# Intelligent text big data analysis display platform

Yingzi Yuan Hongli Zeng Spring Meng

Dec 2022 – Feb 2023 INSIS, Beijing

Collaboration with Gilight Education Technology Co., Ltd



# Project Background

With the continuous development of campus informatization, various campus information systems (such as portal websites, team sites, personal homepages, and campus forums) have become increasingly complex.

These systems generate **massive** amounts of data, but there is a lack of effective integration, management, and analysis.



### **Current Problems:**

- Difficult information retrieval due to scattered data sources and lack of a unified search entry.
- Lack of effective supervision, making it difficult to monitor content quality and detect sensitive or inappropriate content.
- Potential security risks, including personal information leaks and harmful content.
- Limited support for decision-making, as there are no effective tools to extract insights from text data.

# System Architecture



Portal websites, team sites, personal homepages, campus forums

. . .









Exact search
Fuzzy search
Multi-word search
Statistical analysis
Hot topic discovery
Similarity calculation
Update frequency monitoring



User management
Keyword search
Statistics
Visual reports

**Data Sources** 

**Data Storage & Processing** 

**Key Functions** 

**User Interface** 

### **Security Measures**

### **Operations Assurance**

Data masking, encryption, and authorization to ensure data privacy and security.

Monitoring, process management, operational dashboards, and reporting tools to ensure smooth system operation.

### Problems to Solve

# Overall Design Concept

## Key Text Analysis Features

# Campus Text Data Collection & Management

# Campus Text Data Retrieval

**Campus Text Data Analysis** 

### **Data Collection:**

Real-time crawling of all campus information release pages

### **Data Management:**

Text data is pre-processed using word segmentation, stored using the Hadoop ecosystem and MongoDB, and indexed using Elasticsearch.

### **User Interface:**

Based on Java and VUE framework, providing users with functions such as configuration, login, search, data visualization, and export.

### **Hot Topic Discovery:**

Identify trending topics over time and generate ranking lists to help staff follow the latest developments.

### **Monitoring and Alerts:**

Detect sensitive words, inappropriate remarks, and visualize results for administrators to take immediate action.

### **Content Quality Monitoring:**

Analyze publication frequency and readership to assess content quality. Missing content is flagged for review.

### **Relationship Analysis:**

Explore hidden correlations between different groups and content types to improve content recommendation accuracy.

# Home page overview

Through comprehensive event
analysis across the network, heat
index tracking, and comment
analysis, the platform enables
real-time monitoring of published
content across all sites and pages,
while also tracking and counting the
number of detected anomalies to
meet the monitoring requirements of
customers.



# Search & Query

Through keyword extraction, related word analysis, and precise query capabilities, the platform enables multi-functional monitoring of website text, images, and videos, achieving accurate identification of the target information.

- Supports exact search and fuzzy search
- Supports single keyword search and multi-keyword combined search



# **Anomaly Alerts**

Based on pre-configured alert conditions set by the user, the platform can intelligently identify abnormal information.
When content matching the alert criteria is detected, the system will automatically trigger an alert, allowing users to take immediate action.

### Key monitored anomalies:

- Sensitive words
- Broken links
- Empty content
- Zombie websites (inactive sites)



# Rectification Tracking

Based on pre-configured alert conditions set by the user, the platform can intelligently identify abnormal information.

When content matching the alert criteria is detected, the system will automatically trigger an alert, allowing users to take immediate action.

Track rectification progress for each issue, ensuring timely resolution and content quality improvement.

### Status categories:

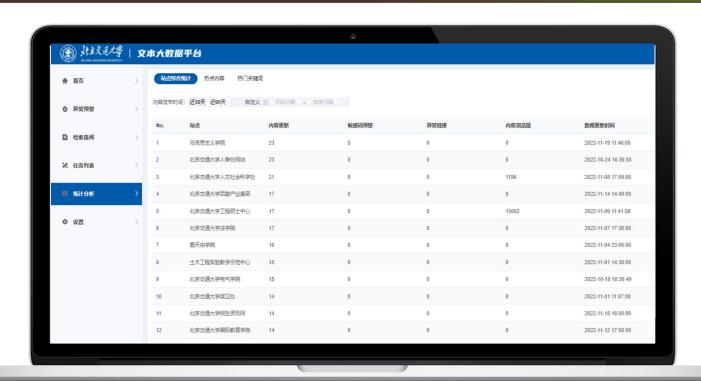
- Pending rectification
- Completed rectification



# Statistical Analysis

### **Key Analysis Functions:**

- Comprehensive site statistics
- Trending content identification
- Popular keyword tracking



# Summary

This platform provides comprehensive campus text data management, including:

- Real-time data collection.
- High-performance search.
- Advanced text analysis and monitoring.
- Visual reporting and decision support.

### It helps the university:

- Monitor public opinion and campus sentiment.
- Support decision-making with data insights.
- Enhance data security and content management.



