

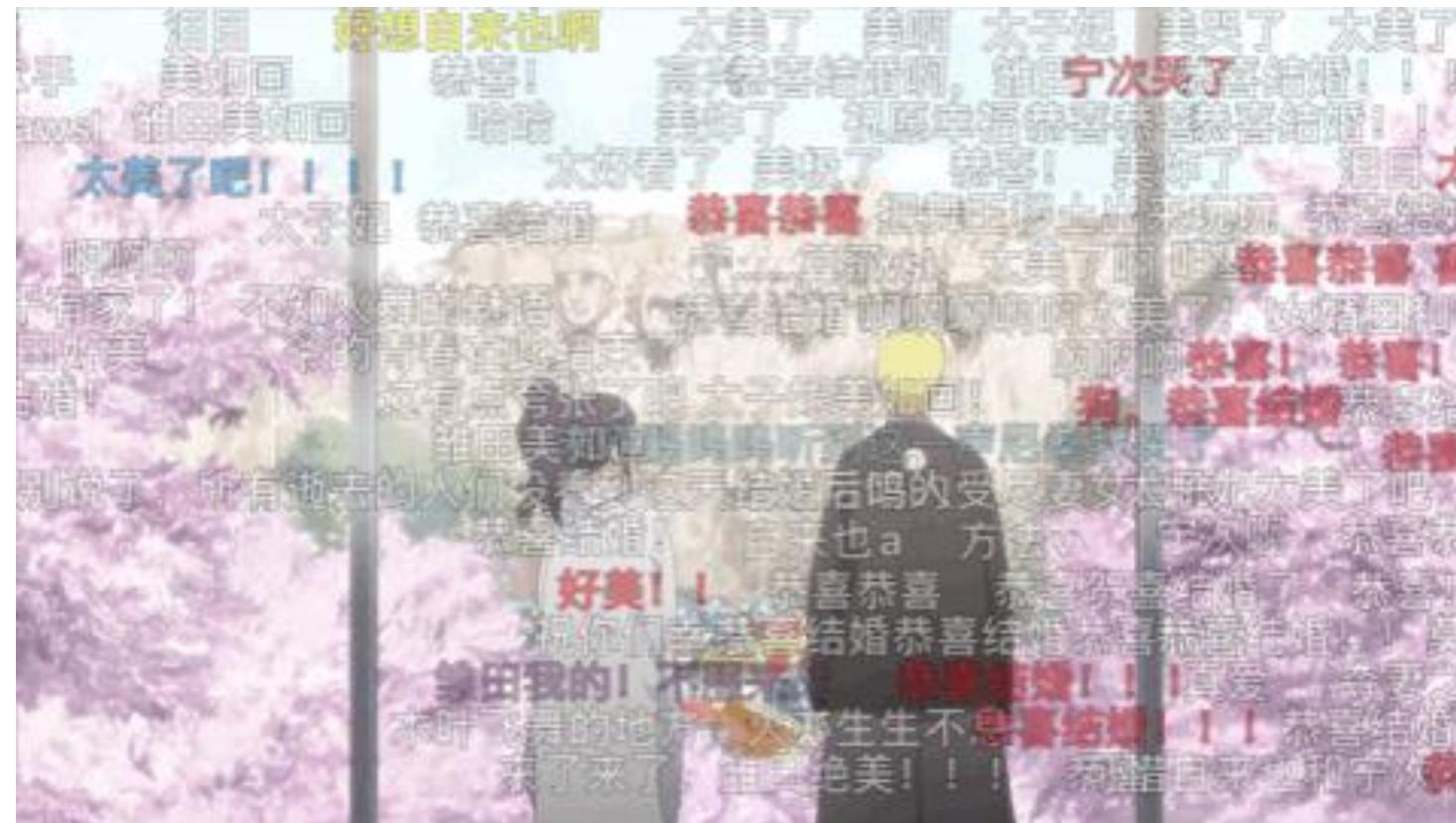
动漫弹幕的文本检索和应用

开题报告

11811721 庄湛

2020/4/25

语料库选择



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

语料库选择



哔哩哔哩弹幕网

api类型	单集弹幕数	访问需求	解决方案
实时弹幕	当日3000	无需登录	长期爬取
历史弹幕	累计10w+	需要登录, 反爬虫	cookies 池



两部较热门动画采用了实时弹幕爬取，约400w条数据
其余1000多部动漫采用实时弹幕爬取，约5000w条数据

<https://api.bilibili.com/x/v2/dm/history?type=1&oid={}&date={}>

<http://comment.bilibili.com/{}.xml>

存储格式

CLANNAD	ISLAND	ReLIFE完結篇	WZ	白色相簿上半篇	冰海战记
Classroom☆Crisis	JOJO的奇妙冒险	REriderD-穿越时空的德理达	X战記	白兔糖	波子汽水
CodeRealize ~创世的姫君~	JOJO的奇妙冒险-不灭钻石	revisions	X战警	百变小樱	玻璃假面
ComicGirls	JOJO的奇妙冒险-黄金之风	Rewrite2ndSeason	X战警进化	百变之星	玻璃舰队
COPCRAFT	JOJO的奇妙冒险-星尘斗士	Re創造主	YAT安心宇宙旅行	百合熊风暴	伯爵与妖精
CROSSFIGHT弹珠人	JOJO的奇妙冒险-星尘斗士-埃及篇	RobiHachi	ZEGAPAIN	百炼霸王与圣约女武神	伯纳德小姐说
DaDaDa	JustBecause	RoomMate	阿尔卑斯山的少女	拜托了老师	博多豚骨拉面团
DancewithDevils	K	RWBY	阿尔蒂	拜托了双子星	博人传 火影忍者新
DARKERTHANBLACK-黑之契约者-	KARNEVAL狂欢节	SACREDSEVEN	阿童木起源	棒球大联盟2	不吉波普不笑
DAYS	KERORO军曹	SA特优生	阿宅的恋爱真难	棒球伙伴	不可思议星球的双胞
D r . S T O N E 石纪元	KRETURNOKINGS	SchoolgirlStrikers	埃罗芒阿老师	棒球英豪TV版	不可思议星球的双胞
DRAMATICMurderOVA	LOSTSONG失落的歌谣	SD高达三国传BraveBattleWarriors	艾莉森与莉莉娅	薄樱鬼	不思议美眉
EVA新世纪福音战士	LoveLiveSchoolIdolProject	SHOWBYROCKShort	爱吃拉面的小泉同学	宝石商人理查德的鉴定	不愉快的怪物庵
FateApocrypha	LoveLiveSchoolIdolProject第二季	SHOWBYROCK第二季	爱丽丝学园	宝石之国	不愉快的怪物庵续
FateEXTRALastEncore	LoveLiveSunshine	SHUFFLE	爱你宝贝	暴力宇宙海贼	彩云国物语
Fatestaynight[UnlimitedBladeWorks]第二季	MacrossPlusOVA	SHUFFLEMEMORIES	爱情泡泡糖	爆TECH爆丸	苍色骑士中文
Fatestaynight[UnlimitedBladeWorks]第一季	Megalobox	SlowStart	爱书的下克上为了成为图书管理员不择手段	爆丸粤语版	苍天航路
Fatestaynight06版	MIX	SOLA	爱丝卡与罗吉的工作室黄昏之空的炼金术士	爆旋陀螺	苍之彼方的四重奏
FateZero第一季	MYSELF;YOURSELF	STAR DRIVER闪亮的塔科特	爱天使传说	爆走猎人	苍之茧
FORTUNEARTERIAL-赤之约定-	NANA	STARRY☆SKY	爱玩怪兽	悲慘世界少女珂赛特	柴犬阿旺的和式生涯
Free-Dive to the Future-	NEWGAME	TARATARI	暗黑破坏神在身边。	被狙击的学园	超次元游戏海王星
GAMERS电玩咖	NOGAMENOLIFE游戏人生	TheRevelation	暗夜第六感	笨女孩	超级机器人大战OG
GANGSTA	OneRoom	ToHeart回忆永恒	暗夜魔法使	比宇宙更远的地方	超级酷乐猫
GOGO575	OneRoom第二季	TRICKSTER	暗芝居:世界黑暗图鉴	碧蓝航线	超能力女儿
GOSICK	OVERLORD	TSUKIPRO	暗芝居第六季	碧蓝之海	超能奇兵
GR铁甲人	OVERLORD III	ULTRAMAN机动奥特曼	暗芝居第七季	变形金刚2008	超人高中生们即便在
H2O苏砂的印记	OZMAFIA	UN-GO因果论	暗芝居第四季	变形金刚大电影	超人战队BARATTA
H2好逑双语物语	PHI-BRAIN神之谜题第一季	UQHOLDER悠久持有者	奥运高手	变形金刚领袖之证美版第一季	超时空骑兵团
HandShakers	pop子和pipi美的日常	URARA迷路帖	八男别闹了	变形金刚三乘合体	超时空世纪02
HelloKitty苹果森林第三季	RagnastrikeAngels	VENUSPROJECT-CLIMAX-	八犬传·东方八犬异闻·第一季	变形金刚微型传说	超时空要塞Frontie
IDOLiSH7-偶像星愿-	RAILWARS-日本国有铁道公安队-	VividStrike	八月的棒球甜心	变形金刚银河之力	超时空要塞Macros
InfiniStratos2	Re: 从零开始的异世界生活 新编集版	VS骑士弹珠汽水40炎	白猫计划零之纪元	便当	超时空要塞ZERO
Infini-TForce	Regalia三圣星	WWW迷糊餐厅	白色相簿2	冰菓	超时空要塞△

<  >

4月14日第一次爬取:

大小: 271 MB (284,595,618 字节)

占用空间: 310 MB (325,726,208 字节)

包含: 26,035 个文件, 26,976 个文件夹

4月30日第二次爬取:

大小: 529 MB (555,043,668 字节)

占用空间: 611 MB (640,708,608 字节)

包含: 51,697 个文件, 31,306 个文件夹

存储格式

名称	修改日期
01_红豆上班了	2020/4/15 17:
02_干了这杯酒下辈子还做红豆	2020/4/15 17:
03_终于要被吃掉了	2020/4/15 17:
04_换个工作东山再起	2020/4/15 17:
05_豆生若只如初见	2020/4/15 17:
06_奇怪的同事	2020/4/15 17:
07_你好酷哦	2020/4/15 17:
08_青青草原上的豆砸	2020/4/15 17:
09_跳绳三缺一	2020/4/15 17:
10_某豆带头翘班	2020/4/15 17:
11_皮皮奶里皮皮豆	2020/4/15 17:
12_红豆放假啦	2020/4/15 17:

名称	修改日期	类型	大小
danmu_0.txt	2020/4/14 16:33	文本文档	125 KB
danmu_1.txt	2020/4/14 16:36	文本文档	127 KB
danmu_2.txt	2020/4/14 16:45	文本文档	122 KB
danmu_3.txt	2020/4/14 16:48	文本文档	122 KB
danmu_4.txt	2020/4/14 16:50	文本文档	131 KB
danmu_5.txt	2020/4/14 16:53	文本文档	132 KB
danmu_6.txt	2020/4/14 16:55	文本文档	128 KB
danmu_7.txt	2020/4/14 16:59	文本文档	125 KB
danmu_8.txt	2020/4/14 17:06	文本文档	127 KB
danmu_9.txt	2020/4/14 17:08	文本文档	134 KB
danmu_10.txt	2020/4/14 17:09	文本文档	46 KB

动漫 - 剧集 - 弹幕文件

文件大小接近

按照时间顺序增加弹幕文件

每条弹幕占据一行

1960	典型的转校生都是怪物
1961	别去救它
1962	好燃
1963	好燃
1964	温馨
1965	可怕吗
1966	害怕
1967	被发现了
1968	今晚刀谁都懂了吧
1969	巴拉能量
1970	这兔头肯定香
1971	战场原啊
1972	王水洗头
1973	橘势大好
1974	钉钉时代
1975	一袋漏糠机打米
1976	咂瓦鲁多
1977	战场原黑仪
1978	斋藤千和
1979	不要随便救陌生人啊
1980	熟练的让人心疼
1981	女神低语
1982	炖了
1983	已经不用害怕了
1984	画面与歌词严重不符
1985	强的一匹
1986	见渊进
1987	强势橘气
1988	UMB
1989	预言家
1990	互攻他不香吗
1991	BGM听着渗的慌
1992	对不起ue
1993	熟练得令人心疼
1994	八嘎呀路

项目计划

项目简介	所使用的方法或工具	预期完成时间
网页爬虫	Requests, RegExp	5.1
弹幕分词	Jieba分词, FP-Tree, KMP	5.4
建立倒排索引	SPIMI	5.5
检索方法	向量空间模型或概率似然模型	5.15
实现任务一：梗百科搜索	Tf-Idf or BM25 Top k rank 或查询似然检索	5.20
实现任务二：动漫推荐	Tf-Idf or BM25, k-means	5.25

分词方法

1. 分词原理

(1) 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG); (2) 采用了动态规划查找最大概率路径，找出基于词频的最大切分组合；(3) 对于未登录词，采用了基于汉字成词能力的HMM模型，使用了 Viterbi 算法。

2. 分词三种模式

(1) 精确模式，试图将句子最精确地切开，适合文本分析；(2) 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；(3) 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

分词方法

完结 撒花
 看 面包 的 眼神
 全体 起立
 朋也 你 怎么 了
 大乔 停止 了 思考
 兄弟 长椅 情
 黄金 精神
 不愧 是 乔鲁诺
 大乔 不 知道 应该 点赞 还是 沉默
 大乔 又 点 了 赞
 比如 压路机
 康一
 大乔为 你 点赞
 nice
 好 了 忘 了 这个 设定
 真就速 A 呀
 指 爽朗 的 偷 了 你 的 行李
 孔乙己 收到 一 张 好人 卡
 乔纳森 也 很 痛心
 父词 子哮
 压路机 可以 吗
 呼叫 由 花子
 孔乙己 你 长 荒木线 了
 民风 淳朴 意大利
 老汉 突然 出现
 大乔 点 了 个 踩
 笑 死
 吼吼
 请 忘 了 这个 设定
 是 艺术
 再 放送
 传统的 开局 先 捧 队友
 嘟噜 噜 噜 噜 噜 噜 噜 噜 噜 噜 噜

```

print ("/.join(jieba.cut("大乔为你点赞")))
jieba.add_word("大乔", freq = 10000, tag = None)
print ("/.join(jieba.cut("大乔为你点赞")))

```

大乔为/你/点赞
大乔/为/你/点赞

```

[60]: print(shorten("23333333"))
print(shorten("wryyyyyyyy"))
print(shorten("ohhhhhhhhhh"))
print(shorten("义勇啊啊啊啊啊"))
print(shorten("木大木大木大木大木大"))
print(shorten("你们有毒吧哈哈哈哈哈哈"))
print(shorten("得得得得得得得得得得得得得"))

```

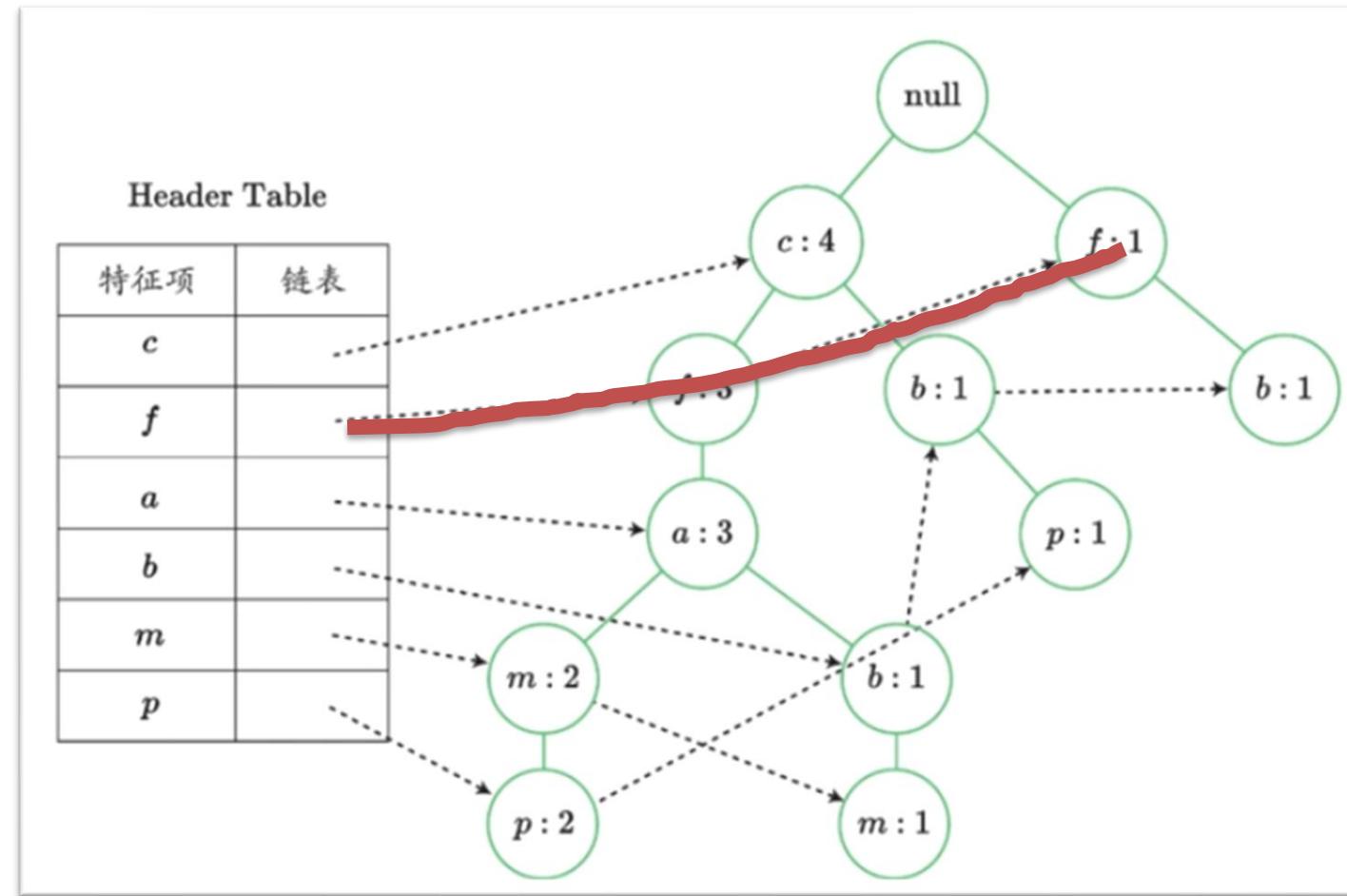
233
wryy
ohh
义勇啊啊
木大
你们有毒吧哈哈
得得

利用后缀数组，做叠词清洗

1. 叠词清洗
2. 转换为字典（弹幕，频率）
3. 收录较短高频词
4. 转换为（单字，频率）

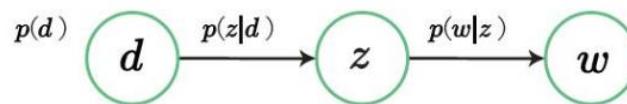
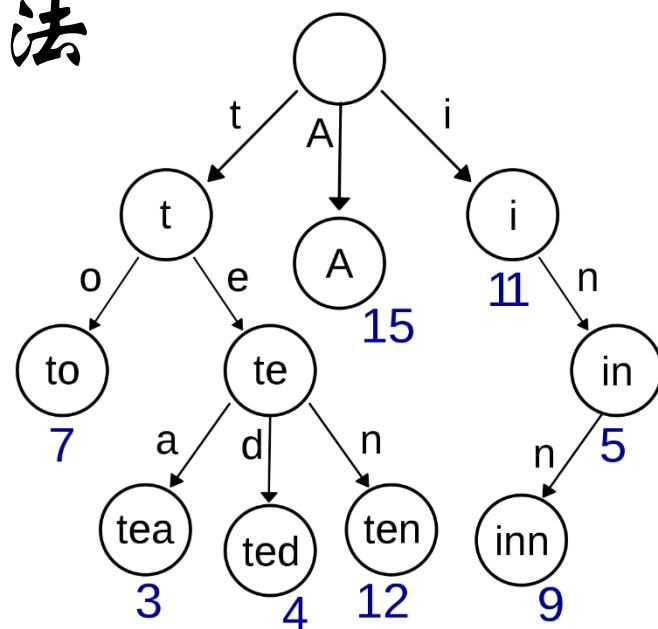
('卖炭翁', 40),
('哈哈', 37),
('百鬼丸', 36),
('满面尘灰烟火色', 35),
('泪目', 32),
('两鬓苍苍十指黑', 26),
('前方高能', 26),
('aws1', 25),
('flag', 22),
('犹豫就会败北', 19),
('工具人', 17),
('好看吗', 17),
('果断就会白给', 17),
('追了', 16),
('护食', 16),
('恶鬼灭杀', 16),

分词方法



Frequent Pattern Tree (FP Tree)

分词方法



$$S(\text{大}) = 305$$

$$S(\text{乔}) = 271$$

$$P(\text{乔}|\text{大}) = 202/305$$

$$P(\text{大}|\text{乔}) = 202/271$$

$$(1) S(a) > 10$$

$$(2) P(b|a) > 1.3 - 0.35\ln(S(b))$$

$$(3) P(a|b) > 0.6$$

```
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
不知	7
道	7
应	5
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

分词方法

```

留下jio印 留个jio印 0.3333333333333333
祝君武运昌 祝各位武运昌隆 0.2222222222222222
下jio印 留个jio印 0.375
下一季再见 期待下一季 0.2857142857142857
鸡心打开 几盆鸡心打 0.3333333333333333
祝武运昌 祝各位武运昌隆 0.25
君武运昌 祝各位武运昌隆 0.25
等下一季 期待下一季 0.3333333333333333
一季再见 期待下一季 0.1666666666666666
为什么 什么时候 0.25
再见了 下季再见 0.25
肝完了 天肝完 0.3333333333333333
个脚印 下脚印 0.3333333333333333
缘再见 下季再见 0.25

```

1. 子串去重
2. 2 gram 相似度去重
3. 导入词典
4. 去停用词, 分词

Result →

```
simplify(new_words)
```

```

['女装大佬', '珍爱上了', '糟糕的台', '下季再见', '盗梦空间', '错误示范', '一天肝完', '撩完就跑', '笑死我了', '下次一定', '美女你谁', '剧场版见', '猝不及防', '前方高能', '完结撒花', '撑不住了', '不娶何撩', '口气肝完', '逻辑鬼才', '到此一游', '优秀员工', '证明我来', '虎狼之词', '精神小伙', '感谢陪伴', '二十六集', '进度条', '下爪印', '香奈乎', '好可爱', '舍不得', '紫藤花', '为什么', '晒太阳', '我来过'],

```

倒排索引

类别 → 文档 → 词项

动漫 → 剧集 → 词项

$$w_{t,d} = (1 + \log tf_{t,d}) \times \log_{10}(N / df_t)$$

$$\text{Weighted Term Frequency} = \frac{(k + 1)tf_{td}}{k * ((1 - b) + b * (\text{len}(D) / \text{avg_doclen})) + tf_{td}}$$

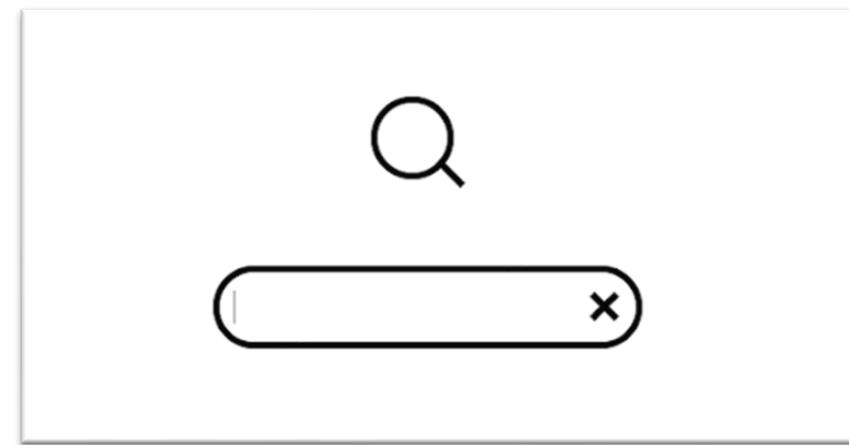
1960	典型的转校生都是怪物
1961	别去救它
1962	好燃
1963	好燃
1964	温馨
1965	可怕吗
1966	害怕
1967	被发现了
1968	今晚刀谁都懂了吧
1969	巴拉能量
1970	这兔头肯定香
1971	战场原啊
1972	王水洗头
1973	橘势大好
1974	钉钉时代
1975	一袋漏糠机打米
1976	咂瓦鲁多
1977	战场原黑仪
1978	斋藤千和
1979	不要随便救陌生人啊
1980	熟练的让人心疼
1981	古神低语
1982	炖了
1983	已经不用害怕了
1984	画面与歌词严重不符
1985	强的一匹
1986	见渊进
1987	强势橘气
1988	UMB
1989	预言家
1990	互攻他不香吗
1991	BGM听着渗的慌
1992	对不起ue
1993	熟练得令人心疼
1994	八嘎呀路

梗百科搜索

梗的意思指动画、电视剧等作品中喜闻乐见的桥段

特点：

1. 小众，代表着部分圈子文化
2. 源于一部作品，但会出现在多部作品弹幕中（玩梗）
3. 一般不符合常规语言句式，但复用性强
4. 传播效率高，有些具有时效性



火之意志 → 火影忍者疾风传

黄金精神 → JOJO的奇妙冒险

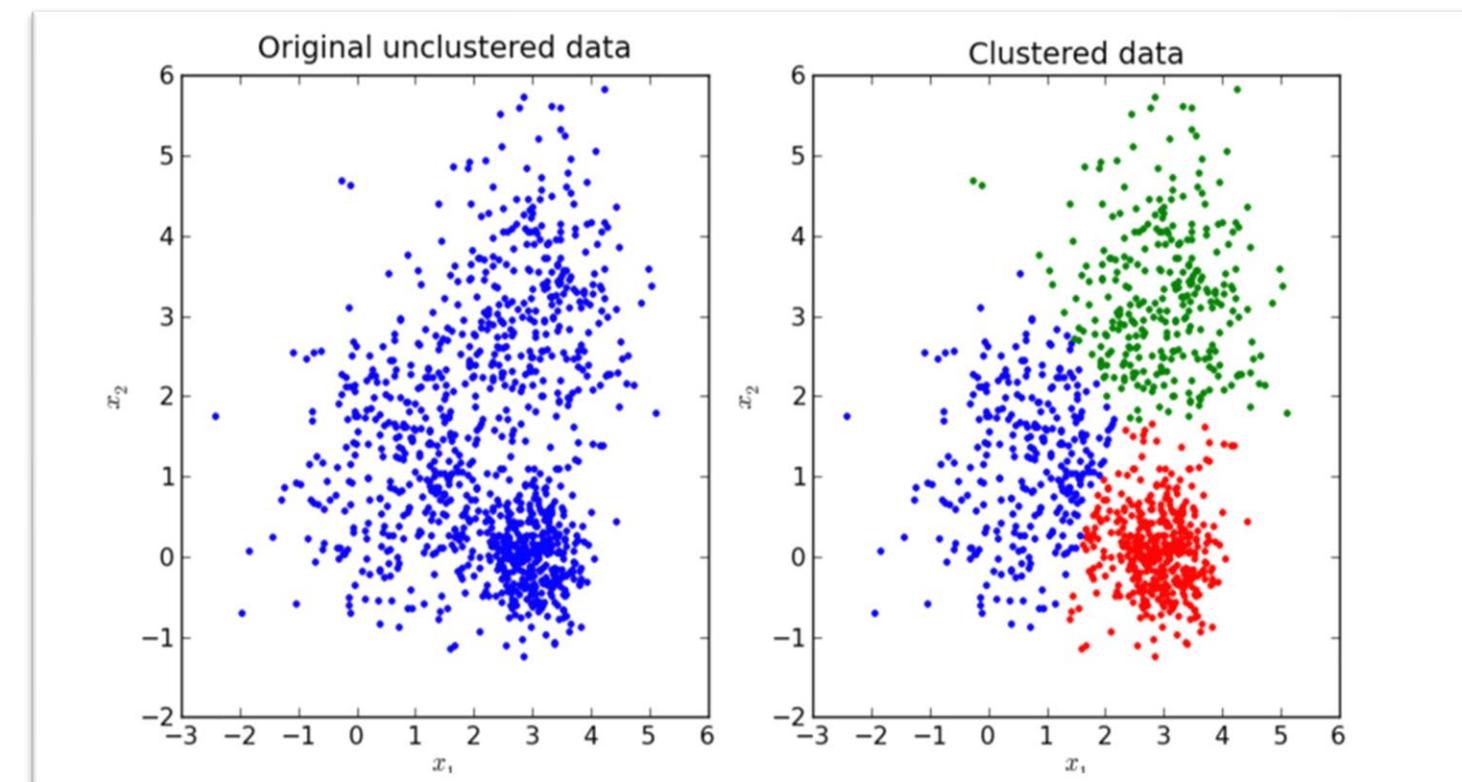
(符号化词汇)

动漫推荐

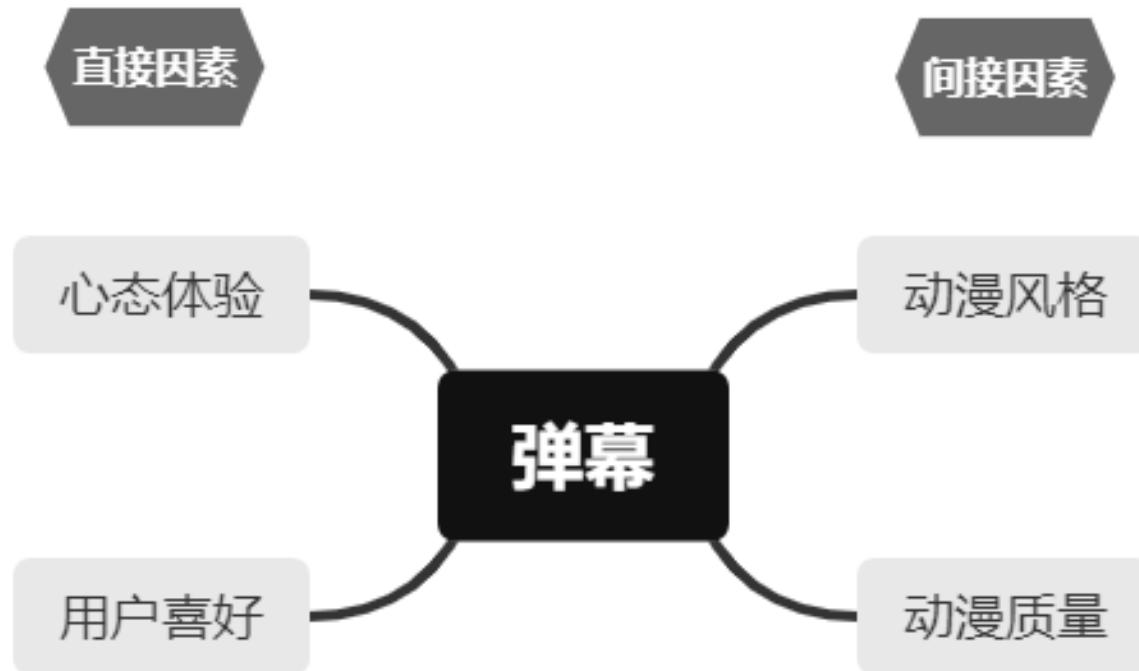
动漫间距离：用余弦相似度或杰卡德相似度刻画

数据存储：用单个文档表示一部动漫的高维向量

聚类方法：K-means



项目意义



对于视频平台：

1. 通过对弹幕的聚类分析，可以将动漫进行初分类。将用户近期观看的动漫或所发的弹幕进行相似度分析，平台可以刻画用户类型，并进行精准推荐。
2. 梗百科也可以作为弹幕文化的一部分促进平台社区文化，能让新用户更快了解社区文化，增加归属感。



哔哩哔哩弹幕网

Thanks

动漫弹幕的文本检索和应用

期末答辩

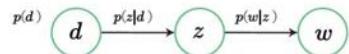
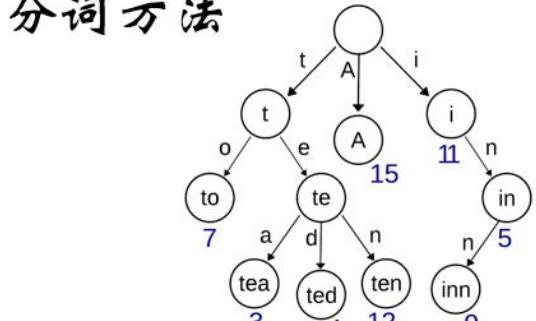
11811721 庄湛

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

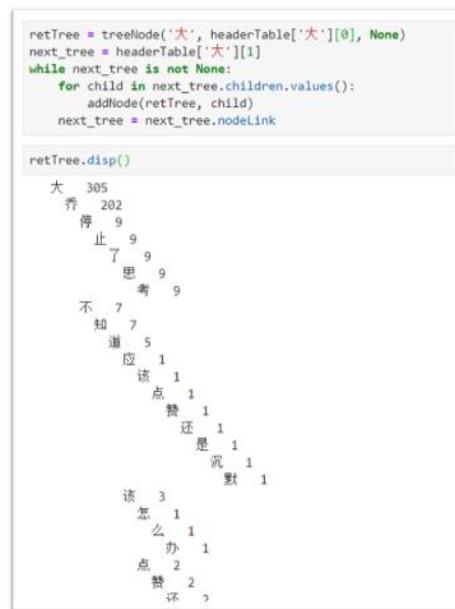


$$\begin{aligned} S(\text{大}) &= 305 \\ S(\text{乔}) &= 271 \\ P(\text{乔}|\text{大}) &= 202/305 \\ P(\text{大}|\text{乔}) &= 202/271 \end{aligned}$$

$$\begin{aligned} (1) \quad S(a) &> 10 \\ (2) \quad P(b|a) &> 1.3 - 0.35\ln(S(b)) \\ (3) \quad P(a|b) &> 0.6 \end{aligned}$$

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:

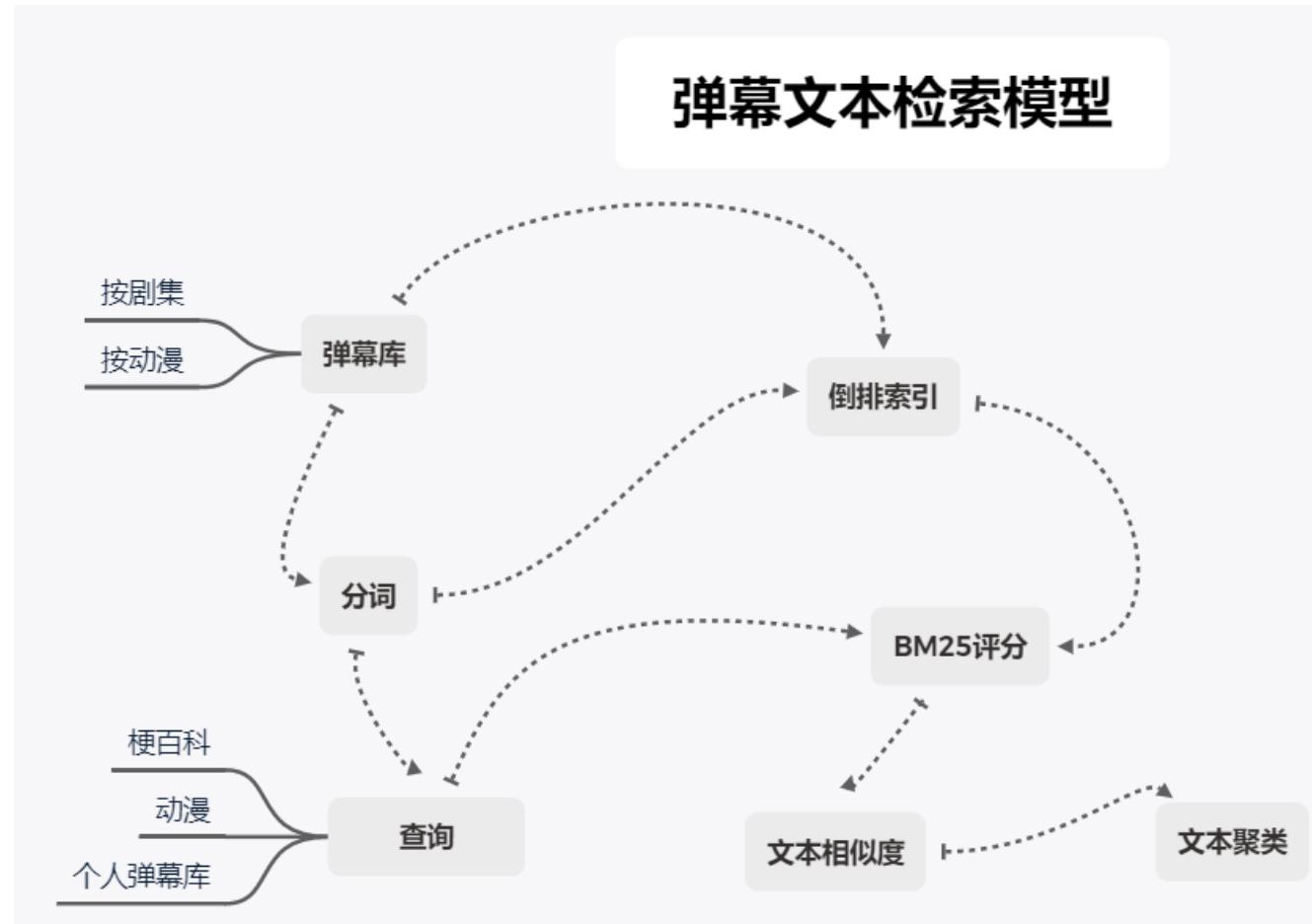
- 语料库总量大
- 单句精简短小
- 新词叠词较多
- 错别字词较多
- 词语较口语化
- 部分词项词频较高
- 不同语境下差异大

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。

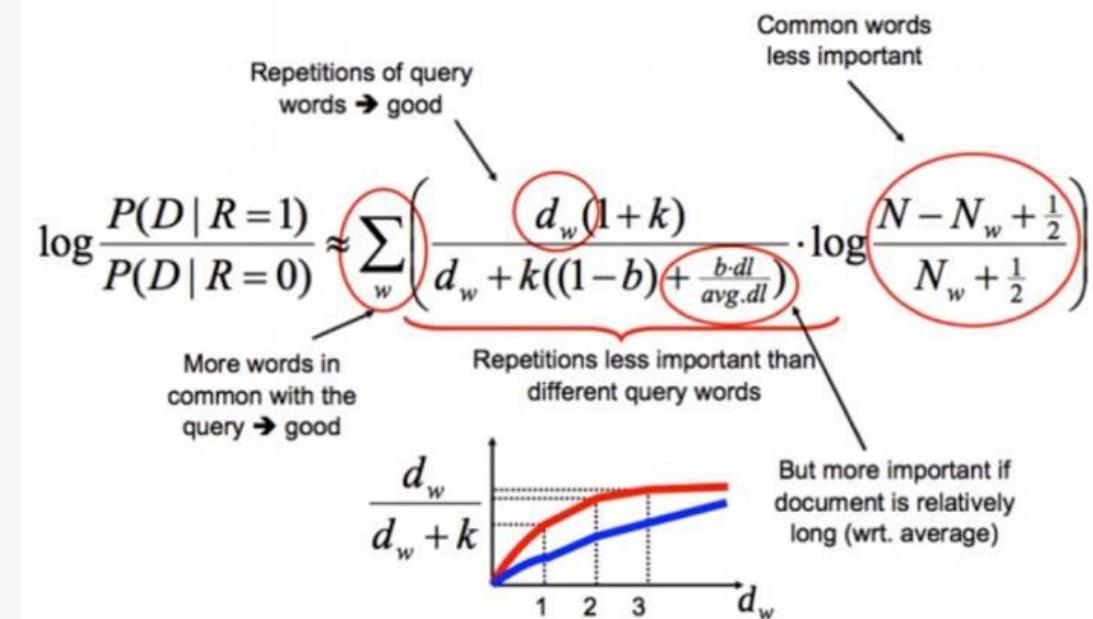
我认为这是一种**有灵魂的分词**, 和以往分析前后凝聚度, 最短路径以及**n-gram**分词法相比, 该方法利用了**语料特点**和**集体习惯**。当人们表达情绪和分享时, 往往利用较短而常用的词汇。在爬取得得到的五千九百多万条弹幕中, 有两千六百多万条弹幕长度小于5。平均弹幕长度远小于正常语料中两个标点间长度。并且由于错字较多、新词较多、较为庞大等特点, 我认为该类语料库有很高的研究价值, 对于这种新型文本的研究可以推动统计自然语言处理学科的发展。

梗百科：



"梗"通常指动漫中一些喜闻乐见的桥段

BM25: an intuitive view



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。

信息检索项目

① 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/... A (1) Bookmarks - A...

其他书签

信息检索项目

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_才虎财阔的豪华海上观光
- OneRoom第二季\01_花坂结衣的序幕
- D r . S T O N E 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”

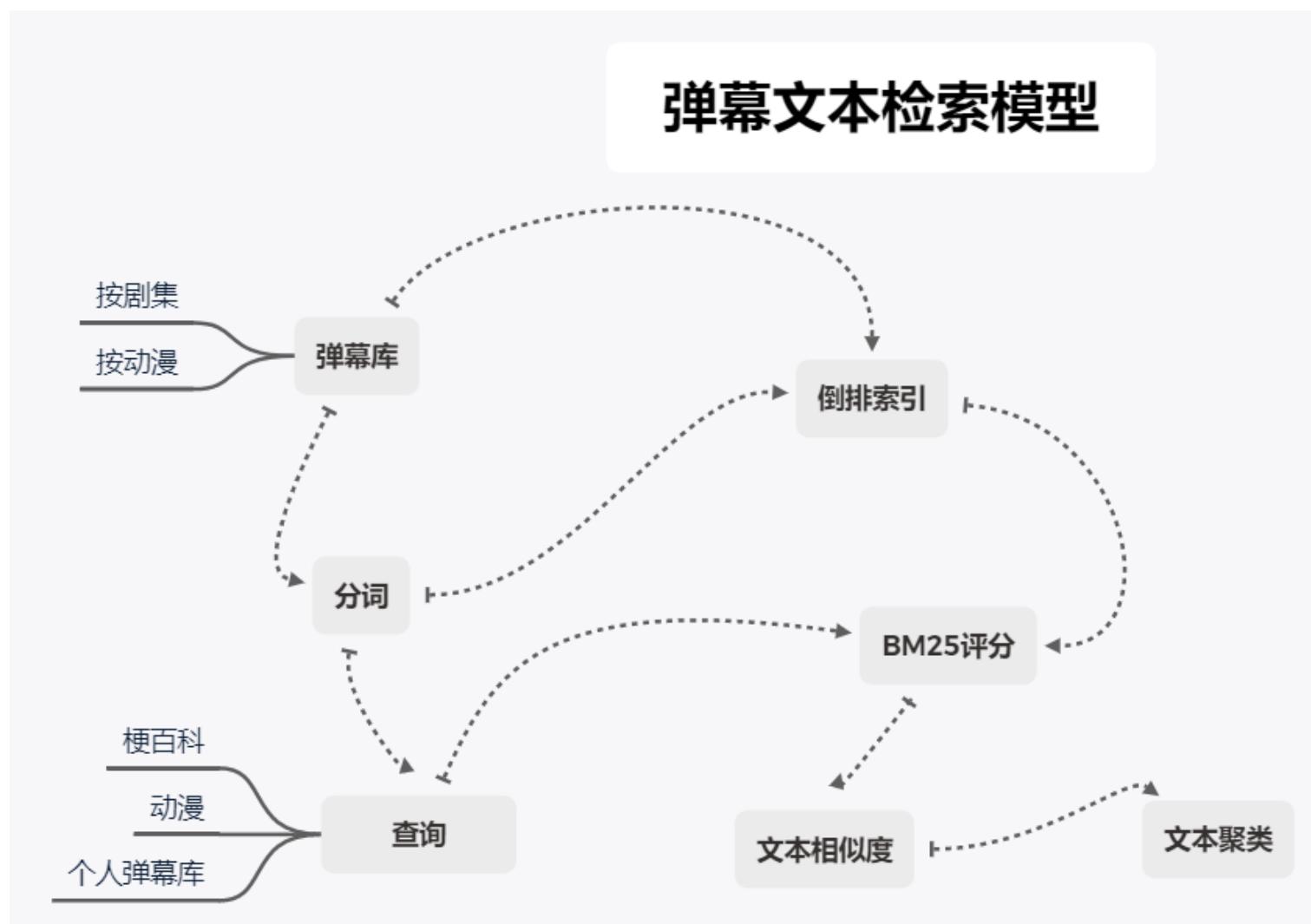




项目意义和价值：

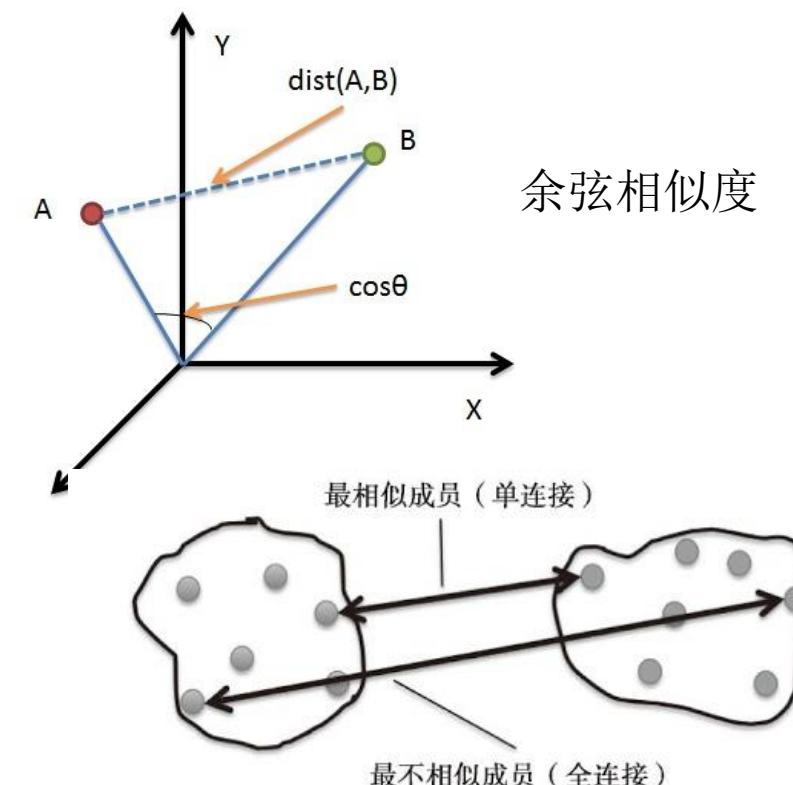
1. 提供不了解“梗文化”的人，一个快速了解的途径，可以很有趣地吸引圈外人士。
2. 给喜欢玩梗的人一个补番推荐或者玩梗指南，让“梗”这种亚文化更有活力。
3. 分析“玩梗”、“造梗”等动漫文化表达出的观众对动漫的期望来研究更多商业价值。以及提供一种研究不同动漫间的联系的途径。





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

(动漫总数, 词项总数)





[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]

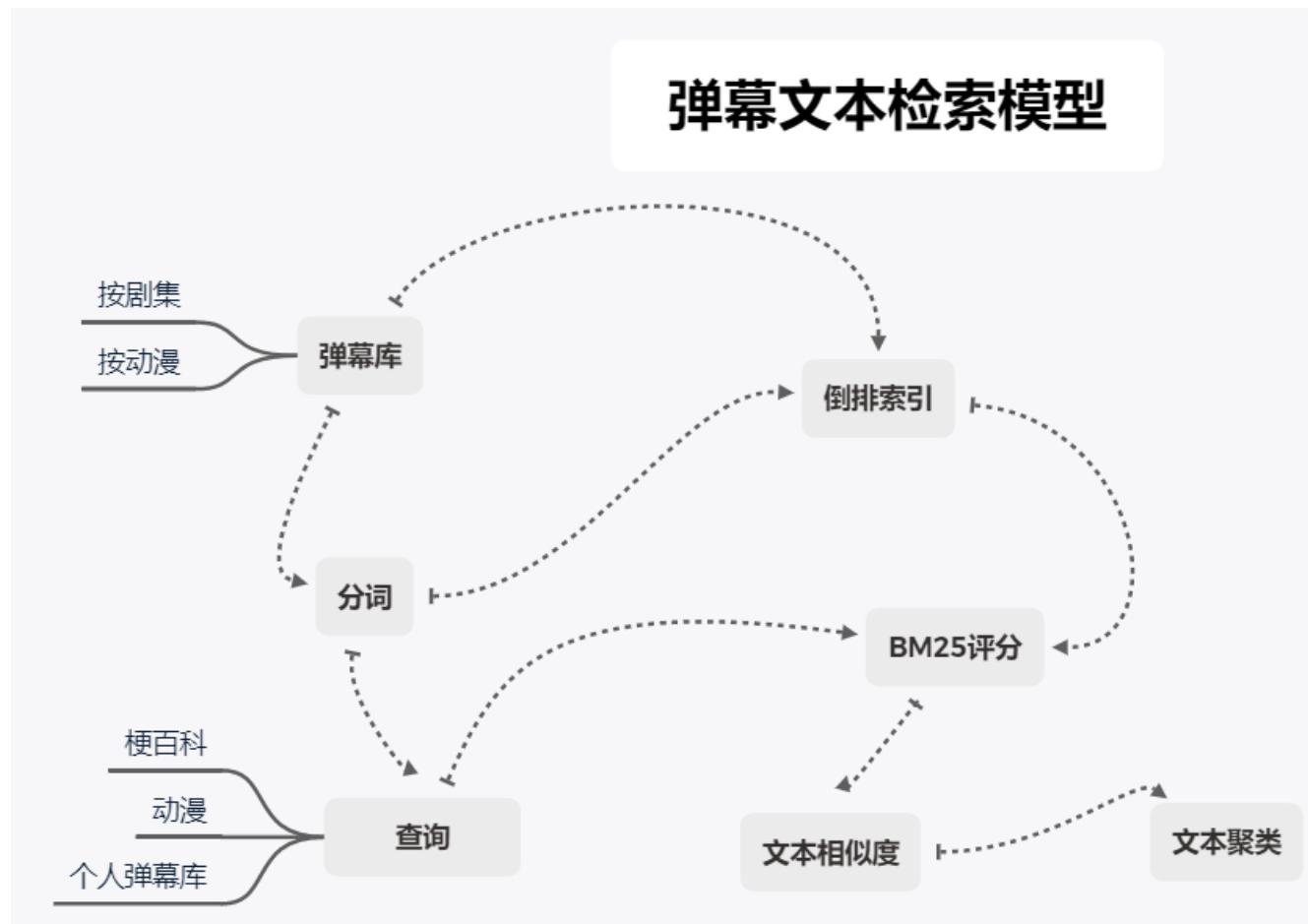
思考：如果将一个人所发的全部弹幕做成一个弹幕集当做一部“动漫”，也可以以此做更加精准的动漫推荐。

[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]

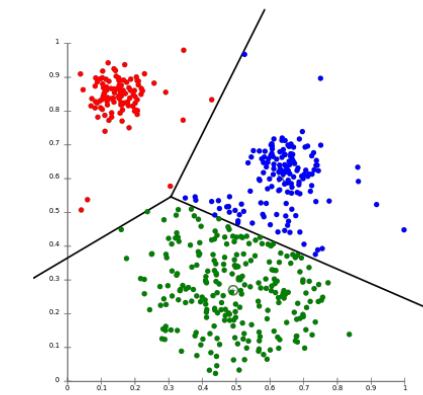




非连续增加惩罚

```

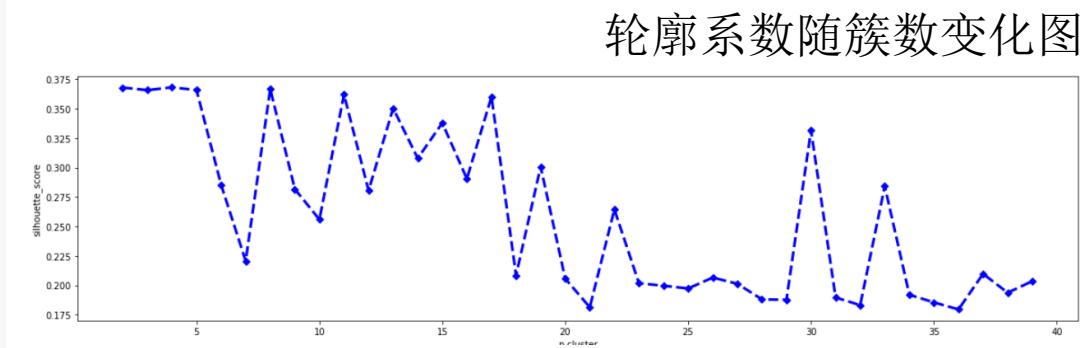
if len(clust_1)>100 and len(clust_2)> 100:
    dis*=1.4
elif len(clust_1)>50 and len(clust_2)> 50:
    dis*=1.2
elif len(clust_1)>30 and len(clust_2)> 30:
    dis*=1.1
elif len(clust_1)>10 and len(clust_2)> 10:
    dis*=1.05
  
```



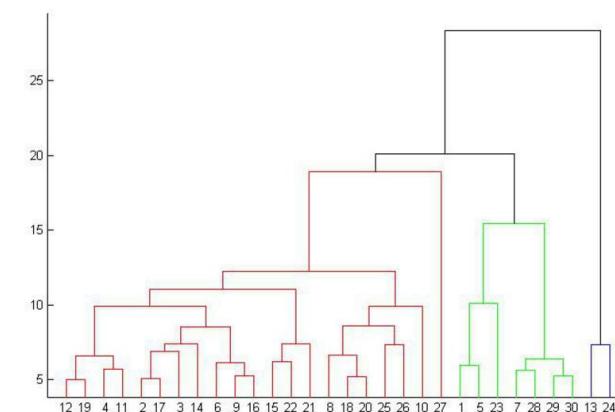
K-Means

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,百合魔风暴,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☆歌舞剧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巴哈姆特之怒玛纳利亚魔法学院,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,小鳩,遥远时空八叶抄,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,四叠半神话大系,有顶天家族,RWBY,拽妹黛薇儿第一季,草莓公园第一季,源氏物语千年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

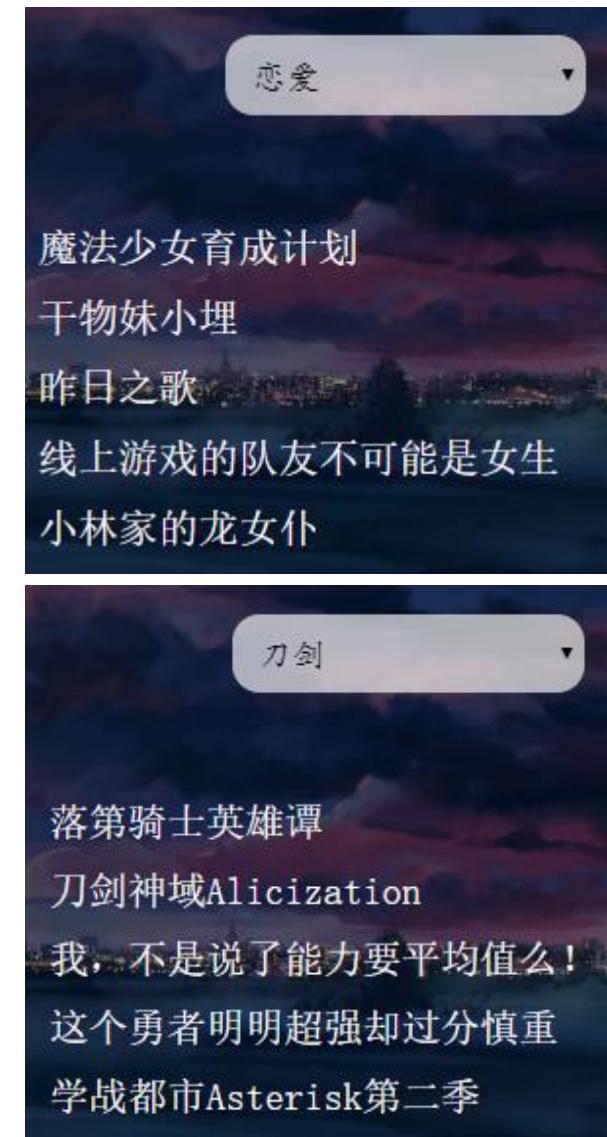
category	animations
国创	{0号宿舍, 我家大师兄脑子有坑特别篇, 斩兽之刃, 凸变英雄LEAF, 萌妻食神, 通灵妃, 剑网3·侠肝义胆沈剑心}
少女、公主、精灵啥	{11eyes, 出租魔法使, Caligula卡里古拉, 境界线上的地平线, 神不在的星期天, CDE, 黑神, [C]THEMO}
偶像	{227, 神推偶像登上武道馆我就死而无憾, ANIMAYELL, LoveLiveSchoolIdolProject, LoveLive}
不确定	{ACCA13区监察课, 江户盗贼团五叶, GANGSTA, 青春生存游戏, DaDaDa, 天使怪盗, 寻找满月, 萩萩公主, }
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, 中二病也要谈恋爱爱, 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。



信息检索项目

介绍 梗百科 动漫推荐 动漫聚类

项目介绍

本项目通过对一千余部动漫弹幕建立词袋模型的倒排索引，并主要利用BM25向量空间模型对检索进行打分，利用层次聚类和K-means进行聚类。通过打分机制和文本间距离的计算，我们设计了三个主要功能：

1. “梗”通常指动漫中一些喜闻乐见的桥段，你了解哪些“梗”呢？在梗百科中，你可以快速得到你想要查询的“梗”的相关出处（会有很多惊喜的）。
2. 每个人都有独特的审美，你了解你的看番风格嘛？在动漫推荐中，我们会为您推荐最属于你看番风格的动漫（这都取决于你们发的弹幕的）。
3. 你喜欢看什么类型的动漫呢？这或许隐藏在你爱发的弹幕中…在动漫聚类中，我们利用动漫间的余弦相似度进行聚类，并简单命名来达到分类的效果。

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

技术栈



HTML



CSS



```
141
142     def jiebaUtil(conn):
143         cursor = conn.cursor()
144         query = """select * from postgres.ir.raw_words"""
145         cursor.execute(query)
146         new_words = cursor.fetchall()
147         for key_word in new_words:
148             if key_word[1] > 2:
149                 jieba.add_word(key_word[0], freq=key_word[1] * 10, tag=None)
150
151
152     def jiebaUtil_file():
153         new_words = []
154         f = open("static\\new_words_all.txt")
155         lines = f.readlines()
156         for line in lines:
157             split = line.strip().split(',')
158             new_words.append((split[0], int(split[1])))
159         f.close()
160         for key_word in new_words:
161             if key_word[1] > 2:
162                 jieba.add_word(key_word[0], freq=key_word[1] * 10, tag=None)
```



感谢观看

开发环境: <http://127.0.0.1:8000/>

生产环境: <http://140.143.30.217:8008/>