**Chapter 3: Methodology**

**3.1 System Design and Architecture**

The Internal Plagiarism Detector is designed as a modular system with the following components:

***1. File Upload Module:*** Handles the upload of assignment submissions in various formats (e.g., `.docx`, `.pdf`, `.rtf`).

***2. File Conversion Module:*** Converts uploaded files into `.txt` format for uniform processing.

***3. Text Processing Module:*** Tokenizes the text and applies TF-IDF vectorization.

***4. Similarity Calculation Module:*** Computes cosine similarity scores between documents.

***5. Reporting Module:*** Generates detailed reports highlighting potential plagiarism.

***6. User Interface:*** Provides a user-friendly interface for academic professionals to interact with the system.

**3.2 File Conversion Process**

The file conversion module is a critical component of the system, as it ensures that all submissions are in a uniform format for processing. The module supports the following file formats:

- .docx: Microsoft Word documents.

- .pdf: Portable Document Format files.

- .rtf: Rich Text Format files.

The conversion process involves the following steps:

***1. File Upload:*** Users upload assignment submissions through the user interface.

***2. Format Detection:*** The system identifies the file format using file extensions and metadata.

***3. Conversion to .txt:*** The system converts the file content into plain text using appropriate libraries (e.g., `python-docx` for `.docx`, `PyPDF2` for `.pdf`).

***4. Text Cleaning:*** The converted text is cleaned to remove unnecessary characters, such as extra spaces, line breaks, and special symbols.

The workflow for file conversion is illustrated below:

***File Upload → Format Detection → Conversion to .txt → Text Cleaning***

**3.3 TF-IDF Vectorization**

The TF-IDF vectorization process transforms the text content of each document into a numerical representation, enabling similarity analysis. The process involves the following steps:

1. Tokenization: The text is split into individual words or tokens.

2. Term Frequency (TF) Calculation: The frequency of each term in the document is calculated.

3. Inverse Document Frequency (IDF) Calculation: The importance of each term across the entire corpus is determined.

4. TF-IDF Score Calculation: The TF and IDF scores are combined to generate the TF-IDF vector for each document.

The formula for TF-IDF is as follows:

***TF-IDF(t, d) = TF(t, d) × IDF(t)***

Where:

* ***t*** is a term.
* ***d*** is a document.
* **TF(*t*, *d*)** is the term frequency of *t* in *d*.
* **IDF(*t*)** is the inverse document frequency of *t*.

The TF-IDF vectorization process is illustrated below:

***Tokenization → TF Calculation → IDF Calculation → TF-IDF Vector***

**3.4 Cosine Similarity Calculation**

Cosine similarity is used to quantify the similarity between two documents based on their TF-IDF vectors. The process involves the following steps:

1. Vector Representation: Each document is represented as a TF-IDF vector.

2. Dot Product Calculation: The dot product of the two vectors is calculated.

3. Magnitude Calculation: The magnitude (Euclidean norm) of each vector is determined.

4. Cosine Similarity Score: The cosine similarity score is computed using the formula:

Cosine Similarity = **A**⋅**B ÷** ∥**A**∥∥**B**∥

Where:

* **A** and **B** are the TF-IDF vectors of two documents.
* **A**⋅**B** is the dot product of the vectors.
* ∥**A**∥ and ∥**B**∥ are the magnitudes of the vectors.

The cosine similarity calculation process is illustrated below:

***Vector Representation → Dot Product Calculation → Cosine Similarity***

**3.5 Similarity Score Interpretation**

The similarity scores generated by the system are interpreted as follows:

- 0.0: No similarity (completely original work).

- 0.1–0.4: Low similarity (common wording or structure).

- 0.4–0.7: Moderate similarity (potential plagiarism, requires investigation).

- 0.7–1.0: High similarity (strong evidence of plagiarism).

The table below provides a detailed interpretation of similarity scores:

|  |  |  |
| --- | --- | --- |
| **Score Range** | **Interpretation** | **Action** |
| 0.0 | No plagiarism detected | No action required |
| 0.1 – 0.4 | Common wording or structure | Review for context |
| 0.4 – 0.7 | Moderate similarity (potential plagiarism) | Investigate further |
| 0.7 – 1.0 | Strong evidence of plagiarism | Apply penalties as per institutional policy |