**总体来讲，在市场前景、技术能力和政策风险方面，有较大的不确定性，需要投入人力来调查。**

# 一、产品形态与市场前景

1、产品形态

公众号，集成到APP，做到车载广播里等

  主要特色为通过网络爬虫爬取国内外最新新闻资讯，摘取主要内容做成短资讯，翻译成中文，通过语音播报的形式呈现给用户。

2、产品特色

   国外最新资讯与动态，初步设想主要针对特定领域

   中文+语音播报

3、市场前景

   这方面需要进一步调查。

# 二、技术手段

需要以下几方面的技术

1、按类别抓取网页（分类、排序（热度））

2、分析（摘要、过滤（去重、删除垃圾新闻）、+人工筛查）

3、翻译（+人工筛查）-------当前翻译质量不达标，建议暂缓，主要针对国外的中文网站

4、语音合成

5、推送

重点在于推送，收集用户和变现：广告？移植到智能硬件中？

难点在于分析：如何提取高质量的摘要。目前有一些开源的算法和框架，测试下来效果并不理想，需要经过人工编辑才能使用。

              或者校企合作，学校在该方向技术比较成熟，如哈工大有专门的项目组。

不清楚业界当前使用的技术，网上没有相关资料。最好能在行业内调查下。

# 三、政策风险

对于时事类政策风险较大，且主观性比较强。目前难以通过技术手段来进行实时的屏蔽。

因此只考虑做一些中性类别的新闻摘要，如体育，科技，经济，人文等，但是不确定是否依旧存在风险。

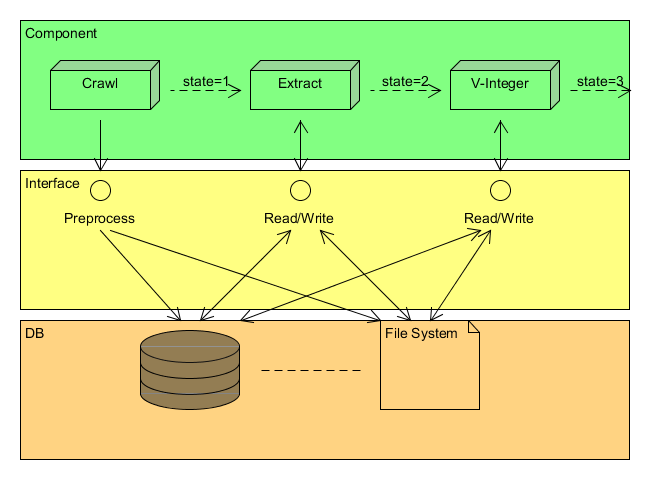
技术手段屏蔽？

# 四、总体框架说明

## 概要

系统分为三大组件：Crawl, Extract, V-Integer, 前后之间通过polling 的方式完成数据的交互。

如下图所示：



## 定时任务

1、每天起一次，扫描**NEWS\_INFO\_TBL，**删除产生时间超过MAX\_RESERVED\_NEWS\_TIME\_THRESHOLD天的新闻、摘要、生成的音频和相关数据结果

2、每隔CRAWL\_NEWS\_PERIOD分钟起一次，根据新闻产生时间和上次爬时间爬取最近一个爬取周期内产生的新闻。

## 其他说明：

1）爬取的网页按类别、爬取周期开始的时间分别保存到不同文件夹中

2）热度更新：对于已经存在的网页，根据URL下载计算热度相关的信息即可，不需要下载完整的网页内容。为了达到此目的，本地维护一张有效时间范围内所有新闻的表：

**表名称：NEWS\_INFO\_TBL**

**表中每项的信息：（网页内容的MD5值，URL，类别，阅读量，评论量，新闻爬取时间，热度更新时间，热度，新闻内容本地存放路径，新闻摘要本地存放路径，新记录标记，变更标记，废弃标记）**

|  |  |  |
| --- | --- | --- |
| 字段名 | 类型 | 说明 |
| id | int | PK |
| rawPath | Varchar(100) | 原始新闻的本地路径 |
| absPath | Varchar(100) | 新闻摘要文本的本地路径 |
| voicePath | Varchar(100) | 新闻摘要语音的本地路径 |
| deployTime | Timestamp | 新闻页上显示的时间 |
| crawlTime | Timestamp | 新闻爬取时间 |
| hot | int | 新闻的热度，是点击率和评论数的函数 |
| hotUpdateTime | Timestamp | 热度更新时间 |
| url | Varchar(200) | 新闻页的远程地址 |
| title | Varchar(100) | 新闻标题 |
| IsIdle | Tinyint | 废弃标记 |
| state | int | 新闻在系统中的状态（1：preprocessed，2: extracted，3: integered，4: deployed） |
| type | Varchar(50) | 新闻类型 |

每个爬取周期内增量爬取网页结束后，对于表中的所有URL下载其计算热度的相关信息并更新到表中。

**标记说明：**

网页内容的MD5值：用于新闻的唯一ID标识，用于辅助生成文件名。

URL：新闻对应网页的网址

类别：新闻类别

阅读量：

评论量：

新闻产生时间：

热度更新时间：

内容本地存放路径：

摘要本地存放路径：

新记录标记：表示是否为本爬取周期新增加的记录，用于后续的摘要提取。

变更标记：表示本爬取周期内该记录是否有修改，只针对老记录有效。在每个爬取周期开始前，所有记录的该标记清为False。

3）增量爬取网页：根据上次的时间进行过滤爬取，只爬取新增的新闻,并将相关信息填到**NEWS\_INFO\_TBL**中.

## 系统全局信息说明

**爬取网站列表**

CRAWL\_WEB\_SITES\_INFO = [‘sohu’, ‘sina’, ‘yaho’…..]

**新闻类别列表**

NEWS\_TYPE\_INFO = [‘sports’, ‘[technology](file:///D:\Program%20Files%20(x86)\Dict\7.2.0.0703\resultui\dict\?keyword=technology)’, ‘economics’…..]

**最大爬取新闻总量**

MAX\_CRWAL\_NEWS\_NUM = 100k

**最大保留新闻总量**

MAX\_RESERVE\_NEWS\_NUM = 1K

**单类别最大保留新闻总量**

SINGLE\_CLASS\_RESERVE\_NEW\_MAX\_NUM = MAX\_RESERVE\_NEWS\_NUM / len(NEWS\_TYPE\_INFO)

**新闻最大保留时间**

MAX\_RESERVED\_NEWS\_TIME\_THRESHOLD = 7(天)

**新闻爬取周期**

CRAWL\_NEWS\_PERIOD = 30(分钟)

# 五、接口说明

处理工作量不大，可单机串行执行。分为如下几个模块，通过API调用的方式串联完成系统任务。

使用增量处理的方式，定时从特定网站抓取指定类别的新闻

## 按类别抓取网页（分类、热度）

定时在指定网站抓取指定类别的新闻，同一类别内的新闻按照热度排序，封装在一个List中，每个（类别标签：新闻List）组成字典中的一项

函数原型：crawlSomeWebPages(urlList, labelList, date)

参数说明：

urlList为一组URL组成的列表，表示从哪些网站抓取新闻

labelList为类别组成的列表

date:最旧新闻的时间，再此之前的不爬

注意：此处的类别为一组通用的标签，如经济、体育、科技等。不同网站的类别标签可能存在差异，爬取网页需要映射到指定网站的类别

返回值说明：

返回一个字典newsDict，每项为（类别标签：新闻List）

新闻List中的一项 = （URL，新闻标题，新闻内容，热度，扩展信息）

热度可以理解为阅读量，表示新闻的重要程度，用于后续的推送顺序。可参考如下公式技术：

热度 = （a\*阅读量 + b\*评论数）/ （当前时间 – 新闻出现时间）

（a b为两者的权重，一般评论数的权重高点）

也可以考虑一些其他的策略，尽量将用户感兴趣的新闻放到前边。

扩展字段用于后续的扩展，可暂时放一个空List

## 分析

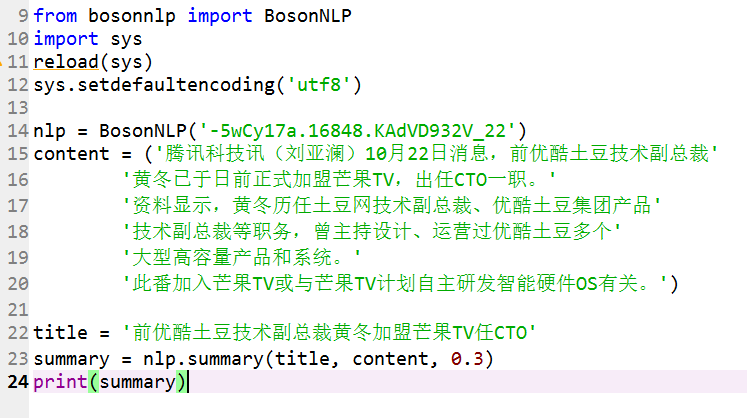
包括提取摘要、过滤（去重、删除垃圾新闻）、人工筛查等步骤

1. 摘要：

该步仅为增量处理。针对上一步增量网页抓取的结果（“新记录标记”为True的记录），提取摘要完成后将”新记录标记”修改为False.按照类别分别处理

并填充表**NEWS\_INFO\_TBL** 中的字段“**新闻摘要本地存放路径**”。

使用bosonNlp的API进行摘要提取，演示代码如下：



函数原型：getSummary(newsDict)

参数说明：

newsDict见 crawlSomeWebPages的返回值说明

返回值说明：

字典summaryDict，每项为（类别标签：新闻摘要List）

新闻List = （新闻标题，新闻摘要，热度，扩展信息）

1. 过滤

该步为全量的处理，对于NEWS\_INFO\_TBL中所有记录进行过滤，提取新闻中热度高的（其他记录需要保留，原因是其热度是动态变更的）。

1. 推送新闻过滤

为了减少语音合成音频的数量，分析阶段按热度和推送的最大数量过滤出本爬取周期内的推送新闻列表，用于语音合成和推送。相关信息保存到

**PUSH\_NEWS\_TBL** = [[网页内容的MD5值，URL， 热度，新闻标题，新闻内容存放路径，新闻摘要, mannulChangeFlag], [网页内容的MD5值，URL， 热度， 新闻标题，新闻内容存放路径，新闻摘要, mannulChangeFlag] .....]，可以保存到CSV中，便于手动修改。

1. 人工筛查

对3）中生成的CSV文件，进行手动修改。并设置,mannulChangeFlag标记为True。

## 翻译（+人工筛查）-------当前翻译质量不达标，建议暂缓，主要针对国外的中文网站

## 语音合成

将标题和摘要合成音频，并保存到本地文件中

扫描**PUSH\_NEWS\_TBL**，检查有无对应的音频文件，如果没有则生成；如果有但是mannulChangeFlag标记为True则重新生成并覆盖原来的，并将mannulChangeFlag修改为False。

音频文件需要按类别分别存放。

**注意:**本阶段需要做到可手动启动,便于人工修改完之后重新生成新闻

## 推送

按类别将**PUSH\_NEWS\_TBL**中热度最高的SINGLE\_CLASS\_RESERVE\_NEW\_MAX\_NUM条新闻推送给用户。

## 用户信息反馈

点击量

定制化

# 附录

## 根据文件内容提取MD5方法：

import hashlib

def getUniqueIDByNews(content):

md5 = hashlib.md5(content).hexdigest()

return md5

def getUniqueIDByFileName(filePath):

md5file = open(filePath, 'rb')

md5 = getUniqueIDByNews(md5file)

md5file.close()

return md5

## 新闻、摘要和语音文件保存的文件名

新闻内容文件名：classLabel\_title\_md5\_content.txt

新闻摘要文件名：classLabel\_title\_md5\_summary.txt

新闻语音摘要文件名：classLabel\_title\_md5\_summary.mp3