1. **数据准备**
2. 爬取网易云音乐共计39013首歌曲信息，筛选出其中热度（评论数）较高的1336首歌曲。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NO. | ID | NAME | SINGER | ALBUM | HOTDEGREE |
| 1 | 186016 | 晴天 | 周杰伦 | 79 | 2023576 |
| 2 | 411214279 | 雅俗共赏 | 许嵩 | 1007 | 1102188 |
| 3 | 471385043 | 暧昧 | 薛之谦 | 635 | 367449 |
| 4 | 418603077 | 告白气球 | 周杰伦 | 321 | 367316 |
|  | … | … | … | … | … |
| 1336 | 405377706 | 想着你的感觉 | 容祖儿 | 1224 | 10054 |

1. 根据已筛选的热门歌曲列表，在新浪微博中爬取分享对应歌曲的微博。筛选出微博数量最多的前750首歌曲微博作训练集。

例如：

{250, 29814898, '可惜没如果', '林俊杰', 71721}：

|  |  |  |  |
| --- | --- | --- | --- |
| NO. | ID | CONTENT | DATE |
| 1 | 29814898 | 该好好和过去道别了～ | 2017-12-29 21:54:16 |
| 2 | 29814898 | 生活总是会给你很多的‘惊喜’，比如走到悬崖边上的时候发现后面的路也没了 | 2017-12-29 21:30:52 |
| 3 | 29814898 | 上帝总是让你以失去什么的方式来让你知道他有多重要 | 2017-12-29 08:56:07 |
| … | … | … | … |
| 1000 | 29814898 | 要说最悲惨的错过，莫过于没有开始已经错过。 | * + 1. 2:26:36 |

1. 爬取750首歌曲的歌词。
2. **数据预处理**
3. FACE++返回的七个情绪为{愤怒, 厌恶, 恐惧, 喜悦, 平静, 悲伤, 惊讶}，通过哈工大信息检索研究中心同义词词林将每一个情绪拓展为词组。

M=

{

[愤怒, 恼怒, 含怒, 气恼, 气愤, 义愤, 激愤, 愤激, 愤慨, 愤然, 愤愤, 悻悻, 愠, 恚, 怒, 气, 恼, 愤, 怒冲冲, 气冲冲, 气乎乎, 气鼓鼓, 恼羞成怒, 怒气冲冲, 怒气攻心, 气哼哼, 懑, 怒目横眉, 忿, 慨, 气呼呼, 愤闷, 生闷气, 一怒之下],

[讨厌, 厌恶, 嫌恶, 憎恶, 痛恶, 恶, 厌, 憎, 嫌, 烦, 厌烦, 腻烦, 腻味, 腻, 头痛, 作呕, 倒胃口, 深恶痛绝, 疾首蹙额, 掩鼻而过, 看不顺眼, 看不惯],

[害怕, 惧怕, 生怕, 怕, 生恐, 畏怯, 畏惧, 恐惧, 疑惧, 恐怖, 胆寒, 忌惮, 畏葸, 惶惑, 失色, 咋舌, 丧胆, 惮, 惧, 畏, 怖, 亡魂丧胆, 闻风丧胆, 丧魂落魄, 魂不附体, 魂飞魄散, 魄散魂飞, 战战兢兢, 惊心掉胆, 毛骨悚然, 提心吊胆, 胆颤心惊, 心惊胆颤, 心惊肉跳, 心胆俱裂, 不寒而栗, 噤若寒蝉, 惊恐万状, 大惊失色, 望而生畏, 担惊受怕, 胆破心惊, 胆战心惊, 心惊胆战, 悬心吊胆, 令人心悸, 悚, 慑],

[高兴, 开心, 愉快, 欢快, 称快, 快活, 快乐, 欢乐, 欢娱, 欢愉, 欢欣, 欢喜, 喜欢, 欣喜, 欣然, 怡然, 陶然, 愉悦, 融融, 乐意, 乐, 欢, 喜, 快, 欣, 怡, 乐滋滋, 乐呵呵, 乐悠悠, 乐融融, 美滋滋, 欢欣鼓舞, 赏心悦目, 如获至宝高高兴兴, 逸乐, 欣悦, 悦, 欣欣然, 快快乐乐, 欢欢喜喜, 欢悦, 喜悦, 其乐融融, 为之一喜],

[心静, 平静, 安静, 宁静, 恬静, 熨帖, 坦然, 安然, 释然, 恬然, 平心静气, 心平气和, 安安静静, 沉心静气, 少安毋躁], [悲伤, 伤悲, 伤心, 伤感, 熬心, 难过, 难受, 悲, 哀, 哀伤, 悲哀, 悲戚, 悲愁, 哀愁, 哀戚, 悲怆, 忧伤, 凄然, 凄怆, 凄恻, 凄惶, 殷殷, 不是味儿, 如丧考妣, 凄风楚雨, 同悲, 可悲, 不好过],

[惊奇, 奇怪, 惊异, 诧异, 奇异, 骇异, 纳罕, 希罕, 咋舌, 惊讶, 惊诧, 骇怪, 怪, 异, 奇, 诧, 惊愕, 愕然, 骇然, 大惊小怪, 讶异, 惊呆, 惊叹, 叹观止矣, 好奇]

}

共计N=216个词。

1. 利用jieba分词将每首歌的歌词、微博进行分词，得词组D。
2. **推荐算法**
3. **使用TF-IDF算法计算每首歌的分词结果，取相关性最高的Top(10\*N)。**

TF-IDF（term frequency–inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。

词频（term frequency，TF）指的是某一个给定的词语在该文件中出现的频率

https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D220/sign=c690d729abec8a13101a50e2c7019157/5ab5c9ea15ce36d3448570f638f33a87e850b177.jpg

逆向文件频率（inverse document frequency，IDF）是一个词语普遍重要性的度量：

https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D220/sign=2b6544f458afa40f38c6c9df9b65038c/a8014c086e061d9552eec4fe79f40ad163d9ca48.jpg

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语：

https://gss2.bdstatic.com/9fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D220/sign=173521e792ef76c6d4d2fc29ad14fdf6/241f95cad1c8a7863bffb9c86509c93d71cf501b.jpg

1. **使用google-word2vec训练中文维基百科(Wikipedia)语料库，生成向量空间模型，将中文词汇转换为向量表示。**

word2vec是谷歌的一个开源NLP工具，利用神经网络，将自然语言中的字词转为计算机可以理解的稠密向量（Dense Vector）。

word2Vec可以将One-Hot Encoder转化为低维度的连续值，也就是稠密向量，并且其中意思相近的词将被映射到向量空间中相近的位置。

1. **计算歌曲与情绪的相关度**

对于每一首歌、每一种情绪，计算每一个情绪词汇与这首歌所有相关词最高的余弦相似度S，取S的平均值作为此情绪与该歌曲的相关度。

词向量W1,W2的余弦相似度cosS(W1, W2)：

歌曲s与情绪词w的相关度S(s, w)：

取S的平均值作为此情绪与该歌曲的相关度：

对每一首歌曲，求出每一种情绪与之相对的相关度：

计算出每个情绪相关性最高的Top30歌曲，存入推荐音乐库。

1. 每一张照片返回的情绪值，是对应七种情绪的七个浮点数，和为100.0，对七个数值分别除以10并向下取整，得到每种情绪需推荐歌曲的数目，从推荐歌曲库相应情绪中取出对应数目的歌曲，返回歌曲列表给前端播放器。