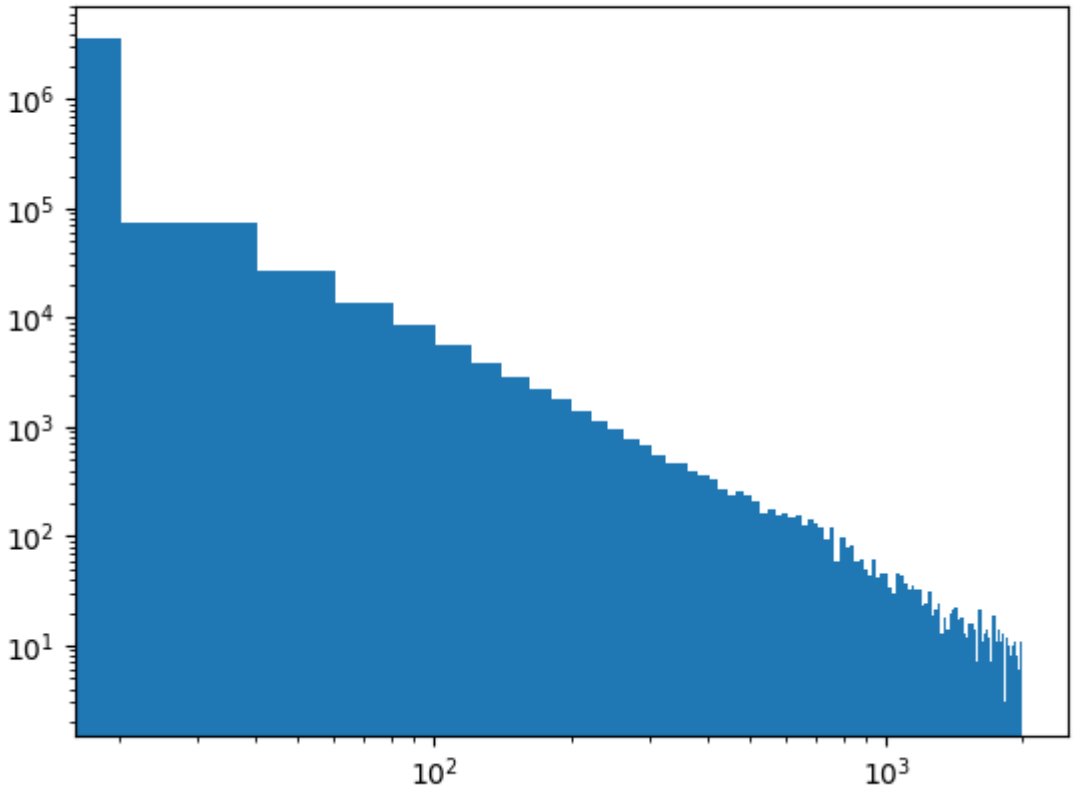
链接结构分析作业

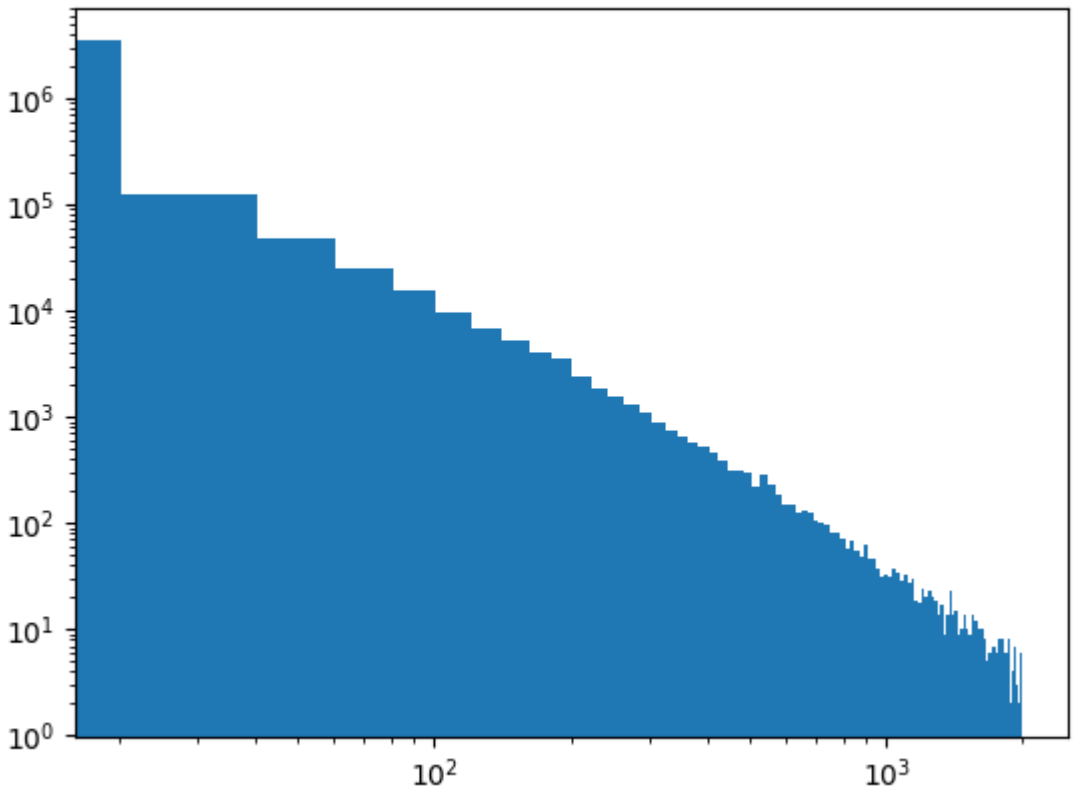
计53 吴昊哲 2015011297

维基百科语料库中出/入链接数分布

入链接分布如下:

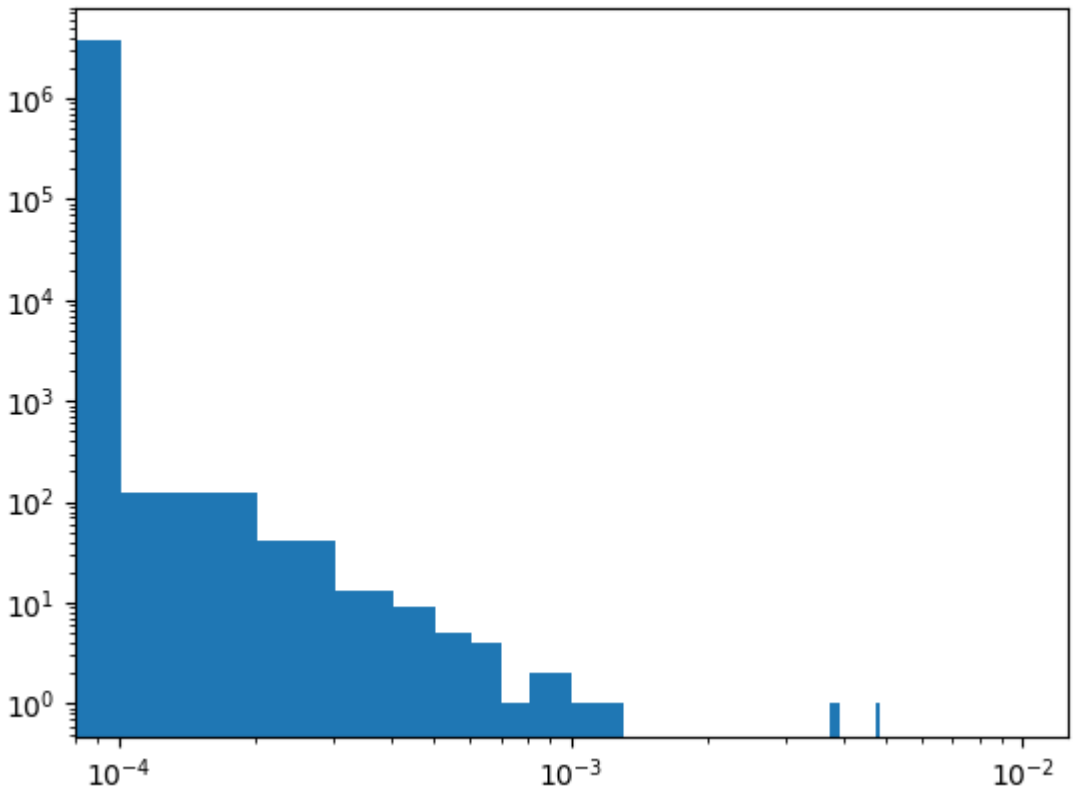


出链接分布如下:



由分布可见大量网页都有着很低的入链接和出链接，有着较高入链接和出链接的网页所占比例极少。

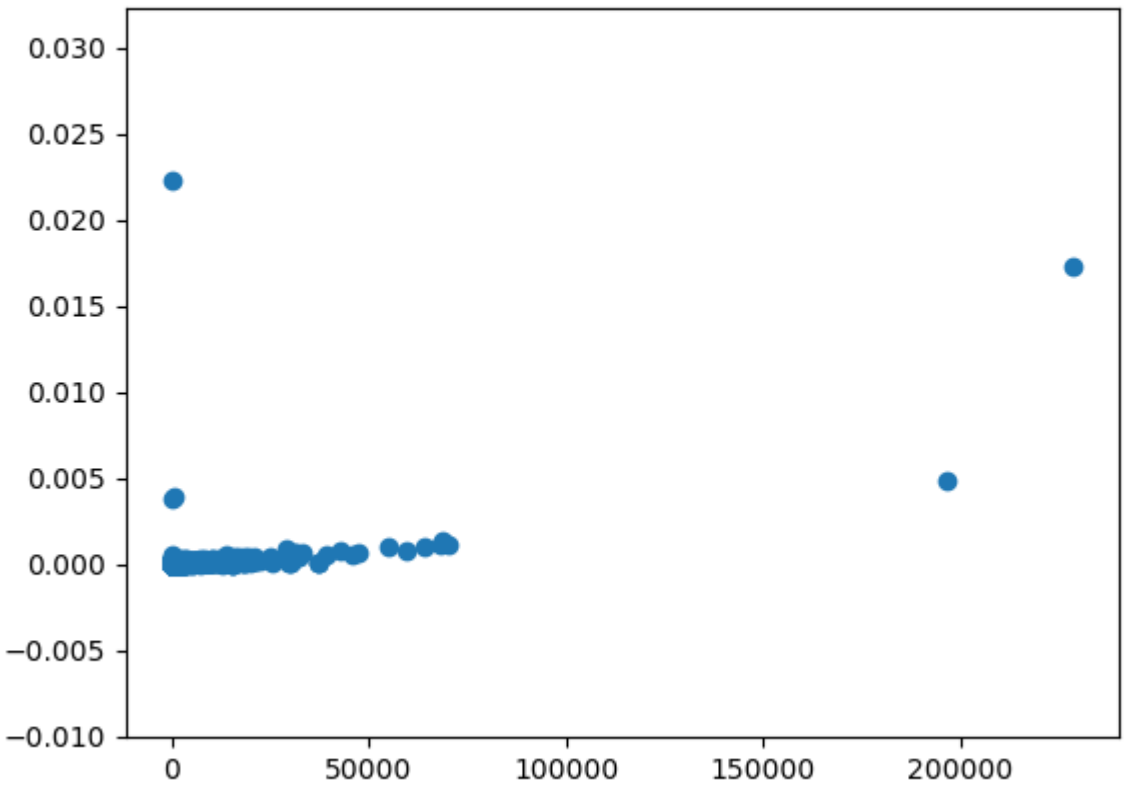
PageRank算法结果分布情况



大部分网页的PageRank评分都很低，只有很少的网页有着较高的PageRank评分。

PageRank与入链接数的关联分析

我们画出了PageRank得分和入链接数的关系图:



可见入链接数和PageRank有着较明显的正相关关系

我们计算了PageRank和入链接数的线性相关系数$r = 0.703$

由Scatter我们可以看出一些Outlier将线性相关系数拉低了。

PageRank得分与相应条目语义内容的分析

我们将最高的最低PageRank得分对应的词条罗列如下:

最高PageRank:

PageRank	词条
0.022	箭头
0.017	←
0.005	维基数据
0.0039	Unicode
0.0038	符号
0.0013	中国
0.0012	美国
0.0011	学名
0.0010	法国
0.00097	市镇

可见最高的PageRank均为名词，有国家（中国、美国、法国），符号（箭头，←，符号，Unicode），以及一些可能会有很多网站指向的词（维基数据）

最低PageRank:

PageRank	词条
1.45191965438e-07	斯科莱布林冰川
1.45191965438e-07	小行星列表/349401-349500
1.45191965438e-07	小行星列表/349501-349600
1.45191965438e-07	小行星列表/349601-349700
1.45191965438e-07	小行星列表/349701-349800
1.45191965438e-07	小行星列表/349801-349900
1.45191965438e-07	小行星列表/349901-350000
1.45191965438e-07	小行星列表/350001-350100
1.45191965438e-07	小行星列表/350101-350200
1.45191965438e-07	小行星列表/349301-349400

最低的PageRank对应的都是一些很生僻的词