零散的思路：

1. 如何获取相关语料？爬虫+手工，多途径知乎+微博+豆瓣+浏览器的搞笑版块+all想到的以及用户投稿
2. 对爬取下来的语料，如何利用策略打分、打标签等处理；对用户的投稿如何识别，系统审核，后期开放审核权限；
3. 对内容的处理：

鉴别一条内容的初始原创性，获取授权；

无法授权，则删除；

1. 类似今日头条的个性化推荐，根据用户喜欢的领域、标签、浏览记录、喜好行为等推荐合适的内容；
2. 用户注册时不要太繁琐，包含基本的邮箱、用户名、密码即可，创建等级机制，靠什么升级？给予默认头像
3. 和脱口秀大会能否进行某种关联

User表：存储用户信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段说明 | 字段名 | 字段类型 | 是否必填 |
| 用户编号 | id | char(21) | 是 |
| 登陆账号 | login\_account | varchar(256) | 是 |
| 密码 | password | varchar(64) | 是 |
| 用户名 | user\_name | varchar(64) | 是 |
| 性别 | sex | char(2) | 否 |
| 电子邮箱 | email\_address | varchar(128) | 否 |
| 手机号码 | phone\_number | varchar(32) | 否 |
| 签名 | signature | varchar(256) | 否 |
| 年龄 | age | int(3) | 否 |
| 兴趣 | interest\_area | varchar(256) | 否 |
| 注册时间 | register\_time | datetime | 是 |
| 最后一次登陆时间 | last\_login\_time | datetime | 否 |
| 注册IP地址 | register\_ip | varchar(15) | 否 |
| 最后一次登陆Ip | last\_login\_ip | varchar(15) | 是 |
| 账户是否可用 | available | smallint(1) | 是 |
| 头像 | avatar | varchar(128) | 否 |
| 所在省份 | province | varchar(32) | 否 |
| 所在城市 | city | varchar(32) | 否 |
| 用户角色 | role | varchar(16) | 是 |

内容信息表：Duanzi

|  |  |  |  |
| --- | --- | --- | --- |
| 字段说明 | 字段名 | 字段类型 | 是否必填 |
| 段子编号 | d\_id | char(21) | 是 |
| 段子标题 | d\_title | varchar(128) | 否 |
| 段子内容 | d\_content | text | 是 |
| 发布时间 | publish\_time | datetime | 是 |
| 发布人编号 | publisher\_id | char(21) | 是 |
| 发布人名称 | publisher\_name | varchar(128) | 是 |
| 是否包含图片 | has\_image | smallint | 是 |
| 内容类型 | d\_type | varchar(64) | 是 |
| 内容标签 | d\_tag | varchar(64) | 否 |
| 该条内容的浏览次数 | d\_brower\_num | int(10) | 是 |
| 该条内容的点赞总数 | d\_nice\_num | int(10) | 是 |
| 该条内容的不喜欢总数 | d\_unnice\_num | int(10) | 是 |
| 该条内容的评论数 | d\_comment\_num | int(10) | 是 |
| 该条内容的分享次数 | d\_share\_num | int(10) | 是 |
| 评分 | d\_score | double | 是 |
| 该条内容的收藏数 | d\_collection\_num | int(10) | 是 |
| 内容是否敏感 | d\_sensitive | smallint(1) | 是 |
| 内容是否被举报 | d\_report | smallint(1) | 是 |
| 内容是否不可用 | d\_available | smallint(1) | 是 |

评论表:Comments

|  |  |  |  |
| --- | --- | --- | --- |
| 字段说明 | 字段名 | 字段类型 | 是否必填 |
| 评论编号 | c\_id | char(21) | 是 |
| 评论内容 | c\_content | varchar(1024) | 是 |
| 评论时间 | c\_time | datetime | 是 |
| 评论人编号 | c\_user\_id | char(21) | 是 |
| 评论的段子编号 | c\_d\_id | char(21) | 是 |
| 评论楼层 | c\_floor | int(10) | 是 |
| 评论是否是回复 | c\_is\_reply | smallint(1) | 是 |
| 评论回复的用户编号 | c\_reply\_user\_id | varchar(64) | 否 |
| 评论回复的用户楼层 | c\_reply\_user\_floor | int(10) | 否 |
| 评论点赞数 | c\_nice\_num | int(10) | 是 |
| 评论是否敏感 | c\_sensitive | smallint(1) | 是 |
| 评论是否可用 | c\_available | smallint(1) | 是 |

举报表

标签表

一周精彩段子表（每时更新）