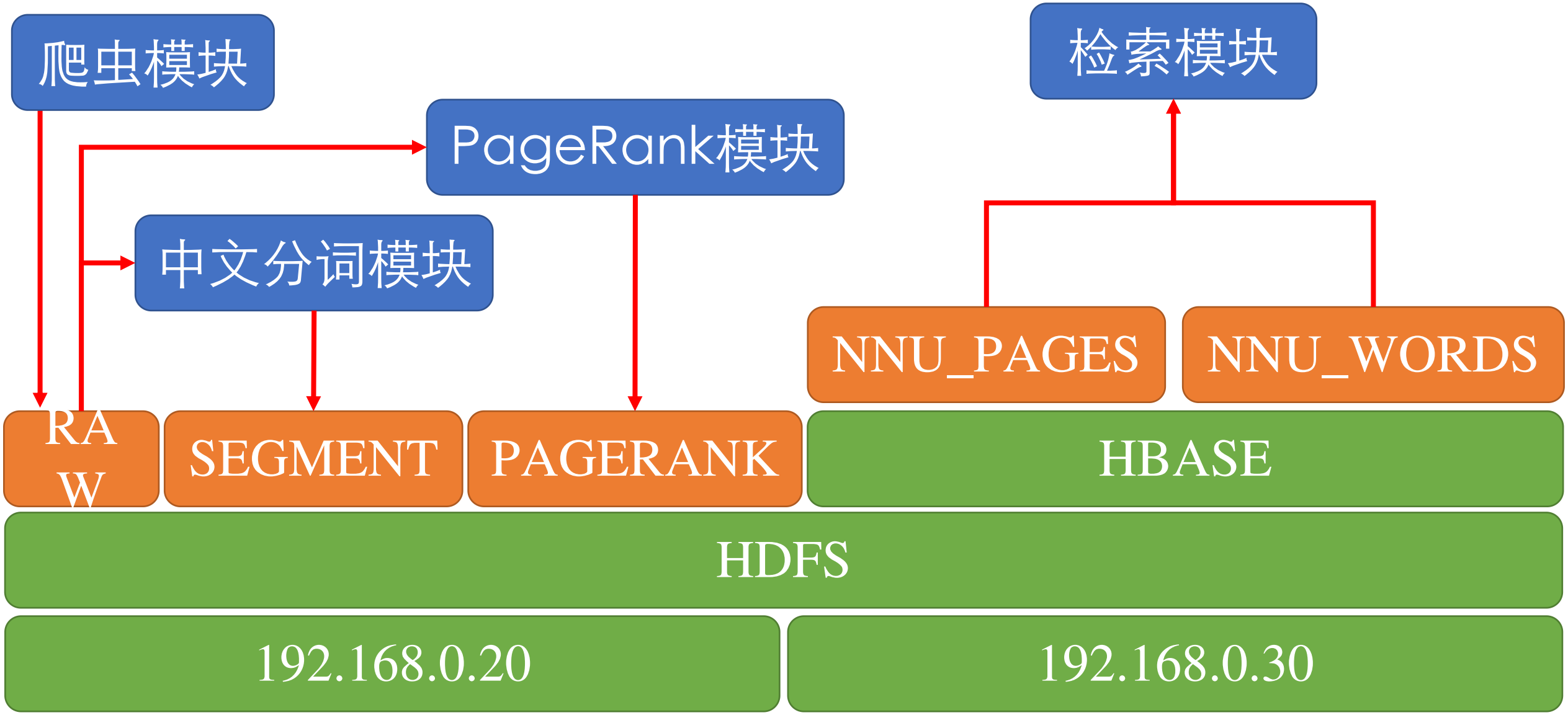
Retrieve NJNU

小型搜索引擎的设计与实现

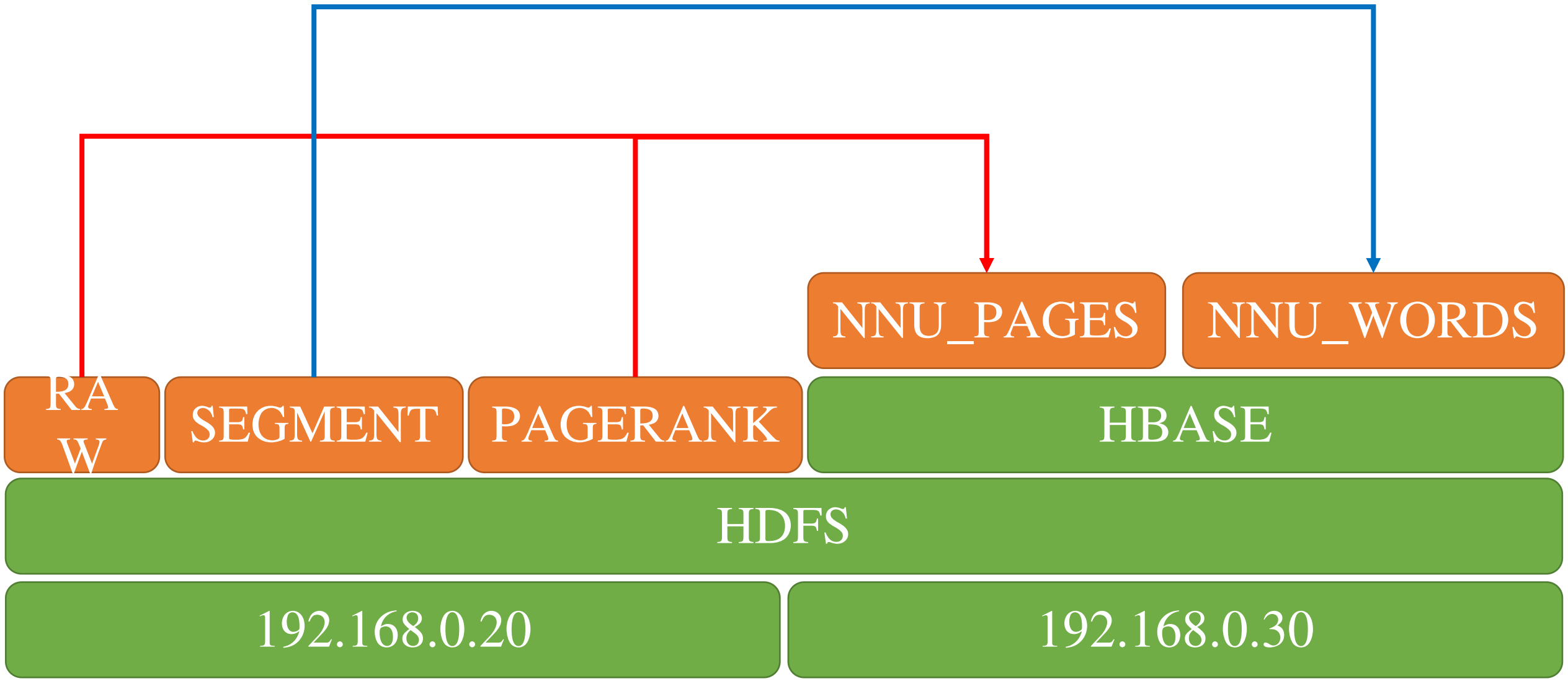
孙振强
南京师范大学
计算机与电子信息学院

目录

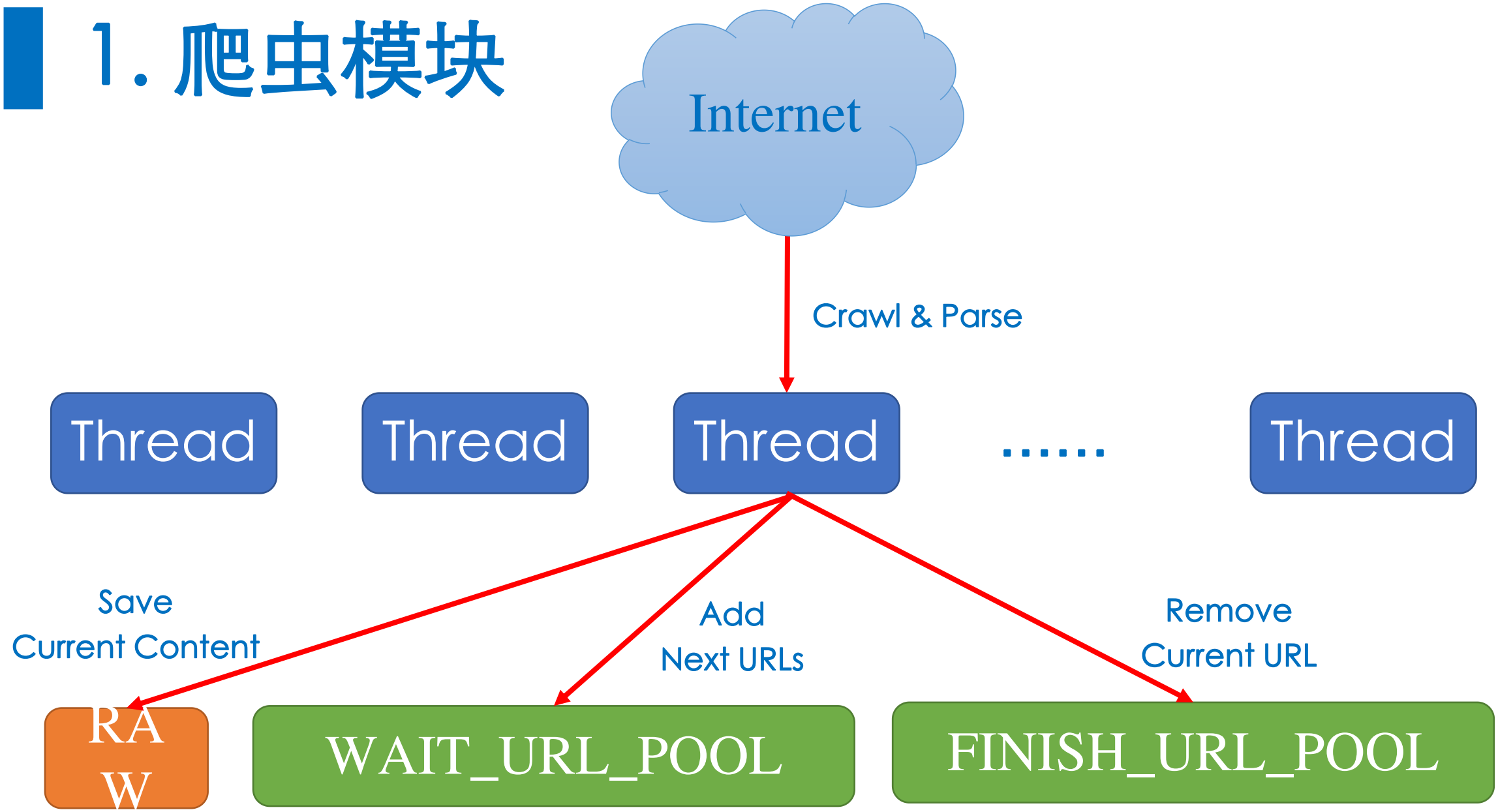
系统架构


爬虫模块
PageRank模块
检索模块
中文分词模块
NNU_PAGES
NNU_WORDS
RA
W
SEGMENT
PAGERANK
HBASE
HDFS
192.168.0.20
192.168.0.30

系统架构


NNU_PAGES
NNU_WORDS
RA
W
SEGMENT
PAGERANK
HBASE
HDFS
192.168.0.20
192.168.0.30

1. 爬虫模块


Internet
Crawl & Parse
Thread
Thread
Thread
.....
Thread
Save
Current Content
Add
Next URLs
Remove
Current URL
RA
W
WAIT_URL_POOL
FINISH_URL_POOL

1. 爬虫模块

- 主站：www.njnu.edu.cn
- 阳光网：news.njnu.edu.cn
- 研究生院：grad.njnu.edu.cn
- 所有28个二级学院：

honors.njnu.edu.cn	jsjyxy.njnu.edu.cn	gjy.njnu.edu.cn	jny.njnu.edu.cn
spa.njnu.edu.cn	sxy.njnu.edu.cn	law.njnu.edu.cn	marx.njnu.edu.cn
jky.njnu.edu.cn	xlxy.njnu.edu.cn	tky.njnu.edu.cn	wxy.njnu.edu.cn
wy.njnu.edu.cn	xinchuan.njnu.edu.cn	sfy.njnu.edu.cn	math.njnu.edu.cn
physics.njnu.edu.cn	hky.njnu.edu.cn	dky.njnu.edu.cn	sky.njnu.edu.cn
energy.njnu.edu.cn	d.njnu.edu.cn	ceai.njnu.edu.cn	env.njnu.edu.cn
hy.njnu.edu.cn	spxy.njnu.edu.cn	music.njnu.edu.cn	msxy.njnu.edu.cn

Num Pages:	198,649
Total Size:	1.2 GB

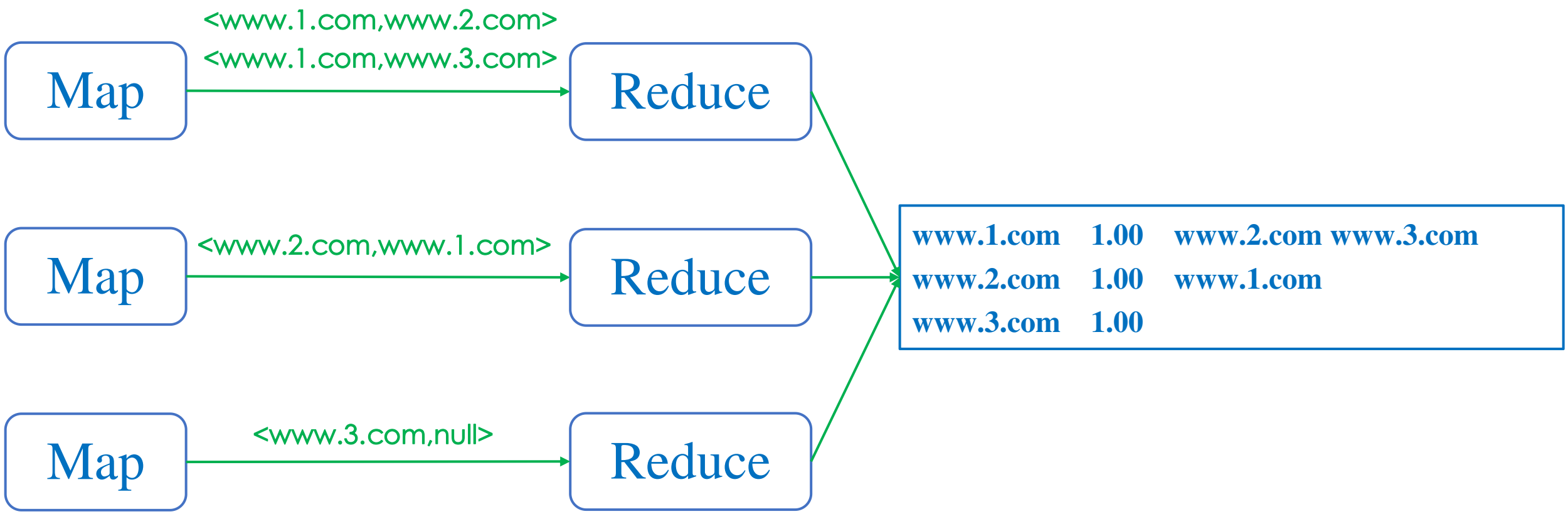
1. 爬虫模块

[0]	[URL] www.1.com	[nextURLs] www.2.com www.3.com	[title] title-1	[body] body-1
[1]	[URL] www.2.com	[nextURLs] www.1.com	[title] title-2	[body] body-2
[2]	[URL] www.3.com	[nextURLs]	[title] title-3	[body] body-3
……	……	……		
[9999]	[URL] www.9999.com	[nextURLs] www.1.com www.8.com	[title] title-9999	[body] body-9999

[0] [URL] http://www.njnu.edu.cn/ [nextURLs] http://www.njnu.edu.cn/index.htm http://www.njnu.edu.cn/info/1171/10561.htm http://www.njnu.edu.cn/info/1113/2784.htm http://www.njnu.edu.cn/xysz.htm http://www.njnu.edu.cn/xxgk.htm http://www.njnu.edu.cn/xysz/rmzd.htm http://www.njnu.edu.cn/xxgk.htm http://www.njnu.edu.cn/xxgk/xxjj.htm http://www.njnu.edu.cn/xxgk/xxzc.htm http://www.njnu.edu.cn/xxgk/bnxs.htm http://www.njnu.edu.cn/xxgk/ysfc.htm http://www.njnu.edu.cn/xxgk/xrld.htm http://www.njnu.edu.cn/xxgk.htm http://www.njnu.edu.cn/xxgk/xymj.htm http://www.njnu.edu.cn/xxgk/xbxg.htm http://www.njnu.edu.cn/xysz.htm http://www.njnu.edu.cn/jyjx.htm http://grad.njnu.edu.cn/ http://jsjyxy.njnu.edu.cn/ http://www.njnu.edu.cn/kxyj.htm http://www.njnu.edu.cn/kxyj/xsyg.htm http://www.njnu.edu.cn/kxyj/xsbd.htm http://www.njnu.edu.cn/kxyj/yjjg.htm http://www.njnu.edu.cn/hzjl.htm http://www.njnu.edu.cn/zsjy.htm http://www.njnu.edu.cn/xysh.htm http://www.njnu.edu.cn/xysh.htm http://www.njnu.edu.cn/xysh.htm http://www.njnu.edu.cn/xysh.htm http://www.njnu.edu.cn/xysh.htm http://www.njnu.edu.cn/xysh/xsst1.htm http://www.njnu.edu.cn/xysh.htm http://www.njnu.edu.cn/xysh.htm http://news.njnu.edu.cn/info/1032/94304.htm http://news.njnu.edu.cn/info/1107/94261.htm http://news.njnu.edu.cn/info/1107/94047.htm http://news.njnu.edu.cn/info/1107/94023.htm http://news.njnu.edu.cn/info/1107/93928.htm http://www.njnu.edu.cn/index/tpyw.htm http://www.njnu.edu.cn/index/nsxw.htm http://news.njnu.edu.cn/info/1107/94261.htm http://news.njnu.edu.cn/info/1107/94210.htm http://news.njnu.edu.cn/info/1107/94047.htm http://news.njnu.edu.cn/info/1107/94015.htm http://news.njnu.edu.cn/info/1044/93907.htm http://news.njnu.edu.cn/info/1107/94312.htm http://news.njnu.edu.cn/info/1044/94308.htm http://news.njnu.edu.cn/info/1044/94305.htm http://news.njnu.edu.cn/info/1107/94303.htm http://news.njnu.edu.cn/info/1107/94302.htm http://news.njnu.edu.cn/info/1044/94283.htm http://news.njnu.edu.cn/info/1107/94274.htm http://news.njnu.edu.cn http://www.njnu.edu.cn/index/tzgg.htm http://www.njnu.edu.cn/info/1098/15185.htm http://www.njnu.edu.cn/info/1098/15596.htm http://www.njnu.edu.cn/info/1098/15574.htm http://www.njnu.edu.cn/info/1098/15569.htm http://www.njnu.edu.cn/info/1098/15556.htm http://www.njnu.edu.cn/info/1098/15553.htm http://www.njnu.edu.cn/info/1098/15552.htm http://www.njnu.edu.cn/kxyj/xsyg.htm http://www.njnu.edu.cn/info/1037/15614.htm http://www.njnu.edu.cn/info/1037/15613.htm http://www.njnu.edu.cn/info/1037/15612.htm http://www.njnu.edu.cn/info/1037/15594.htm http://www.njnu.edu.cn/info/1037/15573.htm http://www.njnu.edu.cn/kxyj/xsbd.htm http://www.njnu.edu.cn/info/1038/15608.htm http://www.njnu.edu.cn/info/1038/15578.htm http://www.njnu.edu.cn/info/1038/15576.htm http://www.njnu.edu.cn/info/1038/15557.htm http://www.njnu.edu.cn/index/fjns/rcpy.htm http://www.njnu.edu.cn/index/fjns/kxyj.htm http://www.njnu.edu.cn/index/fjns/shfw.htm http://www.njnu.edu.cn/index/fjns/whccycx.htm http://www.njnu.edu.cn/index/fjns/gjjlyhz.htm http://www.njnu.edu.cn/index/fjns/sxkyjz.htm http://www.njnu.edu.cn/index/fjns/rcpy.htm http://www.njnu.edu.cn/info/1087/15595.htm http://www.njnu.edu.cn/info/1088/15519.htm http://www.njnu.edu.cn/info/1088/15462.htm http://www.njnu.edu.cn/index/rwgs/bdgs.htm http://www.njnu.edu.cn/index/rwgs/mzyx.htm http://www.njnu.edu.cn/index/rwgs/sdrw.htm http://news.njnu.edu.cn/rwgs.htm http://news.njnu.edu.cn/info/1044/93753.htm http://news.njnu.edu.cn/info/1012/92799.htm http://news.njnu.edu.cn/info/1012/92607.htm http://news.njnu.edu.cn/info/1012/92605.htm http://news.njnu.edu.cn/info/1012/92237.htm http://news.njnu.edu.cn/info/1012/92235.htm http://news.njnu.edu.cn/info/1012/92228.htm http://news.njnu.edu.cn/mtns/mtns.htm http://news.njnu.edu.cn/info/1025/94268.htm http://news.njnu.edu.cn/info/1025/94270.htm http://news.njnu.edu.cn/info/1025/94263.htm http://news.njnu.edu.cn/info/1025/94268.htm http://news.njnu.edu.cn/info/1025/94272.htm http://news.njnu.edu.cn/info/1025/94271.htm http://news.njnu.edu.cn/info/1025/94270.htm http://www.njnu.edu.cn/xxgk/xymj.htm http://www.njnu.edu.cn/info/1162/14241.htm http://www.njnu.edu.cn/info/1162/11794.htm http://www.njnu.edu.cn/info/1162/11717.htm http://www.njnu.edu.cn/dsxx/index.htm http://www.njnu.edu.cn/index/bwcx_ljsm.htm http://honors.njnu.edu.cn/ http://jsjyxy.njnu.edu.cn/ http://gjy.njnu.edu.cn http://jny.njnu.edu.cn/ http://spa.njnu.edu.cn/ http://sxy.njnu.edu.cn/ http://law.njnu.edu.cn http://marx.njnu.edu.cn/ http://jky.njnu.edu.cn/ http://xlxy.njnu.edu.cn/ http://tky.njnu.edu.cn/ http://wxy.njnu.edu.cn/ http://wy.njnu.edu.cn/ http://xinchuan.njnu.edu.cn/ http://sfy.njnu.edu.cn/ http://math.njnu.edu.cn/ http://physics.njnu.edu.cn/ http://hky.njnu.edu.cn/ http://dky.njnu.edu.cn/ http://sky.njnu.edu.cn/index.asp http://energy.njnu.edu.cn http://d.njnu.edu.cn/ http://ceai.njnu.edu.cn http://env.njnu.edu.cn/ http://hy.njnu.edu.cn/ http://spxy.njnu.edu.cn http://music.njnu.edu.cn/ http://msxy.njnu.edu.cn/ http://www.njnu.edu.cn/index/bcskb.htm http://www.njnu.edu.cn/index/fcxx.htm http://www.njnu.edu.cn/index/xqpmyjtt.htm http://www.njnu.edu.cn/index.htm [title] 南京师范大学 [body] 南京师范大学 English邮件在线 English 书记信箱 校长信箱 学院网站 部门网站 热门站点 图书馆 | 邮件在线 学校概况 学校简介 学校章程 百年校史 院士风采 现任领导 内设机构 校园美景 校标校歌 学院设置 教育教学 师资队伍 本科教学 本科生教育 研究生教育 教师特色教育 留学生教育 继续教育 网络教育 科学研究 学术预告 学术报道 自然科学 社会科学 研究机构 南京师大学报 合作交流 基金会董事会理事会 校友会 联合办学与社会服务 国际交流 南京法语培训中心 招生就业 本科生招生 研究生招生 留学生招生 继续教育招生 自考招生 就业创业 校园生活 校园资源 校园概览 健康与体育 艺术与文化 学生社团 教工活动 平安校园 事务中心 2021毕业季 | 青春光影-镜头记录此刻模样，期待前方光芒万丈 南师大举行庆祝中国共产党成立100周年师生歌咏比赛 江苏省“关爱青少年身心健康”系列宣讲会暨“5·25”大学生心理健康教育活动推进周启动... 著名雕塑家、中国美术馆馆长吴为山教授回母校作专题报告 我校社会科学总论首次进入ESI全球前1% 2021毕业季 | 青春... 南师大举行庆祝中... 江苏省“关爱青少... 著名雕塑家、中国... 我校社会科学总论... 更多 南师新闻 南师大举行庆祝中国共产党成立100周年师生歌咏比赛 06月10日 我校艺术教育中心庆祝建党百年专场演出走进南京财经大学 06月07日 江苏省“关爱青少年身心健康”系列宣讲会暨“5·25”大学生心理健康教育活动推进周启动仪式在南师大举行 05月26日 南师大文学院开展“与雨花英烈隔空对话”系列活动 05月23日 南师大新闻与传播学院党史学习教育邂逅中国戏剧“梅花奖” 05月16日 06月16日 江苏省高校资产与采购工作业务培训会（南京班）在我校举行 06月15日 南京师范大学师生走进军中党校共话百年党史 06月15日 我校举办“百年大党与21世纪马克思主义理论创新”高端学术论坛 06月15日 中宣部理论局致信感谢我校 06月15日 我校第三十四次学生代表大会、第二十一次研究生代表大会召开 06月13日 我校召开国家社科基金重大项目开题论证会 06月11日 我校获大学生心理健康教育工作多项荣誉称号 更多新闻请访问阳光网 通知公告 校办 04月14日 声 明 近日，网络上出现以南师大名义进行学历教育招生收费的网站、抖音或各类宣传，其均为虚假内容。我校依法并在江苏省教育厅的要求与指导下，在学校官网www.njnu.edu.cn发布学历教育相关招生信息。任何涉及到我校学生学历教育培养的招生活动，必是经过江苏省教育厅批准后进行的。请谨慎选择，防止被骗。特此声明。南京师范大学2021年4月12日下面为网友举报的假冒南师

2. PageRank模块 Step 1: Build Graph

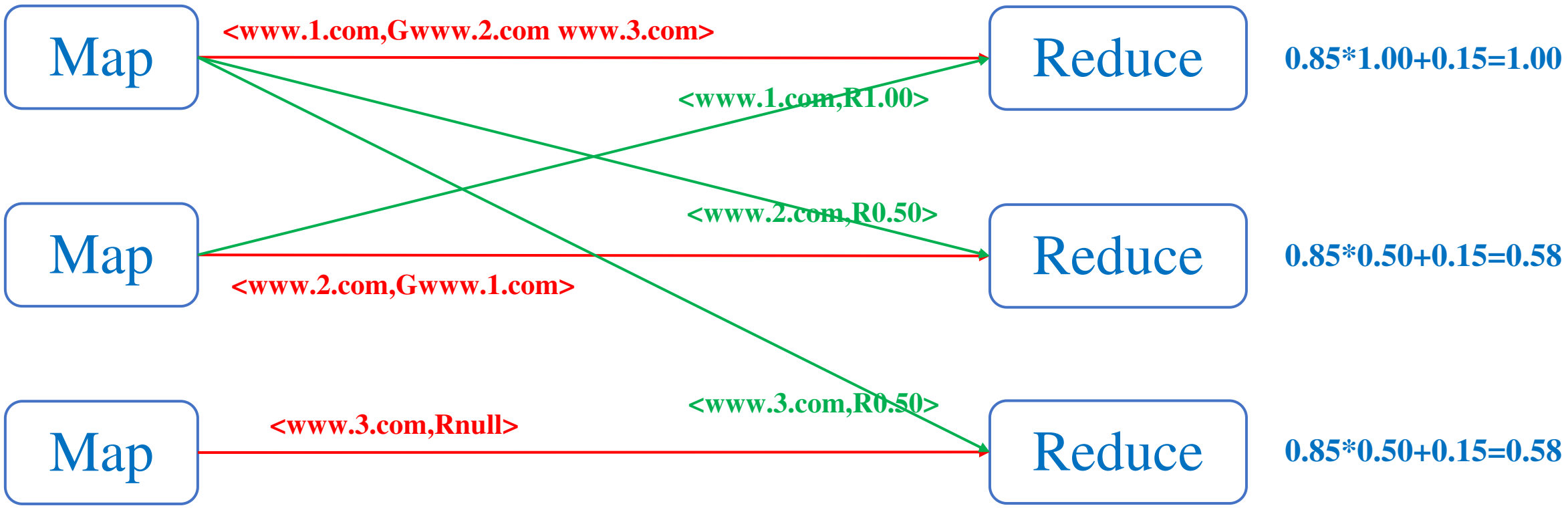
[0]	[URL] www.1.com	[nextURLs] www.2.com www.3.com	……
[1]	[URL] www.2.com	[nextURLs] www.1.com	……
[2]	[URL] www.3.com	[nextURLs]	……



2. PageRank模块 Step 2: Calculate

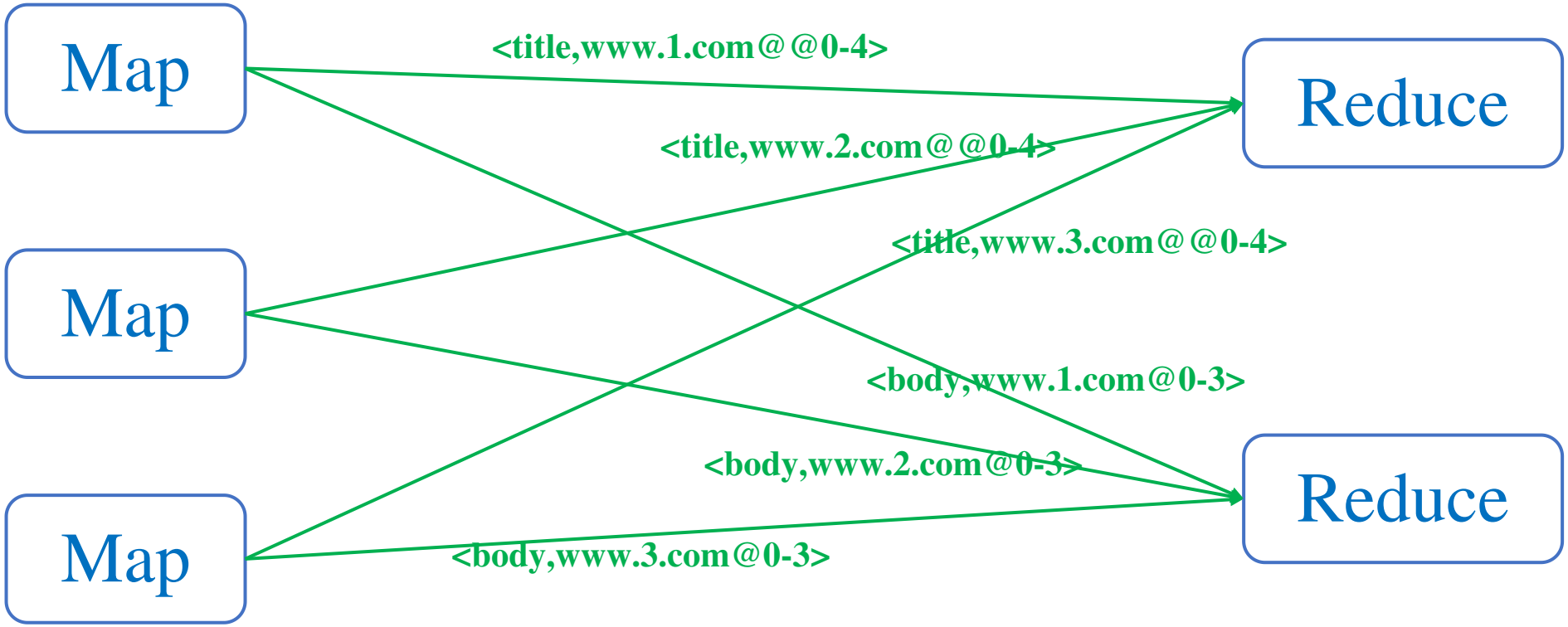
www.1.com	1.00	www.2.com www.3.com
www.2.com	1.00	www.1.com
www.3.com	1.00	

www.1.com	1.00	www.2.com www.3.com
www.2.com	0.58	www.1.com
www.3.com	0.58	



3. 中文分词模块

[0] [URL] www.1.com …… [title] title-1 [body] body-1
[1] [URL] www.2.com …… [title] title-2 [body] body-2
[2] [URL] www.3.com …… [title] title-3 [body] body-3



title www.1.com@@0-4 www.2.com@@0-4 www.3.com@@0-4
body www.1.com@0-3 www.2.com@0-3 www.1.com@0-3

结巴分词(java版)　https://github.com/huaban/jieba-analysis

README.md

结巴分词(java版) jieba-analysis

首先感谢jieba分词原作者fxsjy，没有他的无私贡献，我们也不会结识到结巴 分词. 同时也感谢jieba分词java版本的实现团队huaban，他们的努力使得Java也能直接做出效果很棒的分词。

不过由于huaban已经没有再对java版进行维护，所以我自己对项目进行了开发。除了结巴分词(java版)所保留的原项目针对搜索引擎分词的功能(cut~~forindex~~、cut~~forsearch~~)，我加入了tfidf的关键词提取功能，并且实现的效果和python的jieba版本的效果一模一样！

（以下内容在基于jieba-java版本README.md的基础上，加入了对我新加入的tfidf关键词提取模块的相关说明）

简介

支持分词模式

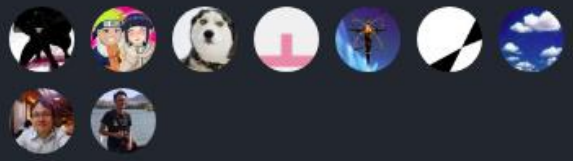
- Search模式，用于对用户查询词分词
- Index模式，用于对索引文档分词

特性

- 支持多种分词模式

Used by 771

Contributors 9



Languages

Java 99.9%　Shell 0.1%

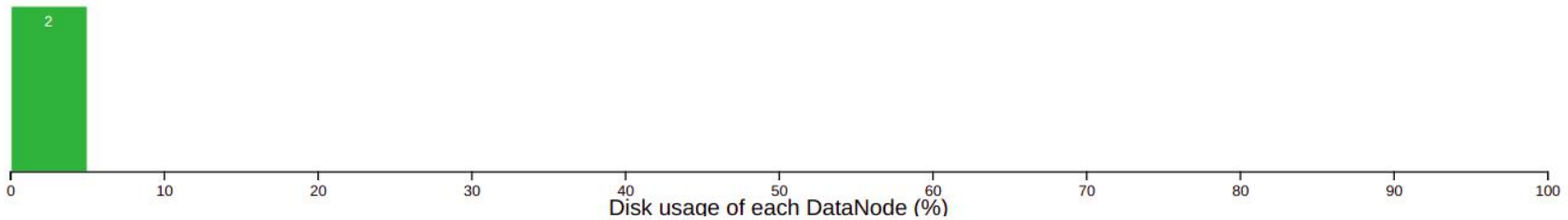
一两件 http://news.njnu.edu.cn/info/1046/83435.htm@885-888 http://news.njnu.edu.cn/info/1016/37681.htm@905-908 http://gjy.njnu.edu.cn/xyxwnry.jsp?urltype=news.NewsContentUrl&wbtreeid=1023&wbnewsid=1337@1240-1243 http://news.njnu.edu.cn/info/1044/31521.htm@508-511 http://tky.njnu.edu.cn/info/1022/3457.htm@1133-1136 http://news.njnu.edu.cn/info/1046/47572.htm@427-430 http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=9480@978-981 http://news.njnu.edu.cn/info/1045/79984.htm@943-946 http://news.njnu.edu.cn/info/1045/16096.htm@780-783 http://jsjyxy.njnu.edu.cn/info/1032/3857.htm@1290-1293
一两倍 http://jny.njnu.edu.cn/info/1078/1494.htm@5207-5210
一两分 http://xinchuan.njnu.edu.cn/info/1016/2522.htm@688-691 http://news.njnu.edu.cn/info/1046/55396.htm@403-406 http://news.njnu.edu.cn/info/1045/12562.htm@1451-1454 http://news.njnu.edu.cn/info/1038/4465.htm@702-705 http://news.njnu.edu.cn/info/1046/44282.htm@559-562 http://news.njnu.edu.cn/info/1045/23516.htm@1045-1048 http://sfy.njnu.edu.cn/info/1016/5746.htm@866-869 http://xlxy.njnu.edu.cn/info/1100/6378.htm@1829-1832 http://news.njnu.edu.cn/info/1046/58997.htm@641-644 http://news.njnu.edu.cn/info/1045/18111.htm@368-371 http://news.njnu.edu.cn/info/1046/40451.htm@513-516 http://news.njnu.edu.cn/info/1046/47339.htm@355-358 http://news.njnu.edu.cn/info/1046/47326.htm@333-336 http://xlxy.njnu.edu.cn/info/1084/4124.htm@1782-1785
一两分钟 http://xlxy.njnu.edu.cn/info/1084/4124.htm@1782-1786 http://xlxy.njnu.edu.cn/info/1100/6378.htm@1829-1833 http://news.njnu.edu.cn/info/1045/18111.htm@368-372 http://news.njnu.edu.cn/info/1046/58997.htm@641-645
一两千 http://news.njnu.edu.cn/info/1025/70030.htm@1846-1849 http://sfy.njnu.edu.cn/info/1027/6856.htm@1597-1600
一两千元 http://sfy.njnu.edu.cn/info/1027/6856.htm@1597-1601
一两口 http://sfy.njnu.edu.cn/info/1021/7833.htm@4340-4343
一两句 http://jsjyxy.njnu.edu.cn/info/1039/3458.htm@5219-5222 http://news.njnu.edu.cn/info/1046/56138.htm@398-401 http://sfy.njnu.edu.cn/info/1016/1791.htm@800-803 http://sfy.njnu.edu.cn/info/1027/4575.htm@800-803 http://news.njnu.edu.cn/info/1044/31258.htm@848-851 http://news.njnu.edu.cn/info/1046/43301.htm@843-846 http://news.njnu.edu.cn/info/1045/81798.htm@531-534 http://news.njnu.edu.cn/info/1046/49256.htm@522-525 http://news.njnu.edu.cn/info/1046/41129.htm@512-515 http://dky.njnu.edu.cn/info/1101/3633.htm@2720-2723 http://sfy.njnu.edu.cn/info/1016/3936.htm@924-927 http://news.njnu.edu.cn/info/1046/42410.htm@714-717 http://news.njnu.edu.cn/info/1046/51175.htm@668-671 http://news.njnu.edu.cn/info/1046/54036.htm@437-440 http://news.njnu.edu.cn/info/1045/74523.htm@369-372 http://news.njnu.edu.cn/info/1046/89175.htm@2781-2784 http://news.njnu.edu.cn/info/1036/81020.htm@707-710 http://news.njnu.edu.cn/info/1021/60244.htm@405-408 http://jsjyxy.njnu.edu.cn/info/1040/5689.htm@1076-1079 http://news.njnu.edu.cn/info/1025/66831.htm@509-512 http://news.njnu.edu.cn/info/1025/66836.htm@532-535 http://news.njnu.edu.cn/info/1045/74139.htm@413-416 http://news.njnu.edu.cn/info/1035/5600.htm@707-710 http://jsjyxy.njnu.edu.cn/info/1032/2407.htm@1025-1028 http://news.njnu.edu.cn/info/1042/38576.htm@1466-1469 http://sxy.njnu.edu.cn/info/1206/2245.htm@1113-1116 http://sxy.njnu.edu.cn/info/1206/2250.htm@951-954 http://news.njnu.edu.cn/info/1046/57686.htm@812-815 http://news.njnu.edu.cn/info/1046/46149.htm@677-680 http://news.njnu.edu.cn/info/1045/15989.htm@607-610 http://jsjyxy.njnu.edu.cn/info/1032/3193.htm@1038-1041 http://news.njnu.edu.cn/info/1039/3415.htm@505-508 http://jsjyxy.njnu.edu.cn/info/1031/1266.htm@1037-1040
一两句话 http://sfy.njnu.edu.cn/info/1027/4575.htm@800-804 http://sfy.njnu.edu.cn/info/1016/1791.htm@800-804 http://news.njnu.edu.cn/info/1046/56138.htm@398-402 http://news.njnu.edu.cn/info/1045/74523.htm@369-373 http://sfy.njnu.edu.cn/info/1016/3936.htm@924-928 http://jsjyxy.njnu.edu.cn/info/1039/3458.htm@5219-5223 http://news.njnu.edu.cn/info/1045/74139.htm@413-417
一两名 http://news.njnu.edu.cn/info/1046/44247.htm@768-771 http://news.njnu.edu.cn/info/1045/7709.htm@488-491 http://grad.njnu.edu.cn/info/1031/1569.htm@882-885 http://news.njnu.edu.cn/info/1045/12248.htm@348-351
一两周 http://jsjyxy.njnu.edu.cn/info/1040/5599.htm@1141-1144 http://news.njnu.edu.cn/info/1025/68076.htm@422-425 http://news.njnu.edu.cn/info/1025/68097.htm@424-427 http://news.njnu.edu.cn/info/1045/20620.htm@438-441 http://jsjyxy.njnu.edu.cn/info/1034/7117.htm@920-923 http://news.njnu.edu.cn/info/1046/54510.htm@567-570 http://jsjyxy.njnu.edu.cn/info/1034/6719.htm@671-674 http://jsjyxy.njnu.edu.cn/info/1034/6719.htm@1191-1194 http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=9489@1035-1038 http://news.njnu.edu.cn/info/1035/5611.htm@772-775 http://news.njnu.edu.cn/info/1036/81026.htm@772-775 http://physics.njnu.edu.cn/info/1146/4582.htm@1436-1439 http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=10392@1006-1009 http://jsjyxy.njnu.edu.cn/bksgz/xszz/54.htm@3669-3672 http://jsjyxy.njnu.edu.cn/bksgz/xszz/53.htm@754-757 http://physics.njnu.edu.cn/info/1130/2613.htm@1355-1358
一两场 http://news.njnu.edu.cn/info/1025/69337.htm@933-936 http://news.njnu.edu.cn/info/1045/27077.htm@569-572 http://news.njnu.edu.cn/info/1045/9715.htm@1844-1847
一两块 http://news.njnu.edu.cn/info/1025/67638.htm@1245-1248
一两声 http://news.njnu.edu.cn/info/1046/47768.htm@769-772 http://sfy.njnu.edu.cn/info/1016/2020.htm@2685-2688 http://news.njnu.edu.cn/info/1045/77553.htm@2287-2290 http://jsjyxy.njnu.edu.cn/info/1037/5862.htm@1132-1135
一两天 http://news.njnu.edu.cn/info/1025/93027.htm@821-824 http://sxy.njnu.edu.cn/info/1194/1835.htm@770-773 http://news.njnu.edu.cn/info/1025/68093.htm@574-577 http://news.njnu.edu.cn/info/1025/68109.htm@567-570 http://news.njnu.edu.cn/info/1025/68705.htm@1824-1827 http://news.njnu.edu.cn/info/1025/68221.htm@894-897 http://news.
@@@

http://192.168.0.20:9870/dfshealth.html#tab-datanode

Datanode Information

In service | Down | Decommissioning | Decommissioned | Decommissioned & dead
Entering Maintenance | In Maintenance | In Maintenance & dead

Datanode usage histogram



In operation

Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
desktop-brown:9866 (192.168.0.20:9866)	http://desktop-brown:9864	2s	211m	101.17 GB	201	2.12 GB (2.09%)	3.2.2
slaker:9866 (192.168.0.30:9866)	http://slaker:9864	1s	185m	915.4 GB	138	2.12 GB (0.23%)	3.2.2

Showing 1 to 2 of 2 entries Previous 1 Next

http://192.168.0.20:8088/cluster

All Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	F
4	0	0	4	0	<memory:0, vCores:0>	<memory:8192, vCores:8>	<mem

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nod
1	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocatior
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocat GPUs
application_1623904995757_0004	root	Segment	MAPREDUCE	default	0	Thu Jun 17 17:53:09 +0800 2021	Thu Jun 17 17:53:10 +0800 2021	Thu Jun 17 18:09:20 +0800 2021	FINISHED	SUCCESS	N/A	N/A	N/A	N/A
application_1623904995757_0003	root	Segment	MAPREDUCE	default	0	Thu Jun 17 17:23:13 +0800 2021	Thu Jun 17 17:23:13 +0800 2021	Thu Jun 17 17:37:49 +0800 2021	FINISHED	SUCCESS	N/A	N/A	N/A	N/A
application_1623904995757_0002	root	Segment	MAPREDUCE	default	0	Thu Jun 17 17:04:23 +0800 2021	Thu Jun 17 17:05:10 +0800 2021	Thu Jun 17 17:22:45 +0800 2021	FINISHED	FAILED	N/A	N/A	N/A	N/A
application_1623904995757_0001	root	Segment	MAPREDUCE	default	0	Thu Jun 17 16:48:49 +0800 2021	Thu Jun 17 16:48:51 +0800 2021	Thu Jun 17 17:05:02 +0800 2021	FINISHED	FAILED	N/A	N/A	N/A	N/A

Showing 1 to 4 of 4 entries

Table NNU_PAGES example

ID	crawl:title	crawl:body	crawl:nextURLs	pagerank:value
www.1.com	title-1	body-1	www.2.com www.3.com	1.64
www.2.com	title-2	body-2	www.1.com	0.88
www.3.com	title-3	body-3		0.32

Table NNU_WORDS example

ID	segment:URLs
hello	www.1.com@@1-5 www.2.com@23-27 www.3.com@0-4
hadoop	www.2.com@@2-7
hbase	www.1.com@@2-6 www.3.com@101-105

http://192.168.0.20:16010/master-status

APACHE HBASE

Home | Table Details | Procedures & Locks | HBCK Report | Process Metrics | Local Logs | Log Level | Debug Dump | Metrics Dump | Profiler | HBase Configuration

Region Servers

Base Stats | Memory | Requests | Storefiles | Compactions | Replications

ServerName	Start time	Last contact	Version	Requests Per Second	Num. Regions
desktop-brown,16020,1623905035535	2021-06-17T04:43:55.535Z	2 s	2.4.3	0	3
slaker,16020,1623905033041	2021-06-17T04:43:53.041Z	1 s	2.4.3	2	4
Total:2				2	7

Backup Masters

ServerName	Port	Start Time
Total:0		

Tables

User Tables | System Tables | Snapshots

2 table(s) in set. [Details]. Click count below to see list of regions currently in 'state' designated by the column title. For 'Other' Region state, browse to hbase:meta and adjust filter on 'Meta Entries' to query on states other than those listed here. Queries may take a while if the *hbase:meta* table is large.

Namespace	Name	State	Regions							Description
			OPEN	OPENING	CLOSED	CLOSING	OFFLINE	SPLIT	Other	
default	NNU_PAGES	ENABLED	4	0	0	0	0	0	0	'NNU_PAGES', {NAME => 'crawl'}, {NAME => 'pagerank'}
default	NNU_WORDS	ENABLED	1	0	0	0	0	0	0	'NNU_WORDS', {NAME => 'segment'}

http://192.168.0.20:16010/table.jsp?name=NNU_PAGES

APACHE HBASE

Home | Table Details | Procedures & Locks | HBCK Report | Process Metrics | Local Logs | Log Level | Debug Dump | Metrics Dump | Profiler | HBase Configuration

Table Regions

Base Stats | Localities | Compactions

Name(3)	Region Server	ReadRequests (6,670)	WriteRequests (0)	StorefileSize (1.22 GB)	Num.Storefiles (9)	MemSize (0 MB)	Start Key	End Key	Region State
NNU_PAGES,,1623942271679.b3753ca6a79240e0738ce26a079c47b7.	desktop-brown:16030	3,753	0	562 MB	5	0 MB		http://news.njnu.edu.cn/info/1045/13656.htm	OPEN
NNU_PAGES,http://news.njnu.edu.cn/info/1045/13656.htm,1623946055344.1c2ba3b928b13ab1a45665c4b6bd2eed.	slaker:16030	1,081	0	332 MB	2	0 MB	http://news.njnu.edu.cn/info/1045/13656.htm	http://news.njnu.edu.cn/sylm/tt/318.htm	OPEN
NNU_PAGES,http://news.njnu.edu.cn/sylm/tt/318.htm,1623946055344.15b05099fa788b78f702b5a0da3d0ecf.	slaker:16030	1,836	0	356 MB	2	0 MB	http://news.njnu.edu.cn/sylm/tt/318.htm		OPEN

http://192.168.0.20:16010/table.jsp?name=NNU_WORDS

APACHE HBASE

Home | Table Details | Procedures & Locks | HBCK Report | Process Metrics | Local Logs | Log Level | Debug Dump | Metrics Dump | Profiler | HBase Configuration

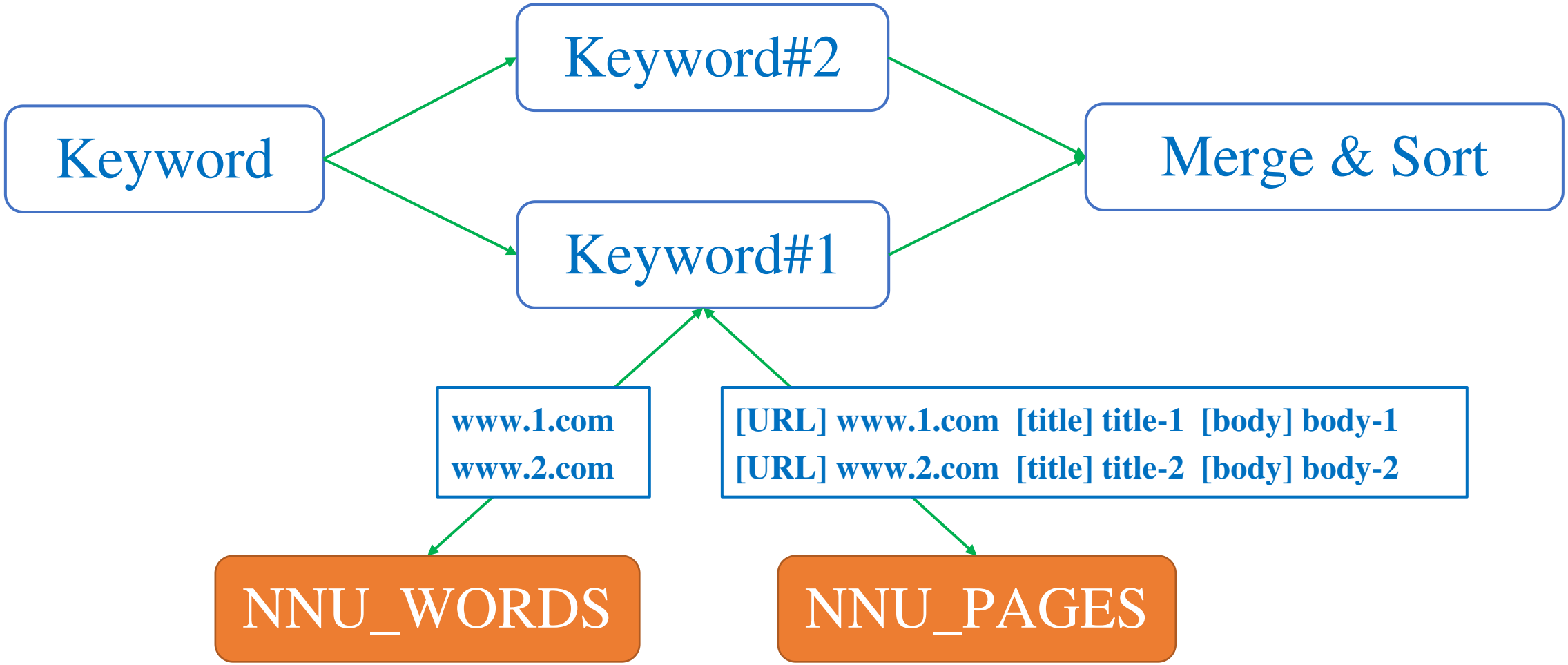
Table Regions

Base Stats | Localities | Compactions

Name(4)	Region Server	ReadRequests (94)	WriteRequests (557,795)	StorefileSize (4.05 GB)	Num.Storefiles (10)	MemSize (0 MB)	Start Key	End Key	Region State
NNU_WORDS,,1623944740629.12bfa5e0dd16dfac3f769c2821c1af1c.	slaker:16030	19	240,673	1.41 GB	3	0 MB		\xE5\x90\x8D\xE6\xA0\xA2	OPEN
NNU_WORDS,\xE5\x90\x8D\xE6\xA0\xA2,1623944740629.3b473d6b5b60268cda2dc70811846596.	slaker:16030	1	31,963	309 MB	2	0 MB	\xE5\x90\x8D\xE6\xA0\xA2	\xE5\xA9\xB7\xE5	OPEN
NNU_WORDS,\xE5\xA9\xB7\xE5,1623944726940.11d3039bd65eff6d44c558912be49b6f.	desktop-brown:16030	15	71,225	762 MB	3	0 MB	\xE5\xA9\xB7\xE5	\xE6\x97\xB6\xE5\x80\x9A	OPEN
NNU_WORDS,\xE6\x97\xB6\xE5\x80\x9A,1623944726940.76135c0165f5d4f68a899ca8adf1e1d1.	desktop-brown:16030	59	213,934	1.60 GB	2	0 MB	\xE6\x97\xB6\xE5\x80\x9A		OPEN

4. 检索模块

Vue.js + axios.js
Pyhon Django



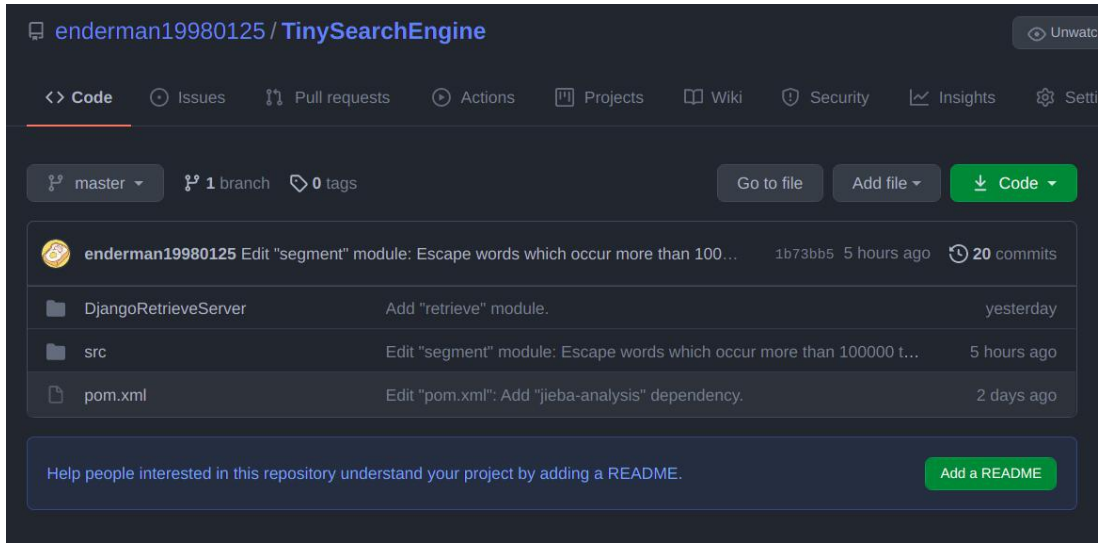
项目演示

Code available at:

https://github.com/enderman19980125/TinySearchEngine/

Project location at:

http://www.conybrown.cn/apps/retrieve-njnu/



http://www.conybrown.cn/apps/retrieve-njnu/search/?keyword=计算机

Retrieve NJNU

计算机 Search

菁菁校园-阳光网

式多元化”暑期社会实践小队的成员... 计算机学院 计算机学院“美丽中国”赴云南昆明石林实践服务 2017年07月29日 7月25日，计算机学院“美丽中国”赴云南昆明石林实践服务小分队驱车至石... 教师教育学院 教师教育学院“大千世界、百家争鸣”六合农村留守儿童关爱教育活动结项仪式 2017年07月29日 7月20日，教师教育学院“大千世界、百家争鸣”六合农村留守儿童关爱教育... 中北学院 中北

http://news.njnu.edu.cn/sylm/jjxy/1430.htm

计算机学院迎新活动-阳光网

计算机学院迎新活动-阳光网 学校主页 访问旧版 阳光首页 人物故事 标点故事 每周一星 十大人物 理论建设 通知公告 学习参考 重要言论 文化建设 通知公告 普法教育 文化品牌 资料下载 媒体南师 南师校报 视觉南师 阳光图片 阳光视频 阳光社区 学院动态 学生活动 讲座信息 体育健身 旅游聚会 社区公告 阳光首页 / 学院风采 供稿 计算机学院 更新时间 2017年09月01日 阅读量 计算机学院

http://news.njnu.edu.cn/info/1044/36114.htm

南京师范大学计算机与电子信息学院/人工智能学院

南京师范大学计算机与电子信息学院/人工智能学院 南京师范大学 English 当前位置：新闻公告>学院新闻>浏览文章 推荐文章 1校工会关于在一线达通网上商城采购 2寻访红色足迹 传承革命基因 争做时 3寻访红色足迹 传承革命基因 争做时 4南京师范大学计算机与电子信息学院 5热烈祝贺我院获批网络空间安全一级 热门文章 12017年上半年计算机学院硕士研究 2017年上半年

http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=10332

http://www.conybrown.cn/apps/retrieve-njnu/search/?keyword=孙振强

Retrieve NJNU

孙振强 Search

南京师范大学计算机与电子信息学院/人工智能学院

等奖，杨光、周青、叶顺龙、吴雨、钱荣涛、孙振强获三等奖，查涵宇、臧家瑞、曹军晓（强化院）、吴奕之（电自院）获优秀奖；Java组的张庆洋获三等奖，沈若欣获优秀奖；Python组的张媛媛（数科院）获三等奖。 版权所有：南京师范大学计算机与电子信息学院/人工智能学院 Copyright © 2020 地址：南京仙林文苑路1号 | 邮编：210023

http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=17919

关于2020年硕士研究生招生本硕贯通培养候选名单公示的通知-南京师范大学研究生院（研工部）

电气与自动化工程学院 电气工程 24 孙振强 06160111 计算机科学与技术学院 计算机科学与技术 25 邵 懿 26160328 环境学院 环境科学 26 陈欣怡 10160620 海洋科学与工程学院 海洋科学 官方微信公众号 南师研究生 南师研招 地址：南京师范大学仙林校区笃学楼 邮编：210023 院长信箱：yjsy@njnu.edu.cn 联系我们 招生办公室：025-85891892 就业招聘：025-85891612 培养办公室：025-85891790 教育管理：025-85891856 学位办公室：025-

http://grad.njnu.edu.cn/info/1054/6222.htm

计算机学院2020级新生开学典礼圆满举行-阳光网

更高的平台上展示自我、探索自我。 而后，孙振强、李虹蕊分别代表研究生和本科新生发言。孙振强从不忘初心、做好职业规划、做好心态上的转变以及处理好人际关系四个方面进行发言。李虹蕊回忆了疫情期间备战高考的过程，表示人生不能甘趋于平庸，要用青春的热烈和萌新的无惧，不断努力，挥洒汗水，做最好的自己，做南师最骄傲的学子，做中国奔向美好未来的助力者。 随后，教师代表计算机科学与技术系主任陆阳教授发言。他表示，进入大学，学习如何学习比知识本身更加重要，课堂学习之外更要注重提升自身

http://news.njnu.edu.cn/info/1044/90928.htm

http://www.conybrown.cn/apps/retrieve-njnu/search/?keyword=鲍培明

Retrieve NJNU

鲍培明 Search

南京师范大学计算机与电子信息学院/人工智能学院

长吉根林教授领衔。《计算机系统基础》由鲍培明副教授领衔。鲍老师主持中国高等教育学会"面向系统能力培养的计算机专业硬件课程群的教学改革与实践"等课题的研究，在计算机系统课程的教学过程中积累了丰富的经验，曾在全国计算机系统类课程研讨大会上分享教学改革经验。《计算机类专业导论》是计算机大类专业的一门重要基础课程，既是计算机类专业的入门课程，又是计算机类专业的学习指南。该课程由教学副院长陈波教授牵头，目前组织了包括院长吉根林教授以及专业负责人在内的学院8位骨干教师任教，多人获得校级教学名师、教学"十佳"等称号。 3. 组织研讨，加强

http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=17512

南京师范大学计算机与电子信息学院/人工智能学院

心得说了很多体会。 接着，优秀班导师代表鲍培明老师根据自己多年任职的心得，分享了学生学业指导和管理方面的一些经验。第一，注意加强与学生的交流互动，了解班里每位同学的思想动态；第二，让成绩优异的同学担任班委，形成示范效应；第三，组织学生上晚自习，加强学风建设；第四，鼓励学生出去实习，锻炼能力，开拓眼界；第五，注意关注考研、就业等信息，为学生未来发展提供指导等。 然后，彭海书记介绍了本科生班导师的职责，强调班导师要注意指导学生树立正确的世界观、人生观和价值观，注意培养学生的科研能力、社会实践能力和创新能力，同时还要关心学生的生活，关

http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=16513

南京师范大学计算机与电子信息学院/人工智能学院

任的三位教学导师：陈波教授、张国强教授和鲍培明副教授，三位青年教师：李峻博士、邵炜世博士和李莹博士，教务处督导李来发老师和学院督导孙燕老师、教师教学发展中心主任李敏老师到场进行指导，计算机学院院长吉根林教授出席了仪式。聘任仪式由教学副院长陈波教授主持。 聘任仪式上，首先，每位教学导师与所指导的新教师都在聘任协议上庄重地签下了名字。接着，每位新教师及教学导师都发表了感言。三位年轻教师都表达了自己向往大学教师这一神圣的职业，同时感谢学院给新教师配备教学导师的举措，并表示一定会珍惜学习机会，认真参加听课学习，努力提高教学能力。三位教学

http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=17513

http://www.conybrown.cn/apps/retrieve-njnu/search/?keyword=吉根林

Retrieve NJNU

吉根林 Search

数科院专家学者走进学生宿舍与同学谈考研话就业[图]-阳光网

学者有数科院党委书记王春生、数科院副院长吉根林教授、杨明教授、张福泰教授、许建华教授、计算机系主任孙燕副教授、曲维光副教授以及在读研究生代表等。活动由策划人数科院党委副书记李宝国主持。 本次活动的宗旨是各位专家学者和同学们具体探讨同学们最为关注的学业热点问题，为同学们的学习释疑解惑，为同学们的未来发展出谋划策……活动得到了很多专家学者的高度关注，也为数科院的"与专家学者同行，做时代先锋学子"系列特色活动拉开了序幕。 吉根林教授进行了精彩的关于学生考研准备的专题报告

http://news.njnu.edu.cn/info/1045/10216.htm

计算机学院迎新活动-阳光网

者们积极投身到迎新活动中。计算机学院院长吉根林、计算机学院党委书记黄菊香、党委副书记彭海、计算机学院副院长曲维光、副院长杨明、副院长蔡继明、副院长王必友以及计算机学院团委书记王海康，胡勇、王水花、刘日晨、张桂英等四位班导师、16及17级辅导员、研究生秘书、本科生秘书、办公室主任等人出席了开学典礼、宿舍看望新生、家长接待等迎新系列活动。 迎新系列活动之初识计科院　　上午八时许，清风抚着细雨，计算机学院学生联合会的部长们带着欣喜与热情开始了迎新的工作。学生会的成员在明理楼下布置好报道处，各个班级分开领取新生材料，同时也设置了咨询处

http://news.njnu.edu.cn/info/1044/36114.htm

南京师范大学计算机与电子信息学院/人工智能学院

团委书记岳嵩，南京师范大学计算机学院院长吉根林，南师大科技园管委会副主任、南师大智慧创意研究院常务副院长钱晓军等出席揭牌仪式。胥口镇政府、大学科技园、创意研究院、南京师范大学等相关部门负责同志也出席了本次仪式。仪式由计算机学院团委书记王海康主持。 在仪式正式开始前，王建一行首先在吴中区委副书记、宣传部部长李朝阳及我校挂职干部、南京师范大学（苏州）大学科技园管委会副主任、苏州智慧创意研究院常务副院长钱晓军等的陪同下参观了大学科技园，并与驻园校友企业负责人代表、计算机学院2004届毕业生、苏州优肯信息工程有限公司总经理王银兵等亲切交

http://ceai.njnu.edu.cn/Item/Show.asp?m=1&d=9418

http://www.conybrown.cn/apps/retrieve-njnu/search/?keyword=端午节

Retrieve NJNU

端午节 Search

教师教育学院12级历史师范成功举行弘扬端午精神活动-南师大教师教育学院

教育学院12级历史师范为了迎接即将到来的端午节，在学明楼成功举行弘扬端午精神主题班会。　　首先主持人向大家提问，端午节有哪些别名，同学们纷纷回答：端阳节、午日节、五月节、五日节、艾节、端五、重午、重五、午日、夏节、蒲节等，没有百度，就回答了问题，相当博学渊源，然后是端午节的由来。有两位同学分别讲述；了传说，本来是夏季的一个驱除瘟疫的节日，后来楚国诗人屈原于

http://jsjyxy.njnu.edu.cn/info/1032/2178.htm

2019年端午节放假调休通知-南京师范大学心理学院

2019年端午节放假调休通知-南京师范大学心理学院 Toggle navigation 首页 学院概况 学院简介 历史沿革 现任领导 师资队伍 教授 副教授 讲师 在站全职博士后 实验人员 行政人员 客座教授 兼职导师 离退休人员 科学研究 科研团队 科研平台 科研项目 科研成果 科研获奖 人才培养 培养计划 本科生 研究生 奖学金计划 教学成果 教师成果 学生成果 教学实践 招生工作 党建工会 招生就业 本科生招生 研究生招生 就业 学生工作 首页 教务通知 本科生 研究生 教管通知 本科生 研究生 学生活动 本科生 研究生

http://xlxy.njnu.edu.cn/info/1058/1990.htm

《文化遗产的释读——象征解码与误读解析》讲座纪要-南京师范大学社会发展学院

作为端午这一非物质文化遗产的文化符号。在端午节中，菖蒲一般象征着“宝剑”，有着斩鬼的作用；而艾叶因形似虎爪，认为有吃鬼的作用。总体而言，菖蒲和艾叶是端午用以辟邪驱鬼的文化符号。此外，陶先生认为，粽子并不像现在端午节中用以缅怀爱国诗人屈原，准确来讲，粽子主要是用来纪念伍子胥。 陶先生再次强调，文化遗产，不论是物质文化、行为文化，还是语言文化、精神文化，都有赖于象征的应用，象征往往成了某些文化模式的精辟概括。中国古代哲学家将“象”、“形”或“象”、“器”相对，《易传》

http://sfy.njnu.edu.cn/info/1021/8623.htm

遇到的困难

- 未完全过滤特殊字符，导致 MapReduce 出错

  没有把不可见控制字符^M过滤掉，而 Hadoop 认为^M是换行符，误将一行切分为多行；

- 中文分词 MapReduce 过程内存溢出

  执行到 Reduce 过程 66% 时发生错误 Error: Java heap space，指定 JVM 参数 -Xmx2048m，并在 mapred-site.xml 中设置 mapreduce.reduce.memory.mb 为 2048 解决；

- HBase拒绝使用连接器登录

  HBase使用2181端口进行集群通信，使用16010端口进行网页监控，使用9090端口开放 thrift 服务器；

自己的收获

- HDFS一地存储*，随处可见

  偶数分片存储在192.168.0.20上，奇数分片存储在192.168.0.30上，但数据对所有到该HDFS的连接可见，*也可能是多地冗余存储；

- HBase自动数据平衡

  所有的数据都从192.168.0.20上迁移到 HBase 中，但192.168.0.30上也存储了大量数据；

- Google + StackOverflow　YYDS(永远滴神！)

  根据 Exception 和 *.log，定位bug，别去百度，上面啥都没有！自己写程序也养成了使用日志模块的习惯；

未来的改进

- **自动化流程**

  目前各模块之间需要手动迁移数据，而真正的搜索引擎应该实现自动化数据流的过程，并及时爬取更新的网页；

- **PageRank 算法的连惯性**

  新添加的网页使用初始值，已添加的网页使用现有值，继续执行PagRank算法，收敛性？(马尔可夫性质)，正确性？

- **多维度的页面排名度量**

  结合关键词的TF-IDF值、页面的PageRank值、关键词数量、关键词位置、标题长度、发布时间等因素，提高页面排序质量；

The End