# SEO Score Classifier Documentation

## Overview

The SEO Score Classifier is a sophisticated system designed to evaluate websites based on their search engine optimization (SEO) implementation. It analyzes various aspects of web pages and provides a comprehensive score indicating the level of SEO optimization.

## Scoring Categories and Weights

The classifier divides SEO factors into four main categories, each contributing differently to the final score:

1. Technical SEO (35% weight)
   * Focuses on technical implementation aspects
   * Highest weight due to fundamental importance in search engine crawling and indexing
2. Content Quality (30% weight)
   * Evaluates content structure and optimization
   * Second-highest weight reflecting content’s crucial role in SEO
3. User Experience (20% weight)
   * Measures factors affecting user interaction
   * Moderate weight acknowledging growing importance of UX in SEO
4. Meta Elements (15% weight)
   * Assesses basic SEO meta tags and elements
   * Lower weight as these are fundamental but not sufficient alone

## Indicators and Data Collection

### 1. Technical SEO Indicators

#### HTTPS Implementation (20 points)

* **Collection Method**: URL scheme analysis
* **Scoring**: Binary (20 points if present, -20 if absent)
* **Rationale**: Security is crucial for modern SEO and user trust

#### Robots.txt (20 points)

* **Collection Method**: Server root file check
* **Scoring**: Binary (20 points if present, -20 if absent)
* **Rationale**: Essential for search engine crawl control

#### Sitemap (20 points)

* **Collection Method**: XML sitemap detection
* **Scoring**: Binary (20 points if present)
* **Rationale**: Aids search engine content discovery

#### Canonical Tags (20 points)

* **Collection Method**: HTML head tag analysis
* **Scoring**: Binary (20 points if present)
* **Rationale**: Prevents duplicate content issues

#### Structured Data (20 points)

* **Collection Method**: Detection of JSON-LD or Schema.org markup
* **Scoring**: Binary (20 points if present)
* **Rationale**: Enhances search result presentation

### 2. Content Quality Indicators

#### Content Length (30% of category)

* **Collection Method**: Text extraction and word count
* **Scoring**:
  + 100 points: ≥1500 words
  + 80 points: ≥1000 words
  + 60 points: ≥500 words
  + 40 points: ≥300 words
  + Below 300: Proportional scoring
* **Rationale**: Longer, comprehensive content typically ranks better

#### Heading Structure (20% of category)

* **Collection Method**: HTML heading tag analysis
* **Scoring**:
  + 40 points: Single H1 present
  + 30 points: H2/H3 tags present
  + 30 points: H4/H5/H6 tags present
* **Rationale**: Proper content hierarchy improves readability and SEO

#### Link Quality (20% of category)

* **Collection Method**: Anchor tag analysis
* **Scoring**:
  + 50 points: Optimal internal/external ratio (60-80% internal)
  + 30 points: Acceptable ratio (40-90% internal)
  + 50 points: ≥10 total links
  + 25 points: ≥5 total links
* **Rationale**: Balanced link structure improves site authority

#### Keyword Optimization (30% of category)

* **Collection Method**: Content analysis against target keywords
* **Scoring**:
  + 25 points each for:
    - Keyword in URL
    - Keyword in title
    - Keyword in meta description
    - Keyword in headers
* **Rationale**: Strategic keyword placement aids relevancy signals

### 3. User Experience Indicators

#### Loading Speed (40% of UX score)

* **Collection Method**: Page load time measurement
* **Scoring**:
  + 40 points: <2 seconds
  + 30 points: 2-3 seconds
  + 20 points: 3-4 seconds
* **Rationale**: Speed directly impacts user experience and rankings

#### Mobile Responsiveness (20% of UX score)

* **Collection Method**: Viewport meta tag detection
* **Scoring**: Binary (20 points if present)
* **Rationale**: Mobile optimization is crucial for modern SEO

#### Navigation Structure (20% of UX score)

* **Collection Method**: Navigation element analysis
* **Scoring**:
  + 40 points: Navigation menu present
  + 30 points: Footer present
  + 30 points: Forms present
* **Rationale**: Good navigation improves user engagement

#### SSL Security (20% of UX score)

* **Collection Method**: HTTPS verification
* **Scoring**: Binary (20 points if present)
* **Rationale**: Security affects user trust and rankings

### 4. Meta Elements Indicators

#### Title Tag (30% of meta score)

* **Collection Method**: HTML title tag analysis
* **Scoring**:
  + 100 points: 50-60 characters
  + 80 points: 40-70 characters
  + 60 points: 30-80 characters
  + 40 points: Other lengths
  + -30 points: Generic titles
* **Rationale**: Optimal title length improves CTR and rankings

#### Meta Description (30% of meta score)

* **Collection Method**: Meta description tag analysis
* **Scoring**:
  + 100 points: 150-160 characters
  + 80 points: 130-170 characters
  + 60 points: 110-190 characters
  + 40 points: Other lengths
  + -30 points: Generic descriptions
* **Rationale**: Well-crafted descriptions improve CTR

#### Social Tags (40% of meta score)

* **Collection Method**: Open Graph and Twitter card detection
* **Scoring**:
  + 50 points: Open Graph tags present
  + 50 points: Twitter cards present
* **Rationale**: Social optimization improves content sharing

## Bonus Points and Adjustments

### Analytics Tools (+5 points)

* **Collection Method**: Analytics script detection
* **Rationale**: Indicates active site monitoring and optimization

### SEO Tool Detection (Perfect Score)

* **Collection Method**: SEO plugin/tool detection
* **Scoring**: Automatic 100 points
* **Rationale**: Indicates active SEO management

### Tool-specific Bonuses

* Technical SEO:
  + Caching tools: +10 points
  + Micro tools: +5 points
* Content Quality:
  + Content tools: +10 points
  + Social tools: +5 points
* User Experience:
  + Micro tools: +5 points

## Classification Thresholds

The final score determines the SEO optimization classification:

* Most Probably Optimized: ≥75 points
* Probably Optimized: ≥45 points
* Probably Not Optimized: ≥20 points
* Most Probably Not Optimized: <20 points

## Implementation Details

The classifier uses BeautifulSoup and lxml for HTML parsing, making the analysis robust against different HTML structures. The system processes each indicator independently, allowing for granular scoring and detailed feedback.

Error handling ensures that individual indicator failures don’t prevent overall scoring, making the system resilient to parsing edge cases.

## Best Practices for Usage

1. Regular Updates
   * Check and update scoring weights periodically
   * Adjust thresholds based on industry standards
   * Monitor and validate classification accuracy
2. Data Collection
   * Implement proper error handling for network requests
   * Use appropriate timeouts for loading time measurements
   * Validate parsed data before scoring
3. Performance Optimization
   * Cache processed results when possible
   * Implement concurrent processing for multiple URLs
   * Use efficient parsing methods for large-scale analysis

## Conclusion

This SEO scoring system provides a comprehensive evaluation of website optimization while remaining flexible enough to accommodate evolving SEO best practices. The weighted scoring system ensures that critical factors have appropriate impact on the final score, while the detailed indicator breakdown provides actionable insights for improvement.