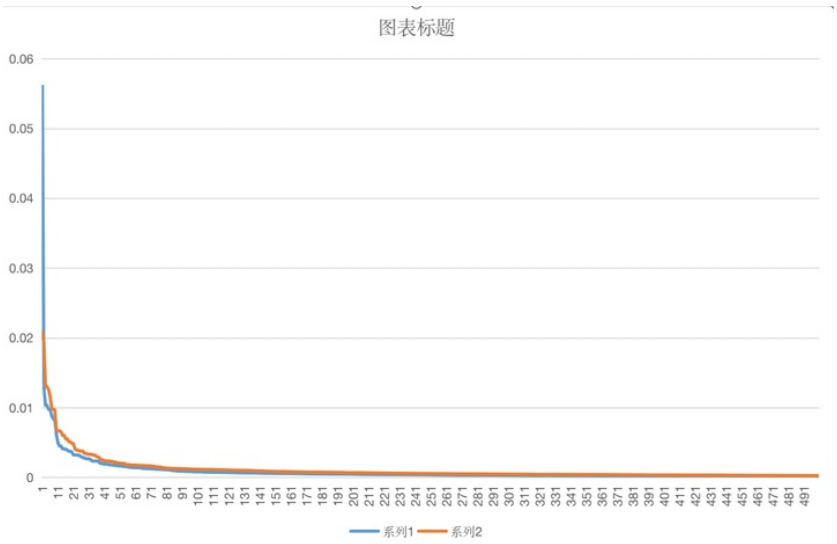
词频出现频率

词频使用频率的曲线图如图所示，蓝色为女性，红色为男性。大于 100 时词频使用区分度不是很高。



关键字使用情况

男性对于关键字的使用较多，女性相对较少。此外从上面曲线图可以看出男性词频较为分散，女性词频使用较为集中。当然这不排除受男性训练数据是从较大规模的男性作者里挑选出来的原因影响

	男	女
数字	7（10 进制）+2（16 进制）	11（10 进制）+1（16 进制）
关键字	30	24
单字符变量	16	17
多字符变量	7	13
布尔变量	2	2

各特征平均值与标准差情况

分别对男女性特征值取了平均值和标准差，实验结果如下：

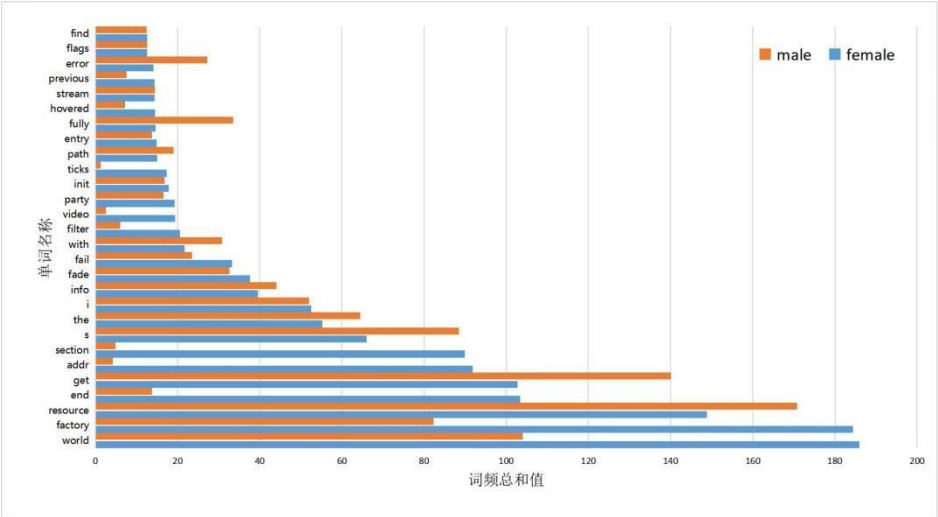
女性使用局部变量较多，驼峰命名法使用相对较多；男性使用三元运算符较多

	男	女
局部变量频率平均值	**0.0912**	**0.12**
局部变量标准差	0.0536	0.0525
关键词频率平均值	0.17956	0.178
关键词频率标准差	0.07756	0.06
函数长度平均值	0.027	0.03
函数长度标准差	0.024	0.025
控制字符平均值	20.105	20.8
控制字符标准差	77.1303	17.756
英语水平平均值	4.6	5.0
英语水平标准差	0.0	0.0
三元运算符平均值	**0.88**	**0.594**
三元运算符标准差	5.7	3.248
空白字符平均值	0.18378	0.1926955
空白字符标准差	0.089	0.10326
OnLineBeforeOpenBranceNumber 平均值	**0.16778**	**0.095**
OnLineBeforeOpenBranceNumber 标准差	4.758	0.081
参数平均值	**0.16778**	**0.09566**
参数标准差	4.758	0.0814
驼峰命名法平均值	**0.00000997**	**0.0000266**
驼峰命名法标准差	0.00040	0.000789
函数个数频率平均值	0.0021915	0.0022899
函数个数频率标准差	0.002315	0.001589
空白行平均值	**0.152317**	**0.14411**
空白行标准差	0.065	0.0677525
注释频率平均值	**28.375**	**25.158**
注释频率标准差	197.39	

女性编程过程中命名规范性较高、布局相对男性规范（驼峰命名法较多、空白行数量较少、大括号较少和前缀代码在一起）。男性更追求编程快捷性（例如三元运算符使用较多，函数参数使用较多、局部变量较少）。这比较符合传统的二元性别的性格特征。

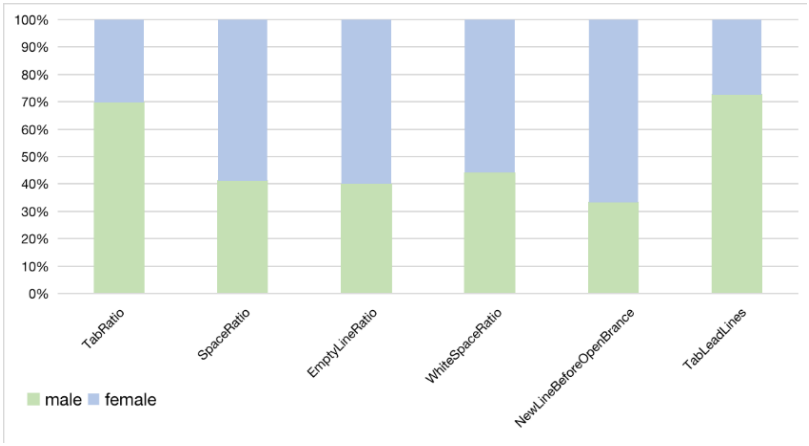
单词使用情况

数据集中男女使用的前 28 个单词，如下图

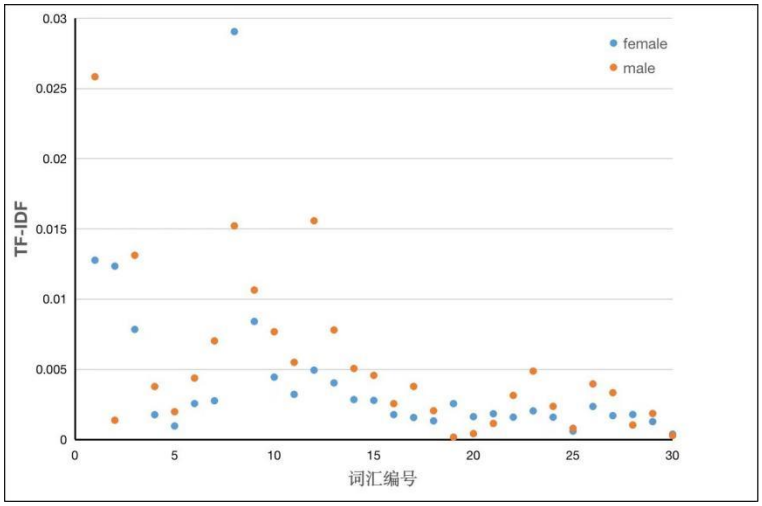


布局特征差异

男女性别在布局特征的表示上确实会存在差异，如下图



TF-IDF 表示结果


0.03
0.025
0.02
0.015
0.01
0.005
0
TF-IDF
0
5
10
15
20
25
30
词汇编号
female
male