关于毕业设计的设计（2019023）

1 事前的调研

1.1 相关文献

1. https://www.geekpark.net/news/232310
2. 一种面向新闻网站的事件跟踪方法
3. 网络舆情热点发现与事件跟踪技术研究
4. 微博热点话题检测与跟踪技术研究_吕伟
5. 基于大数据的互联网热点话题挖掘的研究与实现
6. 互联网中事件检测与跟踪系统设计与实现（重要，也要看这篇文章的参考文献）

   http://gb.oversea.cnki.net/KCMS/detail/detail.aspx?filename=1017060480.nh&dbcode=CMFD&dbname=CMFD2017
7. 一种基于多特征融合的博客文章排序算法

CN102929977A.pdf
1MB

网络舆情热点发现与事...研究.pdf
3.8MB

微博热点话题检测与跟...吕伟.pdf
6MB

基于大数据的互联网热...宗飞.pdf
8.2MB

互联网中事件检测与跟...兆鹏.pdf
9.3MB

一种基于多特征融合的...卢刚.pdf
1.2MB

1.2 关键词

事件检测 知识图谱

在github里搜索关键字，排序大家的常用新闻网站https://github.com/search?p=2&q=新闻爬取&type=Repositories

~~1.3 常见的各个大厂的新闻网站~~

~~网易 新浪财经 每经网 中国证券网 腾讯 搜狐 凤凰网 新浪 东方财富 人民网 新浪体育 网易体育~~
~~可尝试爬取体育新闻~~
已废弃，转战游戏资讯

1.4 其他相关资料

1. 传智播客/黑马程序员mysql,mongodb,redis三大数据库精讲

   https://www.bilibili.com/video/av50382095
2. 2014人民日报词性标注语料库 https://www.aitechclub.com/?

2 团队的资料

下面是一些我们这个毕业设计团队的一些资料的所在地

2.1 github团队仓库

https://github.com/ghost-of-fantasy

2.2 看板

https://trello.com/invite/b/QRUSgZnm/f781bebfdfdbbd397ba962729aa7468d/毕业设计相关事项工作安排

看板里面任务要点开来看详细内容

https://my.oschina.net/huangyong/blog/196883

2.3 石墨文档

在毕业设计的那个文件夹

2.4 服务器信息

plrom.niracler.com:8080

3 设计理念以及原则

恩，设计理念以及原则要先在具体的模块设计前提出，大致的想法呢，先做一个可以帮助我们做标签筛选的工具，然后再将这个工具升级成为

3.1 怎样才是一个特别的软件

这里呢，讲道理就是真正的面向用户的地方。我最近才突然意识到，我一直以来想做的究竟是什么鬼东西。实在是很迷惘，我想做的大概是一个不太一样的新闻资讯平台，但我现在都快要做成一个内容社交平台加数据采集加监控的奇形怪状的应用，究竟是什么回事？？

我倒是想到了一些比较有趣的点子，不如就做一个“有原则的，不太一样的，任性的游戏新闻资讯应用”。所以呢，我就随便设定了一些所谓的原则：

3.2 原则

1. 一天只更新五条新闻
2. 以todolist的形式来看新闻，你要对你看过的并且喜欢的东西负责，所以请帮我们打个标签。
3. 我们不产生内容，也不需要你评论，请你到原网站评论。
4. 待定
5. 原则什么的不能太多，太多就会忘记了，只能有五条

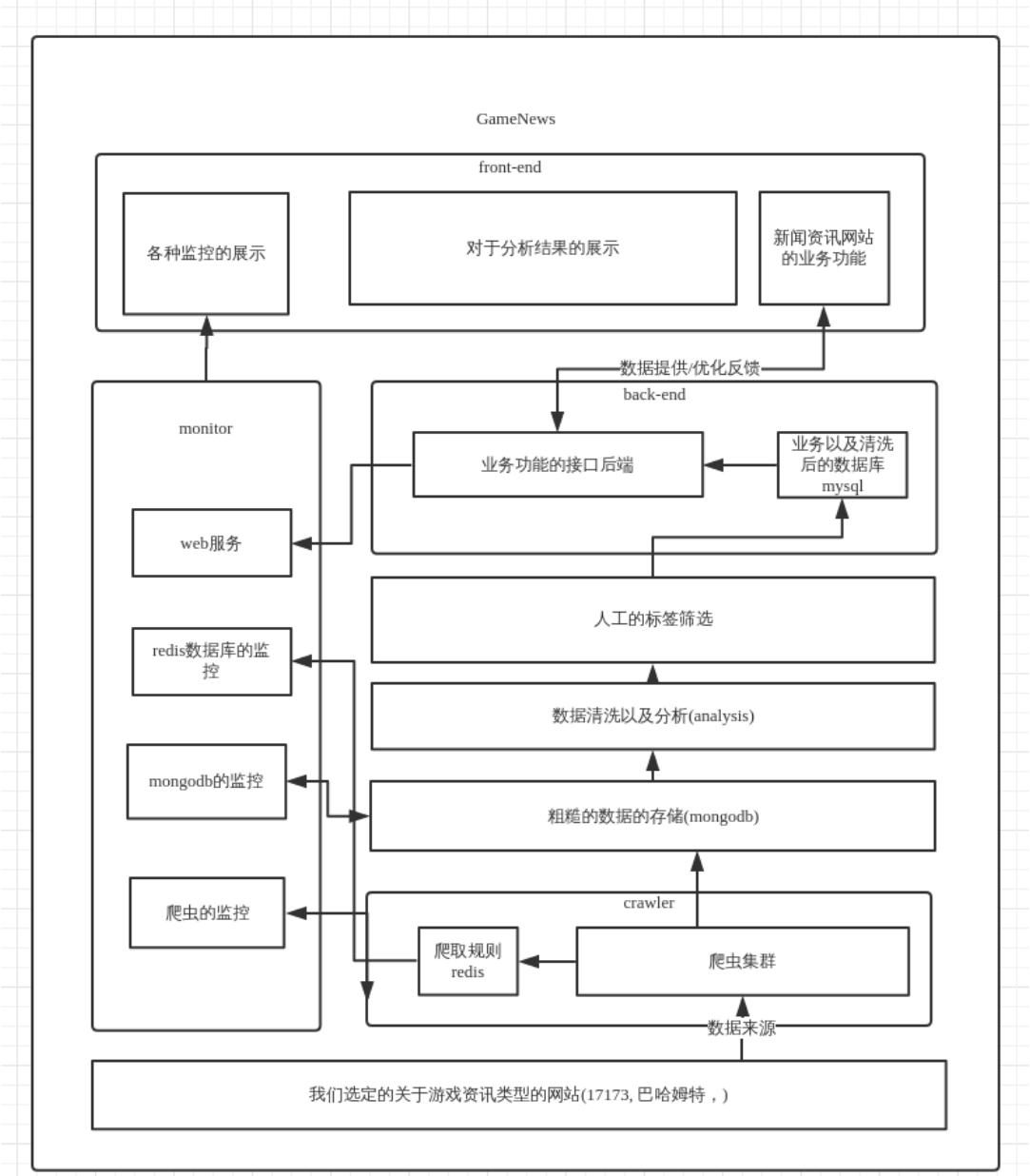
要说我为什么要有这五条原则，因为我们要立志做不太一样的应用。

3.3 总体的功能清单或需求说明书

- [ ] 展示数据(分析好的数据)
- [ ] 可以登录，然后可以关注特定的事件线索
- [ ] 评论点赞分享
- [ ] 生成卡片
- [ ] 提取新闻要素
- [ ] 有属于自己的后台分析
- [ ] 对原始数据的展示
- [ ] 量大的数据用mongo装，量少的用mysql
- [ ] 设计并完成一个数据从爬取到分析再到展示再到更新的一个流程
- [ ] 配置好SSL证书
- [ ] 准备好一个test服务器进行测试
- [ ] 将阿里云用上，作为向外展示的结果?

3.4 不专业的大致架构设计

https://www.processon.com/view/link/5d7201e2e4b0c5c942b8fe52



4 展示模块的后端部分（display-back-end）

这里主要是一些内容社交平台的功能及其接口，数据库是用postgresql，使用的web框架是python的django

4.1 后端开发的流程规约

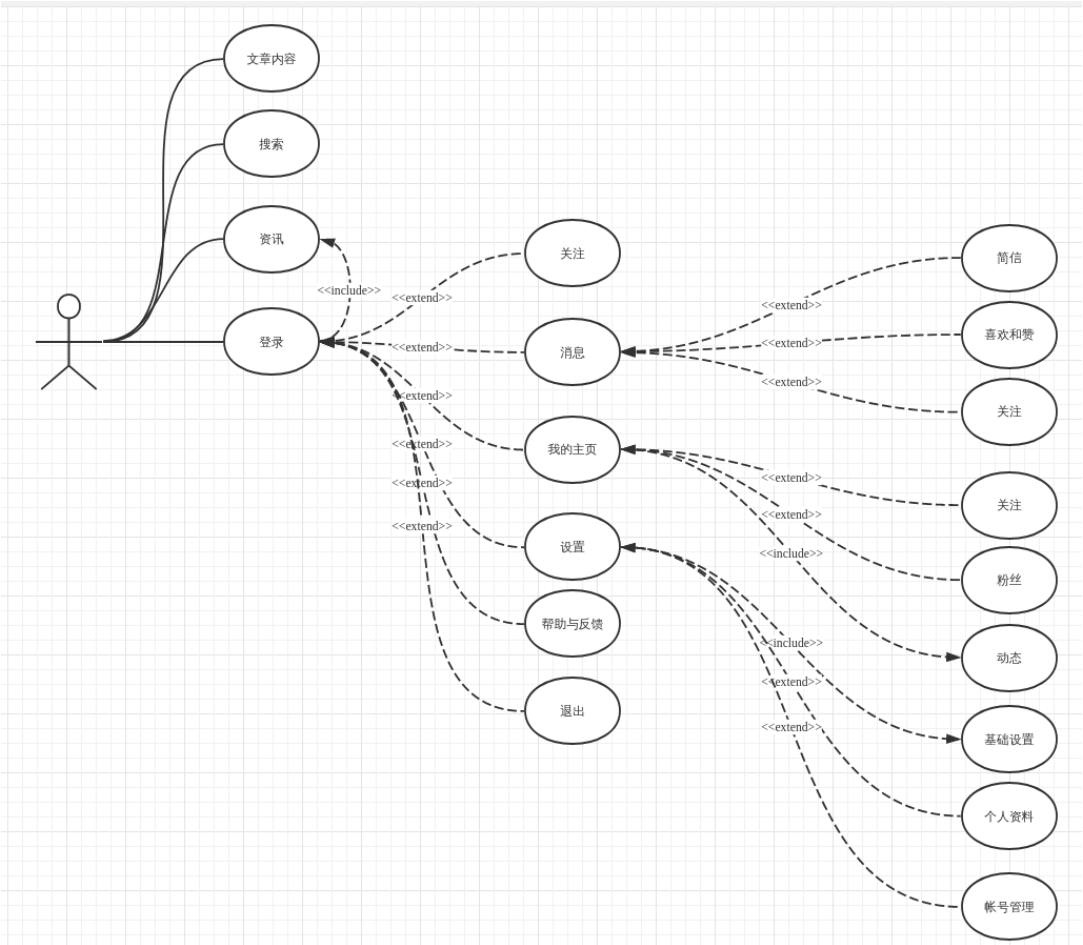
1. 数据库建模
2. 编写表单类
3. 编写视图函数和相关处理函数
4. 在页面中使用Jinja2替换虚拟数据

4.2 技术

django...到时候快结束时再加上

4.3 功能需求

- [x] ~~基于JWT的用户的登录注册功能的接口~~
- [ ] 对新闻的点赞或收藏功能的接口
- [ ] 对新闻的赞或者否的接口
- [ ] 一个类似于TODOLIST的接口
- [ ] 可以修改用户信息的接口
- [ ] 可以修改文章数据的接口
- [ ] 可以用于做文章的标签筛选的接口
- [ ] 记录用户活动流的接口
- [ ] 第三方登录接口
- [ ] redis进行数据缓存
- [ ] 记录文章访问次数
- [x] ~~添加文章的接口~~
- [ ] 对文章进行检索的接口
- [ ] 用户与用户之间互相关注的接口
- [ ] 文章要有一个字段判断文章是否通过标签筛选，通过的话，就无法再被更新
- [ ] 一个游戏实体以及游戏公司实体的管理接口
- [ ] 收藏数
- [ ] 点击数
- [ ] 限制api访问速率
- [ ] sentry错误日志收集
- [ ] 设计更为全面的用户信息以及用户信息编辑



5 展示模块的前端部分（display-front-end）

近段时间，大致是在国庆之前，我应该大致地完成好整个的原型设计。做好原型设计之后呢，就拿着那个原型去跟李宽老师交流

5.1 前端开发的流程规约

1. 根据功能规格书画页面草图（sketching）
2. 根据草图做交互式原型图（prototyping）
3. 根据原型图开发前端页面

5.2 原型设计

看来我还是要用一下下面这些工具来进行原型设计才行了
https://www.axure.com
https://www.mockplus.cn

5.3 功能需求

6 分析模块（analysis）

6.1 如何分词

因为像游戏名字，公司名字的那种信息，我们特别不全，于是就会有一种问题，我们很难能够正确的分词，而且有时候同个公司以及游戏的名字会有不同叫法，那我们该怎么将其聚合起来？

6.2 单独对标题进行聚类

1. 单独对标题进行聚类，我感觉效果也不会太差了（实际上，效果还是一般般）
2. 聚类，按时间线索排序，去重

6.3 事件脉络

通过热门的事件脉络自动追溯，可根据时间线对热点来龙去脉进行全面了解

6.4 文章分类

对文章按照内容类型进行自动分类，这个的话，由于我们没有标签，就挺麻烦的。

6.5 文章标签

对文章的标题和内容进行深度分析，输出能够反映文章关键信息的主题、话题、实体等多维度标签以及对应的置信度

6.6 热点发现

分析当前热点事件、话题热度，自动提供24小时内13个领域热点信息。

6.7 新闻摘要

基于深度语义分析模型，自动抽取新闻文本中的关键信息并生成指定长度的新闻摘要

6.8 功能需求

- [ ] word2vec 本来就有对词进行聚类的方法，所以可以用这个方法帮助找实体
- [ ] 上面这几个待爬网站都是各种实体，现在的比较大的问题就是分词，我们可以将爬取到的实体帮助分词
- [ ] 将书名号里面的内容提取出来，视作是实体

7 监控模块（monitor）

1. 这个监控模块主要是对各种数据库的监控以及后端的服务的心跳检测什么的。
2. 这个监控模块使用flask框架搭建的，主要面向于各种疑难杂症的web开发需求

7.1 功能需求

使用sentry对BUG进行监控。

8 爬虫模块（crawler）

8.1 关于游戏及发行商的实体的爬取

- [x] ~~https://www.3dmgame.com/games/zq_1/~~
- [x] ~~https://ol.3dmgame.com/ku/~~
- [x] ~~https://shouyou.3dmgame.com/zt/~~
- [x] ~~http://ku.gamersky.com/release/pc_199512/~~

要尽量爬多一点信息，连图片也爬取

8.2 候选的待爬取的游戏新闻网站

- [x] ~~17173~~ ~~https://www.17173.com/~~
  - [x] ~~国内新闻~~ ~~http://news.17173.com/dalu/~~
  - [x] ~~全球新闻~~ ~~http://news.17173.com/quanqiu/~~
  - [x] ~~韩国网游新闻~~ ~~http://news.17173.com/world/kor.shtml~~
  - [x] ~~产业新闻~~ ~~http://news.17173.com/chanye/list.shtml~~

- [x] ~~新闻专访 http://news.17173.com/zf~~
- [x] ~~巴哈姆特電玩資訊站 https://www.gamer.com.tw/~~
- [x] ~~GNN新闻 https://gnn.gamer.com.tw/~~
- [ ] 3dmgame https://www.3dmgame.com/news/
- [ ] 电玩巴士 https://www.tgbus.com/
- [ ] 游侠网 https://www.ali213.net/
- [ ] 游民星空 https://www.gamersky.com/news/
- [ ] 机核网 https://www.gcores.com/news
- [ ] 漫资讯 https://www.dongmanzx.com/
- [ ] acg批评 http://www.acgpiping.net/
- [ ] 半次元 https://bcy.net/
- [ ] 果壳网 https://www.guokr.com/scientific/
- [ ] 178网游 http://www.178.com/

8.3 功能需求

- [ ] 要有定时爬取的功能
- [ ] 对mongodb数据库进行增量更新的功能
- [ ] mongodb 的那个时间字段需要用上
- [ ] 或许可以用redais来处理增量更新
- [ ] 将redis爬取设置为可选项

9 测试与检验

1. 进入开发与测试的迭代
2. 调试和性能等专项测试