Advanced Document Analysis System

Objective

This innovative Document Analysis System is crafted to offer an advanced evaluation of text-based documents, combining state-of-the-art techniques in natural language processing and machine learning. The system provides insightful features such as plagiarism detection, similarity scoring, sentiment analysis, topic extraction, writing style evaluation, and customized feedback generation. Its primary goal is to empower users to improve content originality, readability, and emotional impact.

Features

1. **Plagiarism Detection**:
    - Employs advanced algorithms to detect overlapping content in the uploaded document.
    - Highlights plagiarized phrases and sentences with corresponding similarity percentages using enhanced semantic similarity checks.
2. **Similarity Scoring**:
    - Calculates a precise similarity score between the uploaded document and a reference text.
    - Outputs both numerical and visual representations of similarity.
3. **Sentiment Analysis**:
    - Evaluates the overall tone of the document using a combination of polarity, subjectivity, and contextual language models.
    - Provides labels such as Positive, Neutral, or Negative, alongside detailed sentiment insights.
4. **Grading System**:
    - Assigns grades based on originality, sentiment balance, and quality of writing:
        - Grade A: Exceptional Originality
        - Grade B: Good Originality
        - Grade C: Needs Improvement
        - Grade D: Poor Originality
5. **Writing Style Evaluation**:
    - Analyzes textual elements such as sentence structure, vocabulary richness, and linguistic balance.
    - Measures average sentence length, lexical diversity, and stylistic variety.
6. **Topic Extraction**:
    - Leverages advanced topic modeling algorithms to extract key themes and topics from the document.

    - Displays dominant topics with associated keywords for better thematic understanding.
7. **Customized Feedback**:
    - Generates personalized feedback that combines insights from all analyses.
    - Suggests actionable improvements for document originality, tone, and coherence.

Workflow

1. **File Upload**:
    - Users upload a `.txt` file directly into the system for analysis.
2. **Plagiarism Analysis**:
    - Sentences are analyzed against a database of reference materials for semantic and lexical overlap.
    - Results include flagged sections and an overall plagiarism percentage.
3. **Similarity Calculation**:
    - The system computes a similarity score between the document and a user-provided reference text.
    - Outputs both the percentage and areas of high similarity.
4. **Sentiment Evaluation**:
    - Analyzes emotional tone using hybrid NLP techniques.
    - Provides insights into how the document's tone impacts its readability and audience perception.
5. **Writing Style Analysis**:
    - Generates a detailed breakdown of linguistic elements, such as parts of speech, sentence lengths, and vocabulary variety.
    - Compares writing metrics against benchmarks for high-quality documents.
6. **Topic Identification**:
    - Applies Latent Dirichlet Allocation (LDA) to identify core topics.
    - Displays top words for each topic and their relevance to the document.
7. **Feedback Generation**:
    - Consolidates all findings into a tailored feedback report.
    - Suggests specific revisions and strategies for improvement.

Sample Output

Plagiarism Detection:

- **Flagged Sentences**:
    - "Artificial intelligence is a transformative technology for modern industries." (Match: 78%)

Similarity Score:

- **Document Similarity**: 72%

Grading:

- **Assigned Grade**: Grade B (Good Originality)

Sentiment Analysis:

- **Polarity**: 0.32
- **Subjectivity**: 0.48
- **Overall Sentiment**: Neutral

Writing Style Evaluation:

- **Word Count**: 420
- **Sentence Count**: 28
- **Average Sentence Length**: 15 words
- **Vocabulary Diversity**: 0.72

Topic Extraction:

- Topic 1: Artificial Intelligence, industries, innovation
- Topic 2: Data, learning, processes

Feedback:

- "Your document shows good originality with a neutral tone, suitable for informative writing. Writing style metrics indicate clarity and conciseness. However, consider improving vocabulary diversity and refining flagged sections for enhanced quality."

Implementation Details

Libraries and Tools Used:

- `spacy`: For advanced NLP tasks like tokenization and named entity recognition.
- `textblob`: For polarity and subjectivity sentiment analysis.
- `sklearn`: For TF-IDF vectorization, cosine similarity, and LDA topic modeling.
- `nltk`: For additional linguistic processing such as stopword removal.
- `matplotlib` and `seaborn`: For visualizing similarity scores and sentiment distribution.

Key Algorithms:

1. **Semantic Similarity Check**:
    - Combines TF-IDF and sentence embeddings to detect nuanced overlaps in content.
2. **LDA Topic Modeling**:
    - Extracts meaningful topics and keywords using probabilistic techniques.
3. **Sentiment Analysis**:
    - Leverages both lexicon-based and model-based methods for high accuracy.

Use Case Scenarios

1. **Academic Integrity**:
    - Enables educators to assess student submissions for originality and provide constructive feedback.
2. **Content Creation**:
    - Assists writers in crafting unique and engaging articles by identifying tone and style improvements.
3. **Corporate Documentation**:
    - Analyzes sentiment and relevance of reports, presentations, and proposals to align with business goals.

Future Enhancements

- Implement multilingual support for global use.
- Incorporate neural networks for even more accurate plagiarism detection.
- Add integration with cloud-based document storage for seamless file handling.

Conclusion

The Unique and Creative Document Analysis System redefines text evaluation with its comprehensive features and user-focused insights. It bridges the gap between advanced computational techniques and practical applications, making it a powerful tool for students, writers, and professionals alike.