorited
count	37178.000000	37178.000000	37178.000000	37178.000000	37178.000000	37178.000000	3.717800e+04
mean	30.155818	1.260692	386.867637	109.666846	1035.904756	193.361961	5.912669e+02
std	108.300571	12.242286	6951.232235	369.255743	10280.623145	1835.399066	8.577348e+03
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00
25%	0.000000	0.000000	6.000000	9.000000	0.000000	0.000000	0.000000e+00
50%	6.000000	0.000000	23.000000	32.000000	18.000000	5.000000	7.000000e+00
75%	24.000000	0.000000	102.000000	103.000000	250.000000	59.000000	1.050000e+02
max	5847.000000	1067.000000	568275.000000	43909.000000	612928.000000	114131.000000	1.028586e+06

Looking at the 50% percentile, users following '三体' are generally not very active. Large means and standard deviations are mainly resulted from a small portion of very active users or Internet celebrities.