Atitit 歌词成语提取项目 nlp 人工智能项目

目录

[1.1. 流程 首先搜集3w成语词库 1](#_Toc25728)

[1.2. 歌词常用成语400个 按照拼音排序 1](#_Toc2622)

## 流程 首先搜集3w成语词库

放入mmseg测试生效。。

然后遍历700歌词。。获取所有词语。。

然后获取四字词语。。大概416个。。

lyriccheniuExtra.xlsx

可以使用xls自定义文字长度函数len获取。。排序，

去除非成语的四字词语（人工处理

## 歌词常用成语400个 按照拼音排序

阿拉斯加,爱不释手,安非他命,安然无恙,黯淡无光,拔地而起,白头偕老,百里挑一,保持一致,悲欢离合,蹦蹦跳跳,逼不得已,闭月羞花,遍体鳞伤,波涛汹涌,不分彼此,不甘示弱,不顾一切,不好意思,不可或缺,不可收拾,不可思议,不留余地,不能自己,不切实际,不闻不问,不屑一顾,不知不觉,不知去向,不知人间,不知所措,布达拉宫,步履轻盈,擦肩而过,曾经沧海,柴米油盐,超然物外,沉鱼落雁,陈词滥调,成千上万,初恋情人,垂涎欲滴,春暖花开,春去秋来,春夏秋冬,此时此刻,从此以后,从今以后,从来没有,从左到右,寸步不离,大大方方,大摇大摆,大雨滂沱,荡气回肠,得过且过,灯红酒绿,灯火辉煌,地久天长,地老天荒,颠沛流离,点点滴滴,跌跌撞撞,顶天立地,定时炸弹,东南西北,洞房花烛,独守空房,独一无二,独自一人,发明创造,发自内心,翻山越岭,放在心上,飞蛾扑火,分崩离析,分分秒秒,粉身碎骨,奋不顾身,风吹雨打,风风雨雨,风光旖旎,风花雪月,赴汤蹈火,覆水难收,隔岸观火,觥筹交错,孤孤单单,孤身一人,鼓起勇气,鬼迷心窍,国色天香,过眼云烟,海角天涯,海枯石烂,海誓山盟,含糊其词,含情脉脉,行色匆匆,毫不迟疑,好不容易,好景不常,何去何从,和睦相处,黑白电视,轰轰烈烈,忽左忽右,胡思乱想,花好月圆,花花公子,花花世界,恍如隔世,灰飞烟灭,回头是岸,回心转意,浑然天成,或多或少,鸡皮疙瘩,唧唧喳喳,急不可耐,既然如此,坚不可摧,竭尽全力,今生今世,金榜题名,紧急情况,尽管如此,近在咫尺,惊慌失措,晶莹剔透,纠缠不休,卷土重来,可不可以,可有可无,刻骨铭心,坑蒙拐骗,控制不了,口口声声,口是心非,哭笑不得,来日方长,浪费时间,老老少少,泪流满面,泪如雨下,理所当然,历经沧桑,临阵脱逃,麻木不仁,满不在乎,漫不经心,慢慢腾腾,没头没脑,没完没了,每时每刻,美好时光,美景良辰,蒙娜丽莎,梦寐以求,迷惑不解,迷迷糊糊,迷失方向,密密麻麻,面带笑容,命中注定,漠不关心,目眩神迷,那个女人,男男女女,难分难舍,难舍难分,难以承受,难以忘怀,难以置信,内心深处,内心世界,你死我活,宁静致远,努力奋斗,藕断丝连,噼里啪啦,翩翩起舞,漂洋过海,飘飘欲仙,平平淡淡,迫不急待,杞人忧天,千方百计,千家万户,千千万万,千山万水,千言万语,前尘往事,前所未有,卿卿我我,情不自禁,情非得已,晴天霹雳,冉冉升起,热血沸腾,人山人海,任劳任怨,日日夜夜,柔情蜜意,如痴如醉,如此这般,如何是好,如荼如火,如意算盘,若即若离,三心二意,三心两意,山崩地裂,伤痕累累,稍纵即逝,身不由己,什么时候,神魂颠倒,生生世世,声势浩大,声嘶力竭,失魂落魄,十八世纪,十字街头,十字路口,时时刻刻,世界各地,似水流年,视而不见,手足无措,谁是谁非,司空见惯,死得其所,死里逃生,死心塌地,四处奔波,四海为家,随便说说,随波逐流,随心所欲,随遇而安,岁月蹉跎,所有的人,所作所为,谈何容易,坦白从宽,螳臂当车,天崩地裂,天各一方,天花乱坠,天荒地老,天马行空,天南地北,天上人间,天时地利,天涯海角,天摇地动,天衣无缝,天造地设,天长地久,甜言蜜语,铁树开花,童话故事,头晕目眩,推卸责任,弯弯曲曲,完美无缺,万劫不复,万紫千红,忘乎所以,望穿秋水,委曲求全,未尝不可,文化传媒,我们自己,乌兰巴托,无边无际,无地自容,无动于衷,无话不谈,无话可说,无计可施,无可救药,无可奈何,无能为力,无情无义,无人问津,无声无息,无时无刻,无所不知,无所适从,无怨无悔,舞台灯光,误入歧途,吸取教训,熙熙攘攘,喜怒哀乐,下个星期,闲言碎语,现实生活,相见恨晚,相亲相爱,小心翼翼,携手前进,谢天谢地,心烦意乱,心甘情愿,心花怒放,心满意足,心如刀割,心如止水,心心相印,心有灵犀,信誓旦旦,星火燎原,星球大战,惺惺相惜,形影不离,兄弟姐妹,虚情假意,虚张声势,绚丽多彩,烟消云散,言不由衷,炎炎夏日,眼花缭乱,阳光明媚,遥遥无期,夜深人静,一悲一喜,一草一木,一成不变,一错再错,一刀两断,一个男孩,一举一动,一贫如洗,一颦一笑,一腔热血,一去不复,一如既往,一生一世,一往情深,一望无际,一无是处,一无所知,一席之地,一夜之间,依然如故,异想天开,意想不到,引吭高歌,勇往直前,犹豫不决,有权有势,有生以来,有限公司,有缘无份,于事无补,雨打风吹,玉树临风,越来越少,越来越小,越陷越深,在世界上,在所不惜,在所难免,在我心中,怎样才能,展翅高飞,辗转反侧,占为己有,真命天子,真心真意,整个世界,正在进行,支离破碎,只不过是,直截了当,终身大事,周而复始,装模作样,自愧弗如,自始至终,自言自语,自由自在,自作聪明,祖祖辈辈,最重要的,醉生梦死,昨天晚上,左右为难,

/sumdoclist/src/aPkg/segmentMMsegExtraCheniuFromLyric.java

**package** aPkg;

**import** java.io.File;

**import** java.io.IOException;

**import** java.io.StringReader;

**import** java.util.Set;

**import** org.apache.commons.io.FileUtils;

**import** com.alibaba.fastjson.JSON;

**import** com.chenlb.mmseg4j.ComplexSeg;

**import** com.chenlb.mmseg4j.Dictionary;

**import** com.chenlb.mmseg4j.MMSeg;

**import** com.chenlb.mmseg4j.Seg;

**import** com.chenlb.mmseg4j.Word;

**import** com.google.common.base.Joiner;

**import** com.google.common.collect.Sets;

**public** **class** segmentMMsegExtraCheniuFromLyric {

**public** **static** **void** main(String[] args) **throws** IOException {

Dictionary dic = Dictionary.*getInstance*();

Seg seg = **new** ComplexSeg(dic);

Set<String> set=Sets.*newLinkedHashSet*();

String dir="D:\\notdel\\music lyric zip v3 s525";

File[] fa=**new** File(dir).listFiles();

**int** file\_Indx = 0;

**for** (File f : fa) {

file\_Indx++;

System.***out***.println("file\_Indx"+file\_Indx);

String txt="受一股来自中西伯利亚的强冷空气影响，本市出现大风降温天气分节阅读辅以假期";

txt=FileUtils.*readFileToString*(f);

System.***out***.println(txt);

MMSeg mmSeg = **new** MMSeg(**new** StringReader(txt), seg);

Word word = **null**;

System.***out***.println();

**while**((word=mmSeg.next())!=**null**) {

set.add(word.getString());

// System.out.println(JSON.toJSONString(word));

// System.out.print(word.getString()+" -> "+word.getStartOffset());

// //offset += word.length;

// System.out.println(", "+word.getEndOffset()+", "+word.getType());

//

//

}

}

String s=Joiner.*on*("\r\n").join(set);

FileUtils.*write*(**new** File("D:\\lyricCheniuSet.txt"), s);

}

}