**Comparison Algorithm Documentation**

**Index**

1. **Introduction**  
   1.1 Overview of Document Comparison Pro  
   1.2 Purpose of the Comparison Algorithm
2. **Workflow Overview**  
   2.1 Document Upload and Validation  
   2.2 Primary Document Selection  
   2.3 Section-Wise Parsing  
   2.4 Change Detection  
   2.5 Similarity Score Calculation  
   2.6 Report Generation
3. **Detailed Comparison Algorithm**  
   3.1 Document Parsing and Section Identification  
   3.2 Primary Document Selection  
   3.3 Section-Wise Comparison  
   3.3.1 Section Matching  
   3.3.2 Change Detection  
   3.4 Detecting Structural Changes
4. **Similarity Score Calculation**  
   4.1 Overview of Similarity Score  
   4.2 How the Score Is Calculated  
   4.3 Factors Affecting the Similarity Score
5. **Report Generation**  
   5.1 Comparison Report Details  
   5.2 Summary of Differences  
   5.3 Similarity Scores  
   5.4 AI-Generated Insights
6. **Error Handling and Edge Cases**  
   6.1 File Format Errors  
   6.2 Insufficient Documents  
   6.3 Mismatched Sections
7. **Summary**

## **1. Introduction0**

The **Document Comparison Pro** software is designed to compare multiple documents by identifying changes, similarities, and differences. This software is particularly useful for comparing various versions of documents, detecting edits, and generating detailed reports that highlight the nature of those changes.

This documentation explains the inner workings of the **comparison algorithm** that powers the system, guiding users on how the software compares documents and calculates the similarity score.

## **2. Workflow Overview**

The document comparison process involves a series of well-defined steps:

1. **Document Upload and Validation**: Users upload multiple documents (PDFs, DOCXs), and the system validates their format.
2. **Primary Document Selection**: The first uploaded document is treated as the reference (primary document) for comparison.
3. **Section-Wise Parsing**: Documents are divided into sections (such as headings, chapters, paragraphs) for structured comparison.
4. **Change Detection**: The algorithm compares each section of the documents and detects changes (added, removed, or modified text).
5. **Similarity Score Calculation**: The software computes a similarity score for each document based on how closely it matches the primary document.
6. **Report Generation**: The system generates a detailed report, highlighting the changes, similarity scores, and summaries.

## **3. Detailed Comparison Algorithm**

### ****3.1 Document Parsing and Section Identification****

Upon receiving the uploaded documents, the software divides each document into manageable sections. These sections could represent:

* **Headings**: Chapter titles, subsections, etc.
* **Paragraphs**: Text within sections.
* **Other elements**: Lists, tables, and footnotes.

The algorithm automatically extracts the sections, ensuring a structured approach to comparison.

### ****3.2 Primary Document Selection****

* The **first document** uploaded by the user is considered the **Primary Document**.
* All other documents are compared against this primary reference.

### ****3.3 Section-Wise Comparison****

Each document is compared with the primary document on a **section-by-section** basis.

#### ****3.3.1 Section Matching****

The algorithm first ensures that corresponding sections (based on headers or positions) from different documents are matched. If a section is present in one document but missing in others, it is flagged for user review.

#### ****3.3.2 Change Detection****

Once sections are matched, the algorithm examines the content to detect changes. The following changes are identified:

* **Unchanged Text**: Content that is identical between the primary document and the compared document.
* **Added Text**: New content present in the compared document but not in the primary document.
* **Removed Text**: Content that was present in the primary document but is missing in the compared document.
* **Modified Text**: Content that has been edited or altered between the two documents.

Each change is highlighted using different markers:

* **Green**: Added text.
* **Red**: Removed text.
* **Blue**: Modified text.
* **Black**: Unchanged text.

### ****3.4 Detecting Structural Changes****

The algorithm also identifies structural changes such as:

* **Reordering of sections**: If sections in a document have been moved around, this is flagged.
* **New sections**: Sections that are added to the compared document but are missing from the primary document.

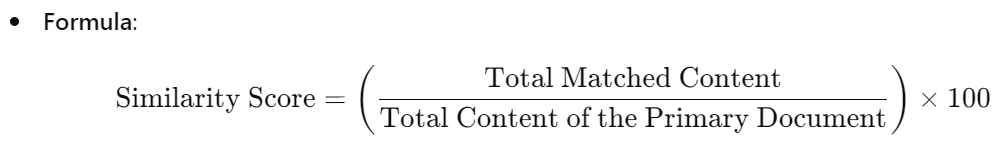
## **4. Similarity Score Calculation**

### ****4.1 Overview of Similarity Score****

The **similarity score** is a percentage-based metric that quantifies how similar the compared documents are to the primary document. This score is calculated based on the textual content, with higher scores indicating more similarity.

### ****4.2 How the Score Is Calculated****

The similarity score is calculated for each document by comparing its sections against the primary document. The process involves:

1. **Total Matched Content**: The number of sections and paragraphs in the compared document that are identical to those in the primary document.
2. **Total Content**: The total number of sections and paragraphs in the primary document.

For example, if a document has 90% of its sections matching the primary document, the similarity score will be **90%**.

### ****4.3 Factors Affecting the Similarity Score****

* **Number of Sections**: Documents with more sections will have a more detailed score.
* **Length of Content**: Longer sections (e.g., full chapters) have more weight in determining the score.
* **Degree of Modification**: Slight modifications to text may not significantly lower the score, while major edits or removals will cause a sharp decrease.

## **5. Report Generation**

After comparison, the software generates a **comparison report** that contains:

* **Document Comparison Details**: Side-by-side comparison of each section, with changes highlighted.
* **Summary of Differences**: A concise summary indicating the most significant changes.
* **Similarity Scores**: A similarity score for each compared document.
* **AI-Generated Insights**: Optionally, an AI summary can be provided for each document, offering a high-level overview of the changes.

## **6. Error Handling and Edge Cases**

### ****6.1 File Format Errors****

If a user uploads documents in incompatible formats (e.g., PDF and DOCX together), the software displays an error and requests the user to upload files of the same format.

### ****6.2 Insufficient Documents****

At least two documents are required for comparison. If fewer than two documents are uploaded, the software will alert the user and halt the process.

### ****6.3 Mismatched Sections****

In cases where sections cannot be matched between documents, the software flags them and presents this information in the comparison report.

## **7. Summary**

The **Document Comparison Pro** software offers a powerful and efficient way to compare documents section-by-section, providing detailed reports and similarity scores. This comparison algorithm is particularly useful for reviewing revisions, ensuring accuracy in content updates, and tracking document changes over time. The software is easy to use and offers clear feedback, making it an essential tool for users who need precise document comparison functionality.