**关键词保障系统项目**

**技术需求**

**关键词保障系统项目组**

**2019年3 月**

编号：

版本：**Ver 1.4.3**

密级：

# 项目概述

## 项目背景

在传统媒体积极变革的大背景下，新闻生产过程对于信息的正确性、权威性提出了更高要求。因此，保障关键内容的准确性，成为了在新闻生产环节过程中至关重要的一环。

关键词保障系统项目主要是为了满足用户需求，在现有新闻业务管理系统进行文稿内容编辑和审核时，提供一套关键词保障系统作为关键字词检测的辅助工具。即用户在文稿内容编辑完成或稿件审核时，点击按钮自动对当前稿件内容进行关键词检测，并将检测的结果呈现给用户进行参考。

## 项目定位

在当前信息技术高速发展、互联网带来信息爆炸的复杂态势下，在对新闻内容生产效率提出更高要求的同时，信息的正确性和权威性也必须得到充分保证。

关键词保障系统项目拟从技术角度出发，提出对节目文字内容进行智能化检测的解决方案，应用于新闻业务流程的文字内容生产环节：

1. 对包括重要领导人姓名职务、重要会议及事件、国家部委名称、禁用词慎用词、错别字等关键内容在内的各项重要数据进行检测，辅助节目审查流程；
2. 系统在实现方式上确保对现有新闻业务系统不产生影响；

## 总体目标

本项目的总体目标是在现有新闻业务管理系统进行文稿内容编辑和审核时，提供一套关键词保障系统作为关键字词检测的辅助工具，实现第一财经网络制播系统内节目文字内容的智能化校验的需求。

本项目将从第一财经频道目前正在使用的非编网络制播系统（索贝）着手，开发关键词鉴别软件，对电视播出相关的文稿内容、字幕内容进行自动审核校对，辅助节目三审流程，达到助力节目内容播出准确、权威、安全的目的。

在功能方面，关键词保障系统的主要涉及以下三个方面：

1. 在检测功能方面：
   * 检测不能出错的内容，如重要领导人姓名、职位、重大会议名称和政治专有名词等。
   * 检测需要谨慎对待的词句，如落马官员姓名，某些特定称谓（如“大陆”、“内地等”），禁用慎用词汇。
   * 检测错别字。
2. 在日志功能方面：
   * 系统对用户每一次的检测行为，记录相关数据，并保留日志。
   * 对日志数据进行汇总呈现，形成系统能力及用户行为分析报表。
3. 在后台管理功能方面：
   * 系统需有弹性的词库管理功能，为SMG新建一套自有词库打下基础。
   * 对涉及用户权限的模块，对用户行为进行管理。

# 项目需求分析

## 第一财经子项目



### 用户需求

#### 用户业务特点

第一财经关键词保障系统针对第一财经（以下也可简称“一财”）用户，一财稿件多以财经类资讯为主，但在重要事件发生时，会产生相关涉政内容的报道。这些报道对重要领导人姓名、职位，重要会议名称、国家部委名称等都有可能涉及，相关的文字可包含在B/S架构的稿件应用编辑区以及调用第三方orad插件制作的字幕板中。

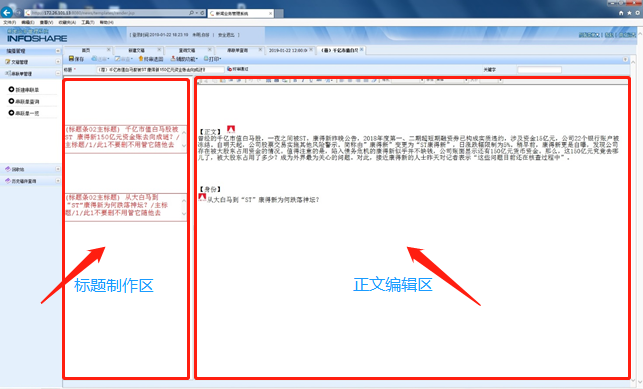
目前第一财经各栏目稿件审核流程为三审，分一审、二审和终审，该审核流程全部人工完成，审核人员资格高低与其岗位重要性相对应，成正比关系。并且在管理上，各栏目会定期对节目质量进行反思和跟踪。

#### 用户对关键词保障系统的需求

融媒体时代节奏快、产出高，对内容生产的效率提出更高要求，但信息的正确性和权威性丝毫不能让步，必须得到充分保证。更进一步，若新闻生产内容涉及重大方针、重要会议、重要人物等政治敏感内容，则影响极为巨大。第一财经作为新闻资讯类的电视新闻媒体，确为各种信息宣传、传播的重要阵地，要进一步加强关键内容准确性的保障力度，以应对在融媒体环境下的新闻生产快速、高效、严格的特征。业务流程仍以严格的人工审核为主，关键保障系统可作为辅助工具，强化审查环节力度。

该系统对第一财经内容生产的关键词保障能力如下：

1. 需要覆盖一财稿件写作系统（即“新闻业务管理系统”）中的稿件正文和标题制作两个区域的文字；并且在三审过程中均可使用。



1. 需要能够识别不能出错的关键词语或表达方式、不能出现的或需慎重出现的关键词语或表达方式，并给出修改建议；
2. 能够识别错别字，并给出修改建议。
3. 界面合理友好，操作便捷，在检测方式和结果显示方面给用户一定的选择自由度。

### 需求分析

#### 功能需求

功能点主要包括以下几点：

1. 文稿文字内容检测。包括文字内容抓取抓取、文稿文字内容关键词检测，关键词检测结果通过颜色标记的方式反馈出来，并给出修改建议。其中关键词包括重要领导人姓名职位、重要会议、国家部委名称、专有名词禁用语慎用语、错别字五大项。

2. 日志功能。该检测系统需提供可追溯的日志功能，每日在每台客户端本地及服务端为每台客户端生成一个日志文件，记录每次点击校验的时间、当前登录的用户名、稿件栏目、稿件标题、设备IP、每次校验后的检测结果及相应时间，以及检出项目。检出项目需连带上下文各10个字记录在日志中，并且所有检出项目在日志中按其在原文中的先后顺序排列。

3. 报表功能。该系统需可以按照时间或用户在服务端生成整个系统的审计分析日志，不仅可总结每台单机的使用情况，也可对所有用户的使用习惯进行全局性的分析。

4. 词库更新功能。该系统需要能够提供在线自动及离线手动更新词库的功能，并记录和体现更新变化的内容以便用户核查。

5. 词库自定义功能。该系统的关键词库主要涉及标准用语和禁用语，需对管理员开放词库入口，已供输入用户自定义的关键词。可提供管理词库的工具，使用户能够按照自身的业务特点设置“白名单”（即合规合法的字、词、组合表达等）和“黑名单”（即自定义的不合规的字、词、组合表达等），并且设置后，在检测系统中立即生效。

6.深度学习功能。系统上线初期，使用的是厂商自有词库及其根据各自的原始海量语料编写的模型，上线后，系统应能够使用SMG定期提供的、脱敏后的、本地化的语料进行模型的离线训练，以通过人工智能手段达到模型本地化的效果，进一步提升系统检测准确率。

7.用户管理功能。该系统用户分为频道用户、总编室用户和管理员用户，频道用户的账号管理跟随原稿件或制作业务系统；对于总编室用户，因涉及词库管理工作，系统需提供工具对该用户账户、密码进行管理；对于管理员用户，因涉及词库优先级选择、报表生成等操作，也将提供工具对该用户账户、密码进行管理。

以上功能点将嵌入下述业务流程中：

* + - 用户保存文稿内容后，点击文字抓取软件中的检测按钮。
    - 文字抓取软件将文字内容通过接口传给关键词检测系统。
    - 关键词检测系统对接收到的文字内容进行分析检测，并将结果反馈给文字抓取软件。
    - 检测出的可能出错的字词以不同颜色标记的方式在文字抓取工具窗口中呈现出来，用户根据检测结果的提示决定是否需要手动进行修正。

#### 性能需求

1. 系统后台能够承受的并发压力。

以第一财经高峰时段50台客户端计（加上2.2融媒体中心子项目中制作高峰时段80台），同时共计130台客户端发起检测请求，每条请求稿件平均字数约500字。从客户端发起检测请求来看，从每次启动检测功能，到检测结果反馈并显示出来，500字以内的稿件耗时需控制在2秒左右，1000字以内稿件耗时需在3秒以内，5000字以内稿件耗时在5秒左右。

1. 关键词保障系统亦不可影响现有“新闻业务管理系统”的各项性能。

现有“新闻业务管理系统”新建文稿<2秒，打开文稿约<2秒，保存文稿<2秒；新建串联单<3秒，打开串联单<3秒，编辑串联单（添加稿件操作）<3秒。

1. 对于检测精度的要求。

对于应检测出来，却没有成功检测出来的错误率应控制在40%以下。对于原本正确，却被检测出来认为错误的误检率，系统根据检测的严格程度不同可进行控制。

#### 系统规模

该系统需以C/S架构实现，需覆盖第一财经非编制播系统目前客户端数量共计135台。

系统按照目前新闻业务管理系统中文稿工作站的总数来设计系统规模，高峰时段并发量可达到50台左右，系统需保证所有用户同时使用该系统时仍然能够快速有效完成关键词的检测。后台服务器采用主备的工作模式，防止因1台服务器故障而导致整个系统不可用的情况发生。

服务器操作系统采用Windows Server和Linux操作系统，windows系统安装操作系统补丁并部署防病毒软件，且所有操作系统仅开启应用服务所需端口，关闭远程访问服务，防止系统因病毒、漏洞等安全因素而影响系统的正常运行。

#### 扩展要求

1. 规模扩展

系统能够根据客户端数量的增加，通过灵活增加服务器数量或提升配置等手段增强系统的负载能力。

2. 业务扩展

系统网络架构能够满足第一财经APP、XNEWS等其他业务模块的接入。

#### 其他

在该系统软件开发过程中：尽量不改变原系统代码。不可植入不相关的信息或功能，以满足台、集团信息安全规范要求。

软件开发不可影响现有终端的软件功能，如精编、审片、配音等软件。

# 项目技术方案

## 3.1第一财经子项目



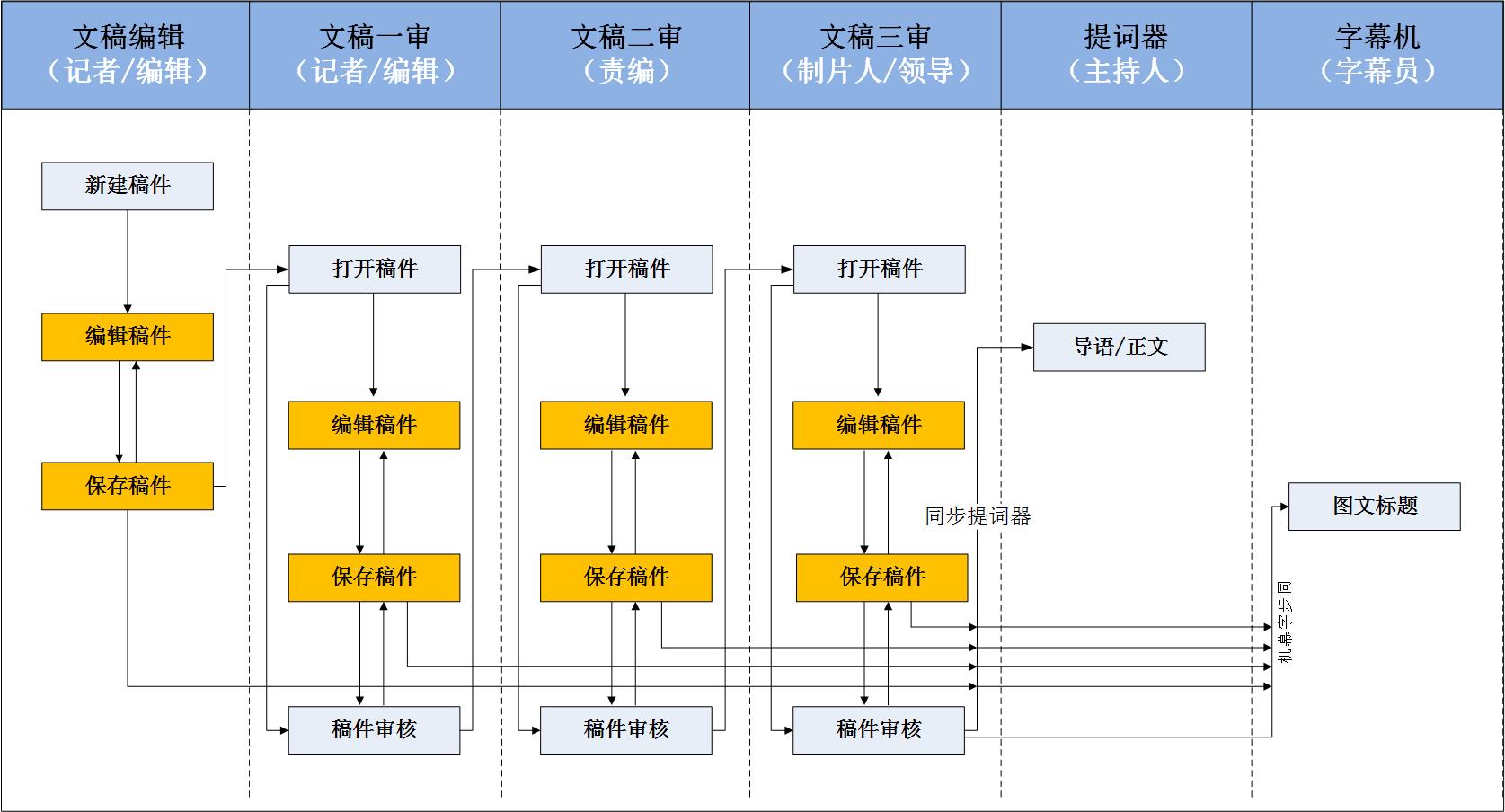
### 3.1.1 流程设计

#### 3.1.1.1 关键词保障流程

关键词保障系统的工作流程始于用户发起检测请求，文字抓取软件将获取web页面内容，经过数据清洗后，将有效文字数据分别传递给两个不同的关键词检测系统进行处理，两个检测系统再将分别处理的结果返回至文字抓取软件，由该软件负责展现结果。在客户端展现结果的同时，系统后台服务端生成基于当次检测请求和结果的相关日志或报告。

#### 3.1.1.2 嵌入关键词保障系统后的用户稿件生产业务

1. 关键词保障系统不改变用户现有的业务流程。只在文字内容生产的写稿（含一审）、二审、三审各环节的工作过程中，提供检测功能，用户可随时，但主要在保存稿件后使用本关键词保障系统进行稿件文字内容的检测。下图中黄色标记模块表示可进行检测的业务环节。



1. 该软件界面最小程度改变目前“新闻业务管理系统”的呈现方式。检测结果可由用户配置默认隐藏与否，若默认隐藏，只向用户给出提示，由用户本人决定是否打开检测结果进行参考。

### 3.1.2 文字抓取软件子系统（软件子系统）

文字抓取软件，在现有新闻业务管理系统（基于IE浏览器的B/S架构）的基础上，开发一个兼容IE的外壳，在文稿编辑、文稿审核的工作站上安装该软件,辅助文稿编辑软件完成关键词的检测。

文字抓取软件主要包括基于第一财经新闻业务管理系统的IE浏览器套壳软件和基于融媒体中心iNEWS文稿系统的接口软件，主要负责获取客户端文稿编辑区域指定的文字内容，并通过接口发送给后台检测系统。其主要功能包括如下几点：

#### 用户界面功能

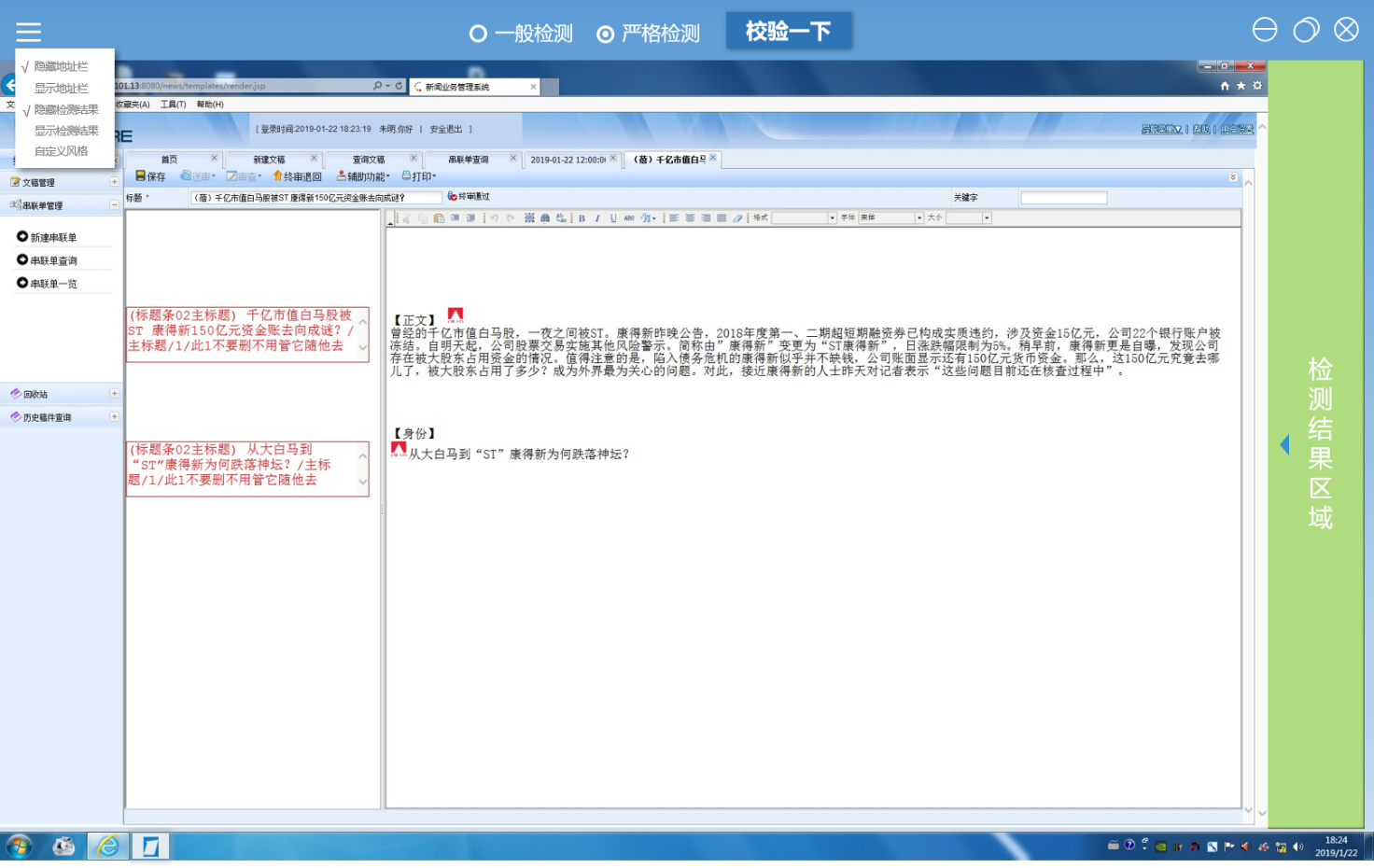
1. 设置“校验”功能按钮，点击后触发文字校验功能，可将稿件文字编辑区的文字内容发送给关键词检测系统进行处理。
2. 校验的宽松程度可选，分为“一般检测”和“严格检测”，一般检测将重点关注国家领导人姓名职位、国家重要部委名称、重要会议和重大活动、禁用语等类别；严格检测除关注一般检测功能中的类型外，还将关注错别字、语法、标点等类别。系统默认使用“一般检测”。
3. 校验结果默认隐藏，但用相应颜色指示检测结果，起到提示作用。
4. 通过双击有关区域，点击按钮或快捷键打开检测结果。
5. 第一财经的“新闻业务管理系统”有主备两个地址分别提供服务，文字抓取软件需考虑这两个地址各自的入口。
6. 用户可自定义界面样式，主要是颜色，也可自定义界面布局，主要包括地址栏的显示与否，及检测结果默认隐藏还是显示。检测结果区隐藏时，隐藏区用相应颜色（红——严重错误、黄——一般错误、绿——无错）指示检测结果，起到提示作用。错误按照严重程度进行分级，一般认为重要领导人姓名职位、重要会议、禁用语为较高级别；国家部委名称、专有名词、慎用语、错别字为普通级别。

文字抓取软件主要UI呈现界面分别如图所示（以下均为参考界面）。

下图（参考）展示了文字抓取软件界面的主备文稿双入口。



下图（参考）展示了在选择“隐藏地址栏”和默认“隐藏检测结果”后的检测界面，在检测结果区以颜色区分错误严重程度的效果，绿色代表无错，黄色代表非严重错误或存在修改建议，红色代表出现严重错误。右侧结果区处于收起状态，显示为绿色，用户可自行决定是否打开检测结果区。



下图（参考）展示了在选择“显示地址栏”和默认“显示检测结果”，并修改自定义界面颜色风格后，检测结果区打开的界面效果。

（1）检测结果区的打开和收起，可通过双击该区域、左右方向键、点击区域左侧小三角等方式进行操作。

（2）结果区打开后，原“新闻业务管理系统”应自动缩小至合适大小，嵌入左侧窗口内。

（3）鼠标悬停在检出项时，应浮现该错误项的修改建议。鼠标在检测结果区点击某检出项时，左侧文稿编辑界面中的光标会移动到编辑区相应文字所处位置，以提示用户错误项在原文中的具体位置。



#### 3.1.2.2日志功能

该检测系统需提供可追溯的日志功能，每日在每台客户端本地及服务端为每台客户端生成一个日志文件，记录每次点击校验的时间、当前登录的用户名、稿件所属栏目、稿件标题、设备IP、每次校验后的检测结果及相应时间，检出项目，以及该次检出由词库A实现还是由词库B实现。检出项目需连带上下文各10个字记录在日志中，并且所有检出项目在日志中按其在原文中的先后顺序排列。

### 3.1.3关键词检测子系统设计（软件子系统）

关键词检测系统包含词库B和词库A两套软件服务，主要提供对文字抓取软件通过接口发送的文字内容进行检测和结果反馈，系统词库可进行更新。

#### 3.1.3.1文字检测系统

其中关键词包括重要领导人姓名职位、重要会议、国家部委名称、专有名词禁用语慎用语、错别字五大项。

* 1. 词库B软件只提供文字输入输出接口，及词库自定义入口。
  2. 词库A要求积累至少4000万篇高质量的文字语料，拆分成1000亿词语语料，采用基于统计算法和深度学习框架相融合的方式，实现高质量的错别字检测功能。

词库A人工智能错别字检测系统要求基于Docker模式进行软件分发，可以做到和云平台、微服务模式无缝集成，可有效地降低企业软件TOC总成本。

* 1. 双词库检测冲突规则

文字经过双词库处理后，获取的检测结果将被进行比对，对字词正确与否的判断将全部包含在以下矩阵内：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **检测系统判别结果** | | **输出结果** | **结果选取** |
| **词库B** | **词库A** |
| 1 | 正确 | 正确 | 正确 | 可配置 |
| 2 | 正确 | 错误 | 错误 | 词库A |
| 3 | 错误 | 正确 | 错误 | 词库B |
| 4 | 错误 | 错误 | 错误 | 可配置 |

对校验结果的判定采取“同判可配置，异判取错方”的原则，即：

* 若词库A和词库B对一个字词的判断均为正确，或均为错误，系统可由配置决定输出哪方的检测结果。
* 若词库A认为该字词无错，词库B认为有错，输出认为有错一方的结果，即词库B。
* 若词库B认为该字词无错，词库A认为有错，输出认为有错一方的结果，即词库A。
  1. 系统检测结果和自定义词库冲突规则

用户自定义的词库优先级最高，包括白名单和黑名单。若上述3）中的系统判定结果和黑、白名单冲突，以黑白名单为准。

#### 词库更新功能

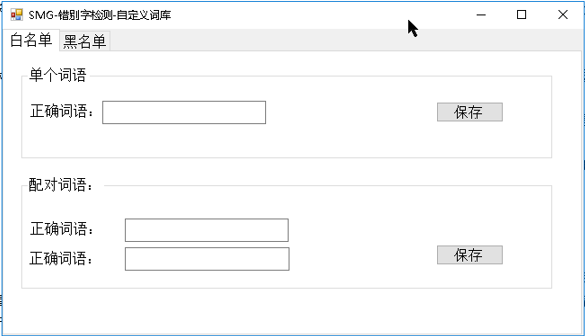
该系统需要能够提供在线自动及离线手动更新词库的功能，并记录和体现更新变化的内容以便用户核查。

### 后台管理子系统设计（软件子系统）

#### 词库自定义功能

1. 提供管理工具。该系统的关键词库主要涉及标准用语和禁用语，需对管理员（总编室用户）开放词库入口，已供输入用户自定义的关键词。

* 通过提供管理词库的工具，使用户能够按照自身的业务特点设置“白名单”（即合规合法的字、词、组合表达等）和“黑名单”（即自定义的不合规的字、词、组合表达等），并且设置后，在检测系统中快速生效。工具界面如下图，图中“单个词语”表示独立的字词，配对词语指的是组合表达，如“国家主席”+“习近平”等。



* 词库管理工具中自定义的内容，通过同时调用两个接口的方式，实现向词库A及词库B校对同时写入。

1. 用户更新词库的行为也可生成日志，日志记录项目包含每次更新的操作账户、更新时间、更新内容、设备IP。
2. 用户自定义后，有核对自定义词汇是否已在检测系统生效的工具，且该工具可生成检测日志。日志记录每次点击校验的时间、当前登录的用户名、设备IP、每次校验后的检测结果及相应时间，以及检出项目。检出项目需连带上下文各10个字记录在日志中，并且所有检出项目在日志中按其在原文中的先后顺序排列。

#### 报表功能

该系统需可以按照时间或用户在服务端生成整个系统的分析报表，不仅可总结每台单机的使用情况，也可对所有用户的使用习惯进行全局性的分析。

1)  用于系统准确率统计分析：用户在使用该系统进行文字检测时，记录每一次检测结果，自动存储用户文稿内容及系统检测结果，通过数据对比分析对多次系统检测出的“错误”进行分析和自定义，提升系统检测率。

2) 用于对系统使用情况进行统计：用户在使用该系统进行文字检测时，记录用户工号/姓名、检测的文稿标题、检测时间等，形成用户使用情况统计表，提高用户对检测系统的重要性的重视程度。

3)   用于稿件准确率的分析统计：检测系统可以记录用户每一次检测的文稿内容及检测结果，结合文稿系统历史操作轨迹，分析用户对错误字词的更正情况并进行稿件准确率统计。

#### 词库优先级自定义功能

针对双词库检测结果冲突的规则设计，若词库A和词库B对一个字词的判断均为正确，或均为错误，可在后台管理系统中进行配置，决定输出哪方的检测结果。

#### 用户管理系统

用户管理系统提供账户配置和简单的登录认证功能，要求密码符合台集团信息安全要求的复杂度，以对用户行为进行约束。

该系统用户分为频道用户、总编室用户和管理员用户，频道用户的账号管理跟随原稿件或制作业务系统，在此系统中不涉及。对于总编室用户，因涉及词库管理工作，系统需提供工具对该用户账户、密码进行管理；对于管理员用户，因涉及词库优先级选择、报表生成等操作，也将提供工具对该用户账户、密码进行管理。

用户管理系统采用基于MySQL数据库，通过web服务的方式，为管理员用户提供用户信息的统一管理。