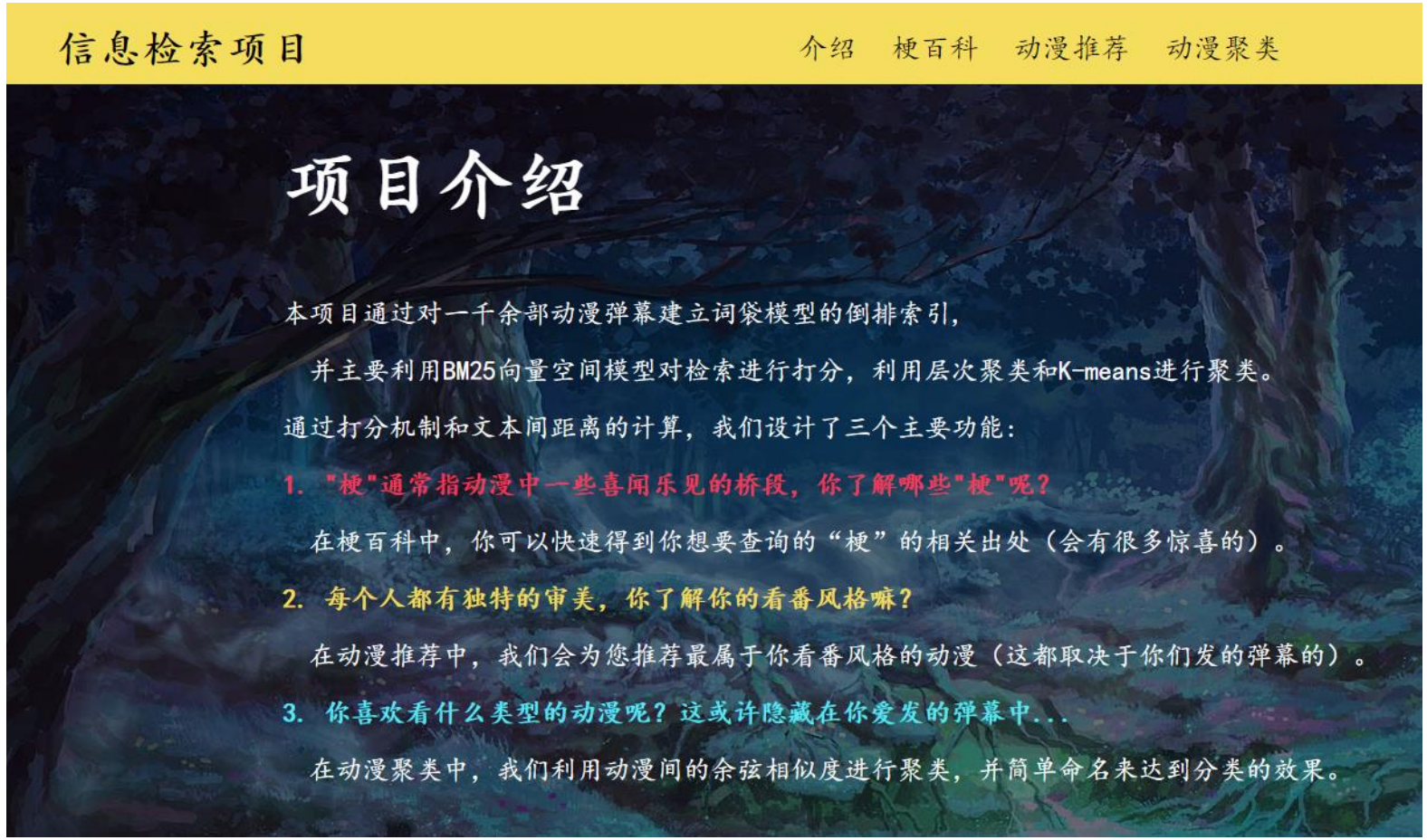
动漫弹幕的文本检索和应用

期末答辩

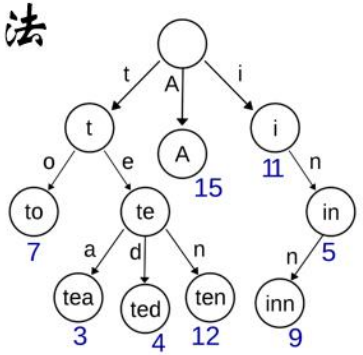
11811721 庄湛

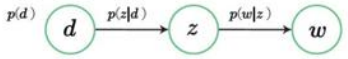
2020/6/20

1. 简短词似然法分词
2. 梗百科
3. 动漫推荐
4. 风格聚类
5. Web设计
6. 演示操作



分词方法





S(大) = 305
S(乔) = 271
P(乔|大) = 202/305
P(大|乔) = 202/271

(1) S(a) > 10
(2) P(b|a) > 1.3 - 0.35ln(S(b))
(3) P(a|b) > 0.6

```
retTree = treeNode('大', headerTable['大'][0], None)
next_tree = headerTable['大'][1]
while next_tree is not None:
    for child in next_tree.children.values():
        addNode(retTree, child)
    next_tree = next_tree.nodeLink
```

```
retTree.disp()
```

```
  大   305
    乔   202
      停   9
        止   9
          了   9
            思   9
              考   9
      不   7
        知   7
          道   5
            应   1
              该   1
                点   1
                  赞   1
                    还   1
                      是   1
                        沉   1
                          默   1
            该   3
              怎   1
                么   1
                  办   1
              点   2
                赞   2
                  还   2
```

利用链式树形结构进行分词，结果并不理想，对于较多的动漫角色名和错字、错词、生词等都没有较好的结果。

根据弹幕特点，设计了另一种分词方法 →

259	雷区蹦迪	2452
260	活着不好	2445
261	太草	2445
262	好好听	2421
263	商业互吹	2394
264	世界真小	2377
265	恭喜	2374
266	好温柔	2367
267	裂开	2366
268	你想多	2342
269	回忆杀	2338
270	感谢土豪	2331
271	好基友	2325
272	痴汉	2312
273	护眼	2302
274	不是	2298
275	熊孩子	2290
276	绯红之王	2289
277	投币	2283
278	好好看	2278
279	假的	2245
280	精准踩雷	2232
281	好人	2230
282	跨服聊天	2224
283	平角裤	2222

	term	freq
1	路飞	823
2	佐助	740
3	鸣人	656
4	萨斯给	569
5	钢铁侠	439
6	炭治郎	353
7	二柱子	281
8	头柱	266
9	辉夜	260
10	喜羊羊	207
11	古河渚	161
12	赖皮蛇	112
13	红孩儿	111
14	吼姆拉	104
15	黑崎一护	97
16	晓美焰	96
17	四宫辉夜	43
18	冈崎汐	36
19	碳治郎	29
20	冈崎渚	15
21	岸本	9
22	炭炭	6
23	虹猫	6
24	炭之郎	5
25	口鸟人	5
26	岸本齐史	4
27	吸氧羊	3

Jieba分词 原始结果：

```
print ("/".join(jieba.cut("大乔为你点赞")))
```

大乔为/你/点赞

利用FP-Growth Tree增加新词：

```
print ("/".join(jieba.cut("大乔为你点赞")))
```

大乔/为/你/点赞

利用简短词似然法增加新词：

```
print ("/".join(jieba.cut("大乔为你点赞")))
```

大乔/为你点赞　　兼语短语更符合理解

语料特点：

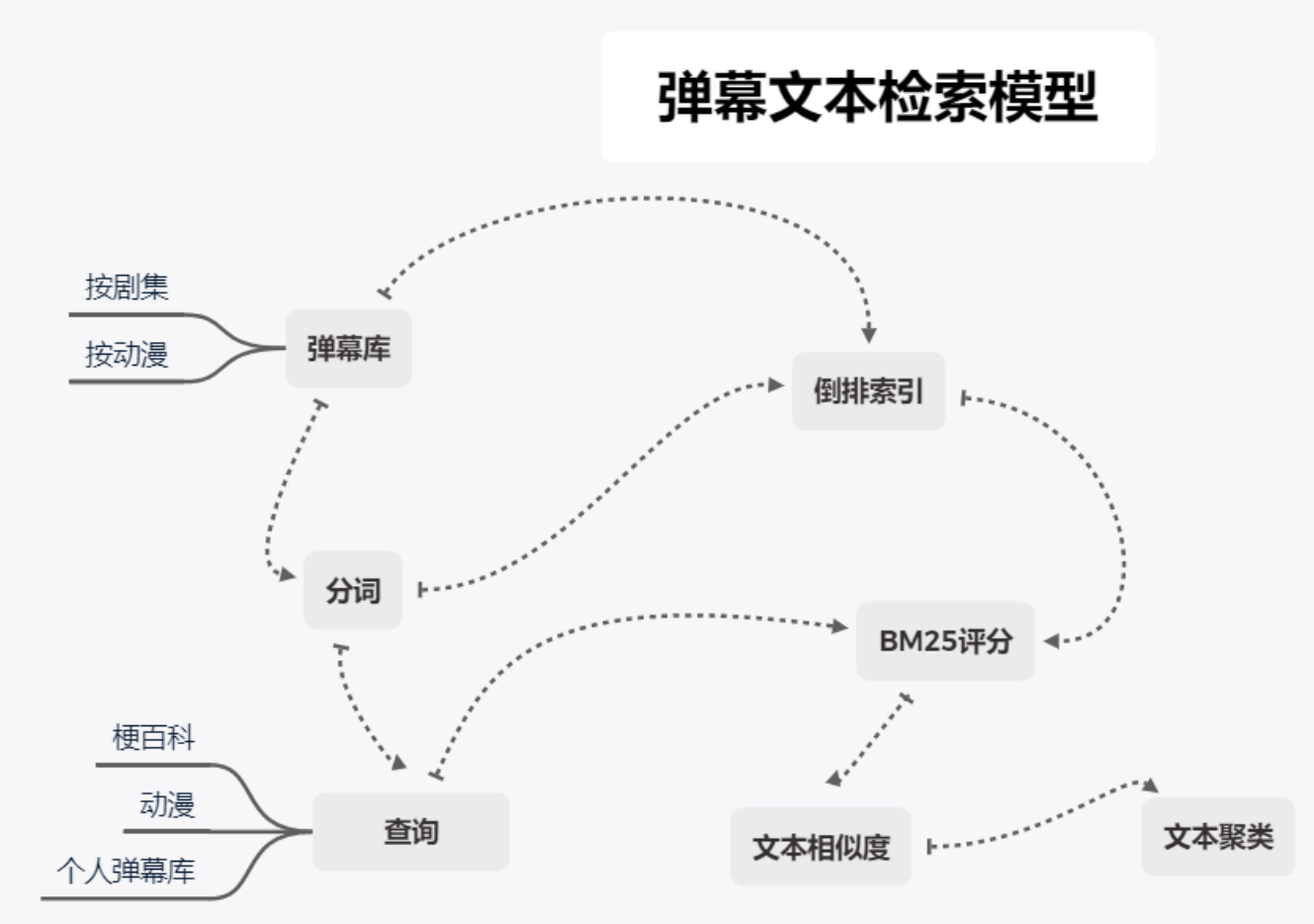
- 语料库总量大
- 单句精简短小
- 新词叠词较多
- 错别字词较多
- 词语较口语化
- 部分词项词频较高
- 不同语境下差异大

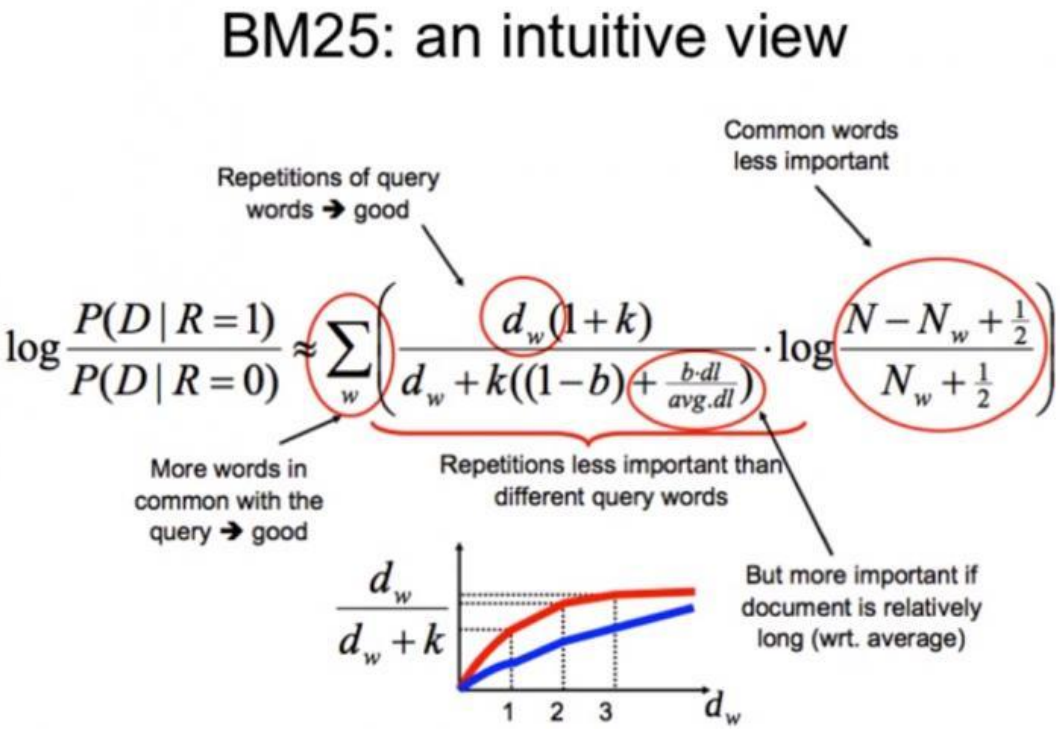
该方法的**优点**：时间复杂度为O(n)， 原理简单，保留信息更多，短语式切分更符合人类理解。

该方法的**缺点**：局限性大，该类语料库较少。我能想到的也只有弹幕语言，或者聊天语料库。

我认为这是一种**有灵魂的分词**，和以往分析**前后凝聚度**，**最短路径**以及**n-gram分词法**相比，该方法利用了**语料特点**和**集体习惯**。当人们表达情绪和分享时，往往利用较短而常用的词汇。在爬取得到的五千九百多万条弹幕中，有两千六百多万条弹幕长度小于5。平均弹幕长度远小于正常语料中两个标点间长度。并且由于错字较多、新词较多、较为庞大等特点，我认为该类语料库有很高的研究价值，对于这种新型文本的研究可以推动统计自然语言处理学科的发展。

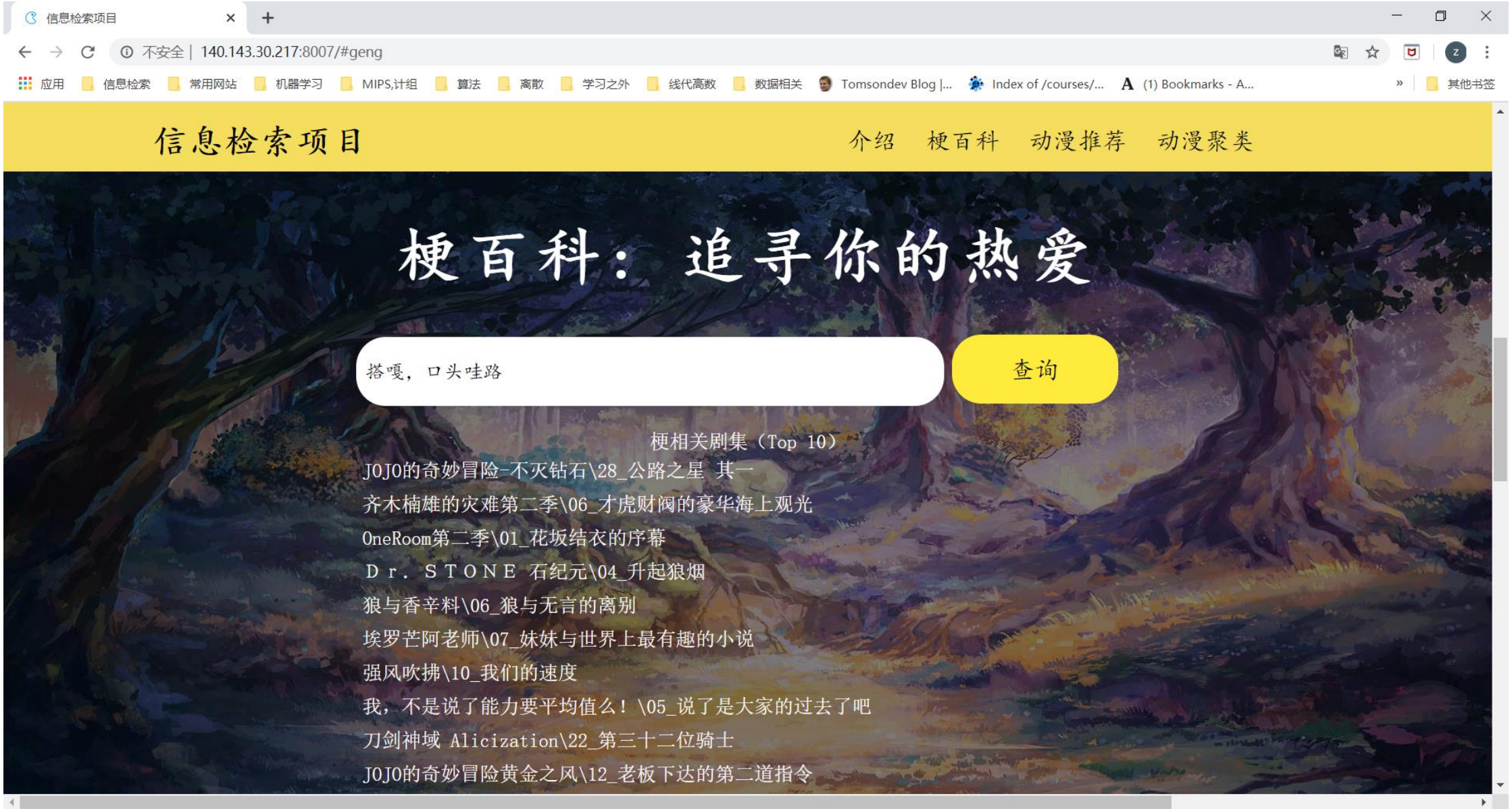
梗百科：



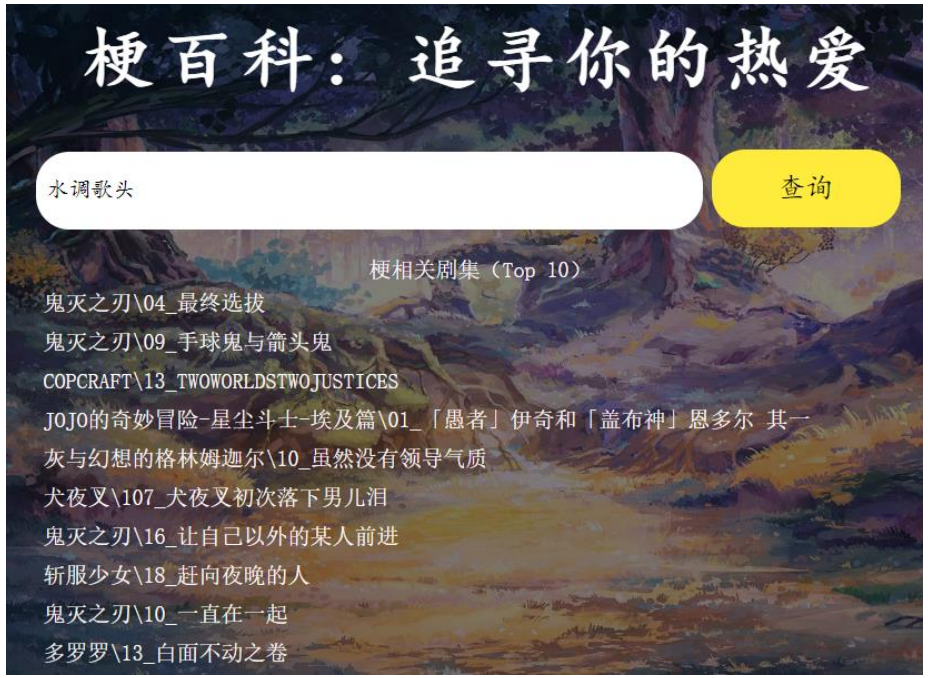


BM25模型是在TF-IDF的基础上，增加了一个TF值的上界约束和增加了对文档长度的考量。

"梗"通常指动漫中一些喜闻乐见的桥段

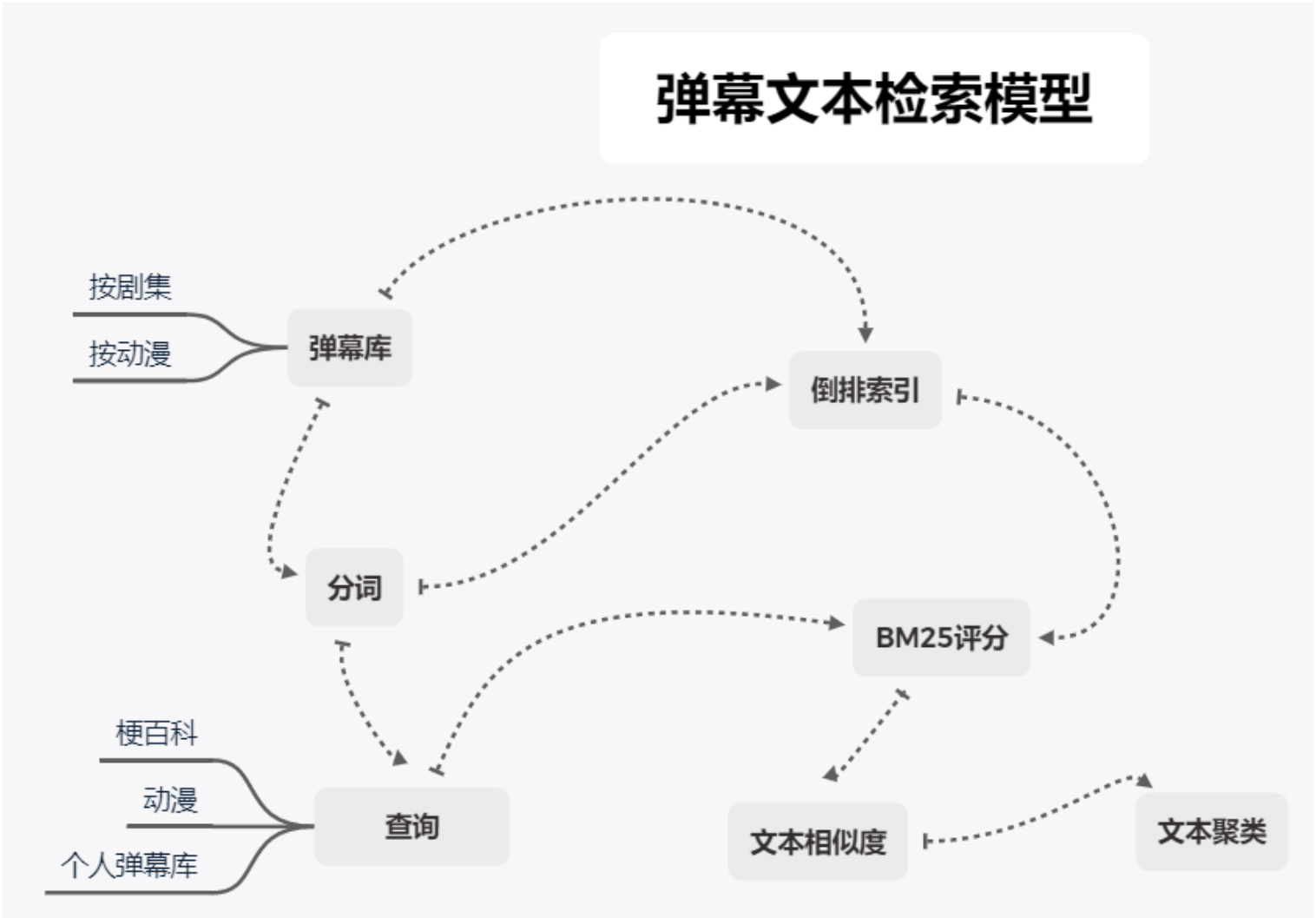

信息检索项目
不安全 | 140.143.30.217:8007/#geng
应用 信息检索 常用网站 机器学习 MIPS,计组 算法 离散 学习之外 线代高数 数据相关 Tomsondev Blog |... Index of /courses/... (1) Bookmarks - A... 其他书签
信息检索项目
介绍 梗百科 动漫推荐 动漫聚类
梗百科：追寻你的热爱
搭嘎，口头哇路
查询
梗相关剧集（Top 10）
JOJO的奇妙冒险-不灭钻石\28_公路之星 其一
齐木楠雄的灾难第二季\06_才虎财阀的豪华海上观光
OneRoom第二季\01_花坂结衣的序幕
Ｄｒ．ＳＴＯＮＥ 石纪元\04_升起狼烟
狼与香辛料\06_狼与无言的离别
埃罗芒阿老师\07_妹妹与世界上最有趣的小说
强风吹拂\10_我们的速度
我，不是说了能力要平均值么！\05_说了是大家的过去了吧
刀剑神域 Alicization\22_第三十二位骑士
JOJO的奇妙冒险黄金之风\12_老板下达的第二道指令

左图为查询结果第一条，也是“梗”出处
右图为查询结果第二条，属于一种“官方玩梗”



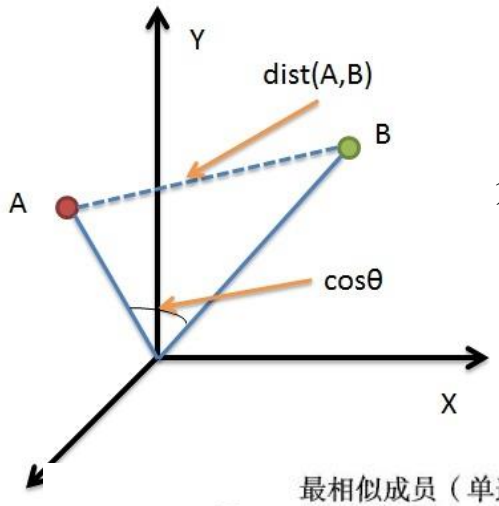
项目意义和价值：

1. 提供不了解“梗文化”的人，一个快速了解的途径，可以很有趣地吸引圈外人士。
2. 给喜欢玩梗的人一个补番推荐或者玩梗指南，让“梗”这种亚文化更有活力。
3. 分析“玩梗”、“造梗”等动漫文化表达出的观众对动漫的期望来研究更多商业价值。以及提供一种研究不同动漫间的联系的途径。

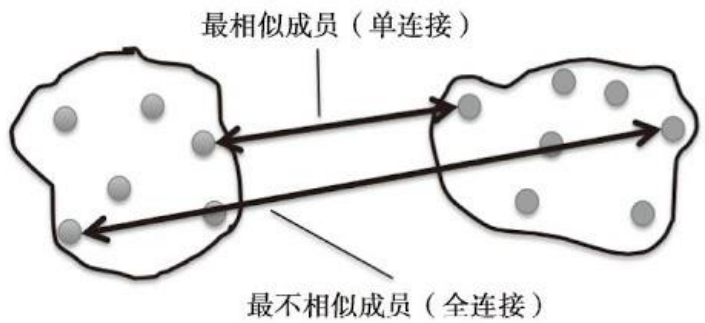


$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

（动漫总数，词项总数）



余弦相似度



[四月是你的谎言，徒然喜欢你，月色真美，萤火之森]　（纯爱类型）

→

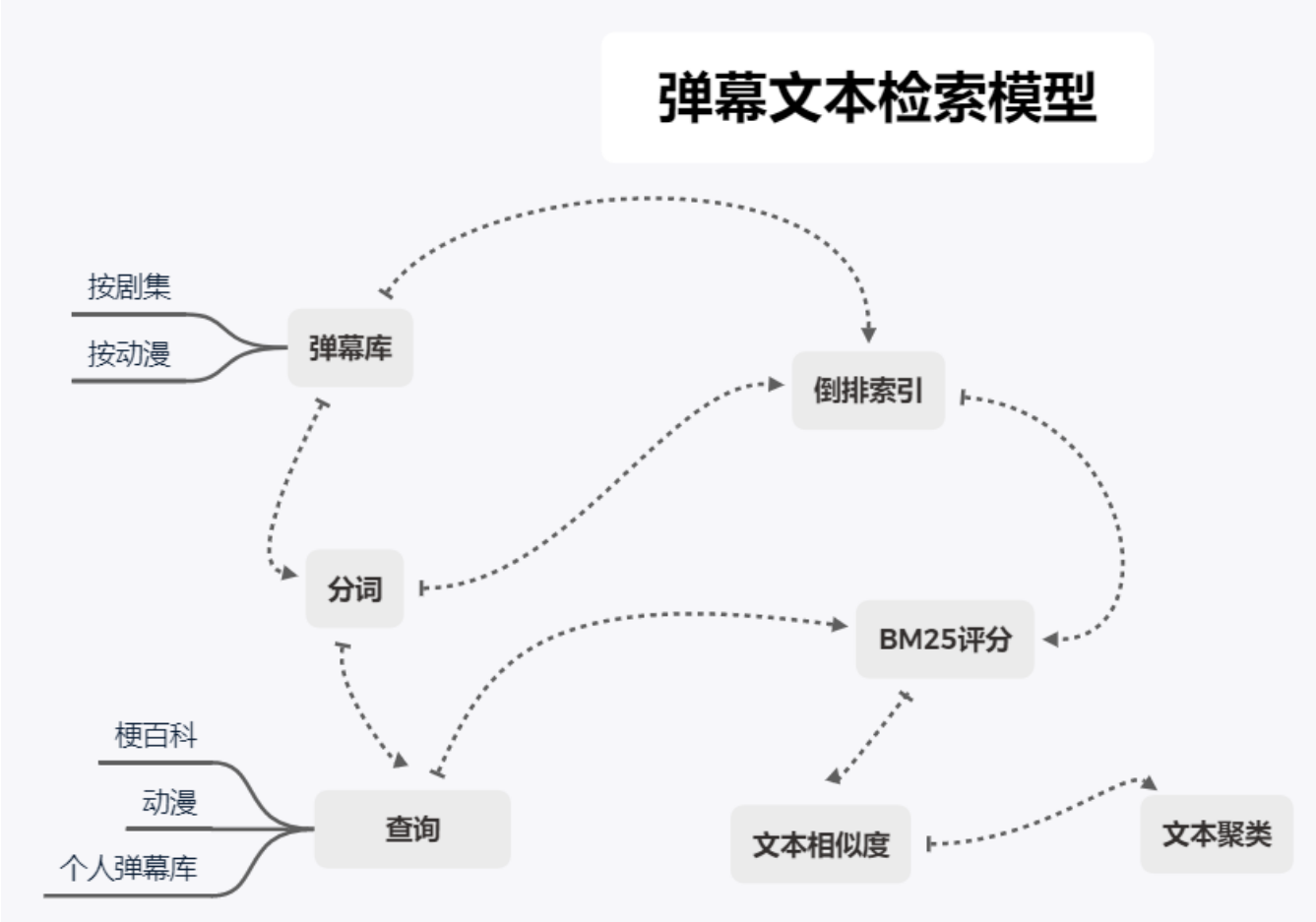
[樱花庄的宠物女孩，秒速五厘米，我们仍未知道那天所看见的花的名字（未闻花名），可塑性记忆，中二病也要谈恋爱]

[乒乓，黑子的篮球，强风吹拂，灌篮少年]

（体育运动类）→

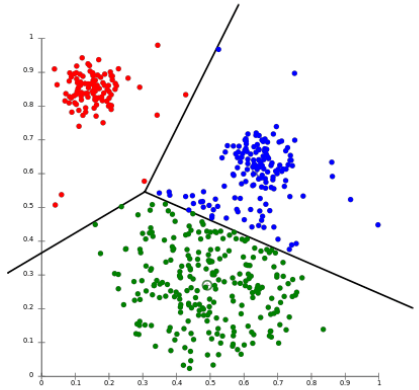
[黑子的篮球第三期，排球少年，足球小将，DAYS]

思考：如果将一个人所发的全部弹幕做成一个弹幕集当做一部“动漫”，也可以以此做更加精准的动漫推荐。



非连续增加惩罚

```
if len(clust_1)>100 and len(clust_2)> 100:
    dis*=1.4
elif len(clust_1)>50 and len(clust_2)> 50:
    dis*=1.2
elif len(clust_1)>30 and len(clust_2)> 30:
    dis*=1.1
elif len(clust_1)>10 and len(clust_2)> 10:
```

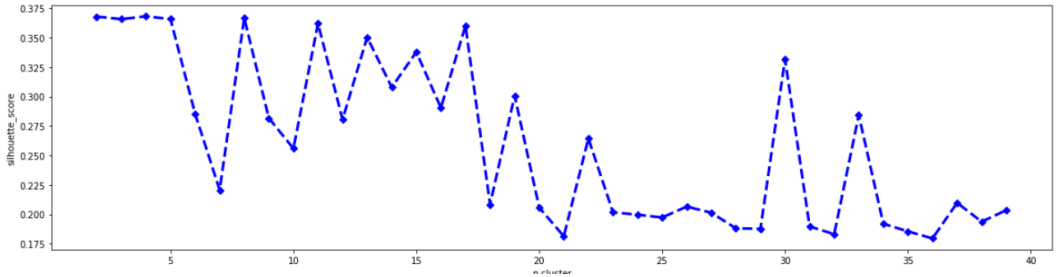


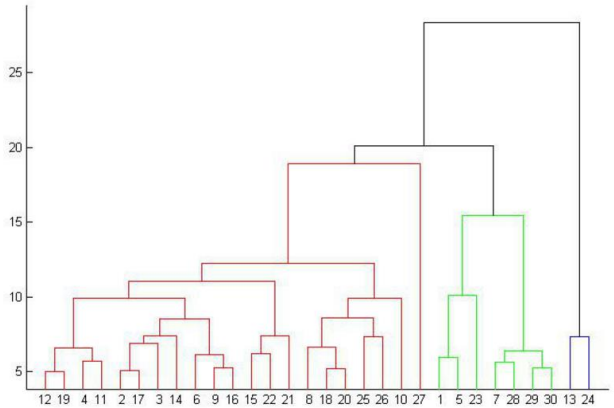
K- Means

需要确定簇数

有高随机性（初始随机）

轮廓系数随簇数变化图





层次聚类

每次合并两个类

0号宿舍,我家大师兄脑子有坑特别篇,斩兽之刃,凸变英雄LEAF,萌妻食神,通灵妃,剑网3·侠肝义胆沈剑心,少年歌行,全职高手第一季,刺客伍六七,罗小黑战记,如果历史是一群喵,王者别闹,请吃红小豆吧,灵笼,镇魂街第二季,风灵玉秀

11eyes,出租魔法使,Caligula卡里古拉,境界线上的地平线,神不在的星期天,CDE,黑神,CHAOS;CHILD,混沌之脑,超自然9人组,时间旅行少女,Infini-TForce,百合熊风暴,历物语,终物语,猫物语黑,斩首循环蓝色学者与戏言跟班,鸦KARAS,宇宙巡警露露子,特别的她,龙的牙医,尸者的帝国,女神异闻录-圣洁之魂-,女神异闻录3剧场版1SpringofBirth,女神异闻录3剧场版4WinterofRebirth,至高指令OAD

227,神推偶像登上武道馆我就死而无憾,ANIMAYELL,LoveLiveSchoolIdolProject,LoveLiveSchoolIdolProject第二季,LoveLiveSunshine,BanGDream,BanGDream第二季,BanGDream少女乐团派对☆PICO,少女☆歌剧RevueStarlight,AnneHappy,三者三叶,斯特拉的魔法,ComicGirls,NEWGAME,SlowStart,请问您今天要来点兔子吗,雏子的笔记,黄金拼图,属性咖啡厅,邻家索菲,恋爱小行星,只要贝尔哲布布大小姐喜欢就好,街角魔族,URARA迷路帖,此花亭奇谭,放学后桌游俱乐部,偶像选举,音乐少女,普通女高中生要做当地偶像,ACHANNEL,若叶女孩,玛纳利亚的密友MysteriaFriends巴哈姆特之怒玛娜利亚魔法学院,次元发电机,番剧茶会

ACCA13区监察课,江户盗贼团五叶,GANGSTA,青春生存游戏,DaDaDa,天使怪盗,寻找满月,萩萩公主,天堂之吻,近所物语,NANA,玩偶游戏,少女革命,他和她的故事,恋爱情结,橘子酱男孩,SA特优生,爱丽丝学园,晨曦公主,东京猫猫,怪盗圣少女,猫眼三姐妹,天国少女,最游记,潘朵拉之心,老虎和兔子,人鱼的旋律,伯爵与妖精,TRICKSTER,妖怪公寓的幽雅日常,舞动青春,十二国记,彩云国物语,不可思议星球的双胞胎公主,不可思议星球的双胞胎公主Gyu,玻璃假面,魔女的考验【中文】,魔女的考验【日语】,光能使者,超级酷乐猫,攻壳机动队STANDALONECOMPLEX,铁臂阿童木2003,变形金刚2008,变形金刚领袖之证美版第一季,复仇者世上最强英雄组合,奇幻贵公子,心灵侦探八云,学校怪谈,怪谈餐厅,X战记,小鸠,遥远时空中八叶抄,怪医黑杰克,青年黑杰克,心灵的声音,悲惨世界少女珂赛特,莎拉公主,阿尔卑斯山的少女,青春歌舞伎,凡尔赛玫瑰,双面骑士,英国恋物语艾玛第二幕,格林童话剧场【中文】,银之匙SilverSpoon第一季,魔法骑士,傀儡师左近,推理之绊,四畳半神话大系,有顶天家族,RWBY,拽妹黛薇儿第一季,脆莓公园第一季,源氏物语千年纪Genji,紫式部源氏物语,我是小甜甜,我家浴缸的二三事,梦幻拉拉,和歌子酒动画,妙手小厨师,英国一家吃在日本,荷包蛋的蛋黄什么时候戳破才好,米老鼠的黑白动画片生涯,伯纳德小姐说,火影忍者舞台剧,无家可归的小孩,圣哥传,佩琳物语

category	animations
国创	{0号宿舍,我家大师兄脑子有坑特别篇,斩兽之刃,凸变英雄LEAF,萌妻食神,通灵妃,剑网3·侠肝义胆沈
少女、公主、精灵啥	{11eyes,出租魔法使,Caligula卡里古拉,境界线上的地平线,神不在的星期天,CDE,黑神,[C]THEM
偶像	{227,神推偶像登上武道馆我就死而无憾,ANIMAYELL,LoveLiveSchoolIdolProject,LoveLive
不确定	{ACCA13区监察课,江户盗贼团五叶,GANGSTA,青春生存游戏,DaDaDa,天使怪盗,寻找满月,萩萩公主,
不清楚	{ACHANNEL,若叶女孩,FORTUNEARTERIAL-赤之约定-,MYSELF;YOURSELF,SOLA,悠久之翼,凉风,
热血、决斗	{AIR,AngelBeats,ISLAND,寻找失去的未来,Rewrite2ndSeason,CAROLE & TUESDAY,CAROL
机甲	{AKB0048第一季,AKB0048第二季,偶像大师,偶像大师灰姑娘女孩,我要成为双马尾,AngeVierge,Re
恋爱	{AngelsofDeath,虚构推理,Charlotte,中二病也要谈恋爱,中二病也要谈恋爱恋,我女友与青梅竹马
历史等	{BASQUASH,NEEDLESS,星界死者之书,魂兽,快盗天使TWINANGEL,萌单,月咏-MOONPHASE-,简单易
冒险	{Butlers～千年百年物语～,鹿枫堂,隐之王,KARNEVAL狂欢节,无法逃离的背叛,幻影少年,八犬传-东
科幻、决斗	{CANDYBOY,青之花,海物语,玉响～hitotose～,LEVELE,兽王星,深渊传说,玛德莱克丝,玲音,科学
运动	{DAYS,足球小将83版,足球小将平成版,足球风云,足球骑士,野狼前锋,H2好逑双物语,棒球伙伴,棒球大
也是科幻决斗？	{D·N·A2他到底失去了什么,铁腕巴迪DECODE,到另一个你的身边去,此时此刻的我,破天荒游戏,魔女猎
未知	{GOGO575,不思议美眉,漫画少女,漫研部,网球并不可笑嘛第一季,网球并不可笑嘛第二季,限制级杀手,
刀剑	{InfiniteStratos2,三坪房间的侵略者,机巧少女不会受伤,漆黑的子弹,绯弹的亚里亚,学战都市Ast
青春	{JustBecause,白色相簿2,只要你说你爱我,好想告诉你第一季,好想告诉你第二季,青春之旅,邻座的怪

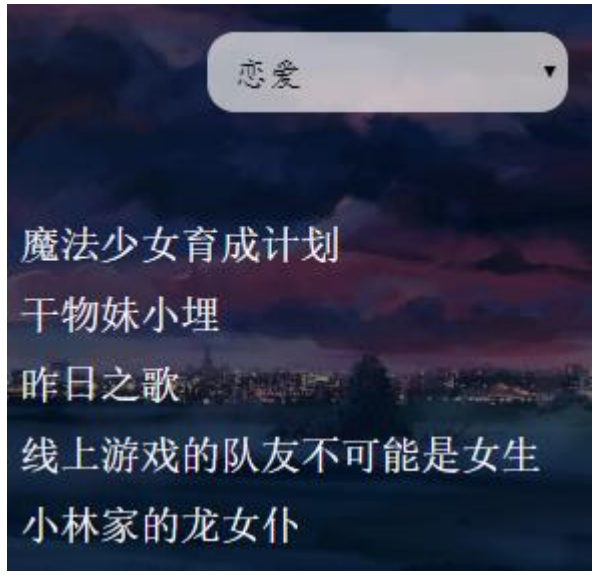
缺点：

1. 没有明显的分组依据（但是可以通过计算该组的平均各词项得分，来推测其主题）。
2. 从1000多个聚类降至20个聚类时间较长。

后续工作：

1. 可以提取各组得分影响权重最大的几个词项，进而分析该组动漫风格。
2. 曾尝试对各动漫词项，采用LDA线性判别分析来分析关键词项。但是由于原始词项向量进行预处理后仍有300w维，没有完成该项工作。后因web开发耗时，未完成进一步处理此项。


动漫聚类：找到你的热爱
国创
给我瞅瞅
随机推荐五部该类动漫
国创
偶像
少女、公主、精灵啥的
国创
热血、决斗
不确定
不清楚
运动
科幻、决斗
恋爱
未知
历史等
机甲
青春
冒险
刀剑
也是科幻决斗？
风灵玉秀
全职高手第
斩兽之刃
剑网3·侠
王者别闹
运动
足球小将83版
足球小将平成版
乒乓
足球骑士
棒球伙伴


恋爱
魔法少女育成计划
干物妹小埋
昨日之歌
线上游戏的队友不可能是女生
小林家的龙女仆


刀剑
落第骑士英雄谭
刀剑神域Alicization
我，不是说了能力要平均值么！
这个勇者明明超强却过分慎重
学战都市Asterisk第二季

技术栈

python™

PostgreSQL

django

HTML

uWSGI

NGINX

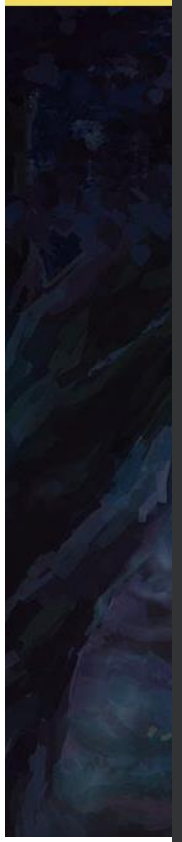
CSS

腾讯云

CentOS

项目应用网址：http://140.143.30.217:8008/ 比本地慢很多

信息



```python
def jiebaUtil(conn):
    cursor = conn.cursor()
    query = """select * from postgres.ir.raw_words"""
    cursor.execute(query)
    new_words = cursor.fetchall()
    for key_word in new_words:
        if key_word[1] > 2:
            jieba.add_word(key_word[0], freq=key_word[1] * 10, tag=None)


def jiebaUtil_file():
    new_words = []
    f = open("static\\new_words_all.txt")
    lines = f.readlines()
    for line in lines:
        split = line.strip().split(',')
        new_words.append((split[0], int(split[1])))
    f.close()
    for key_word in new_words:
        if key_word[1] > 2:
            jieba.add_word(key_word[0], freq=key_word[1] * 10, tag=None)
```

项目应

ngo

SGI

iNX

腾讯云

CentOS

感谢观看

开发环境: http://127.0.0.1:8000/

生产环境: http://140.143.30.217:8008/