### **Task 1: Advanced Data Cleaning with Fuzzy String Matching**

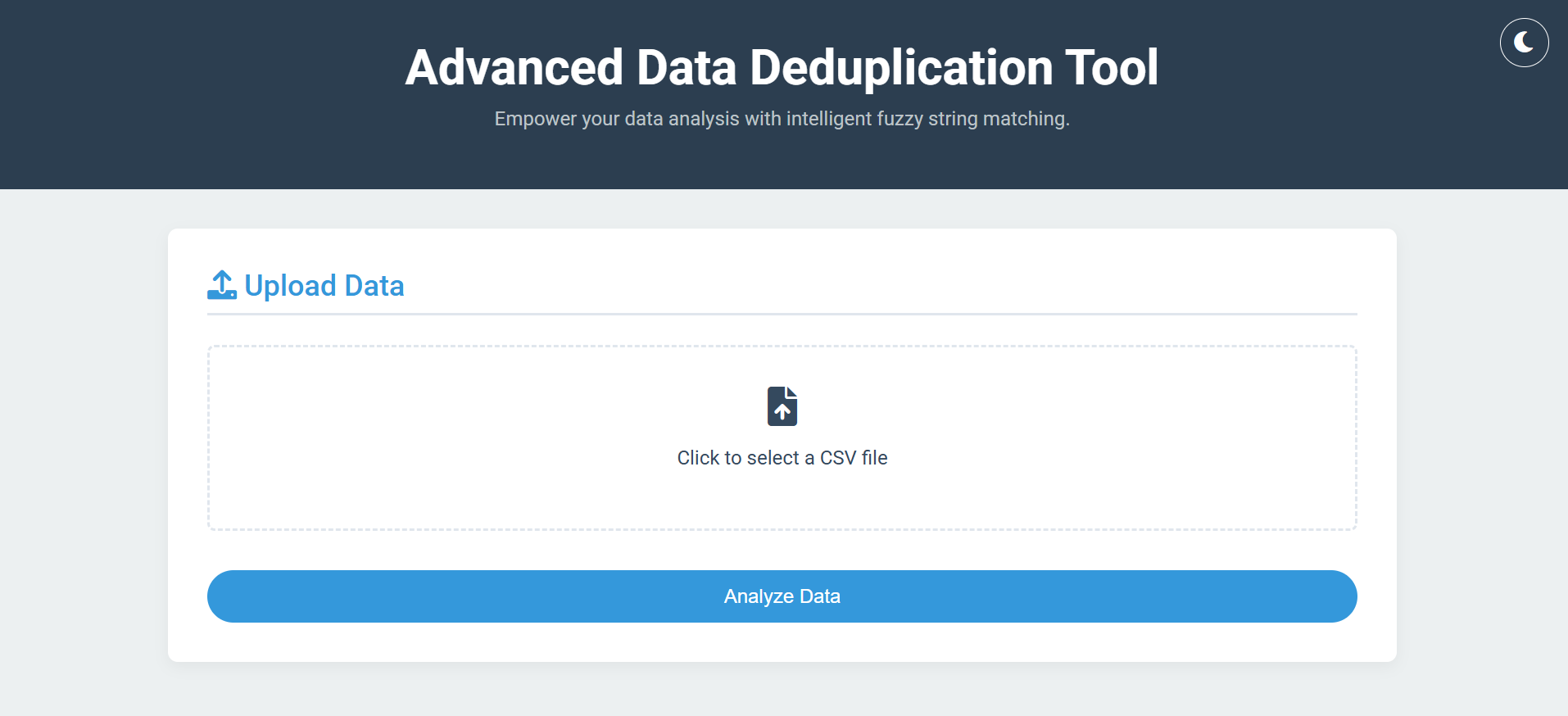
**Objective & Rationale** Data quality is paramount in any analytical project. This application was built to tackle one of the most common data integrity issues: inconsistent text entries. Whether due to typos ("Acme Corp" vs. "Acme Corp."), abbreviations ("Ave" vs. "Avenue"), or formatting inconsistencies, a simple, exact-match search often fails. Our goal was to create a tool that intelligently identifies these partial matches, or "fuzzy" duplicates.

**Our Approach** The application is structured around a robust pipeline built with pandas for data handling and the powerful thefuzz library for its string-matching algorithms.

1. **Automated Column Detection**: Upon uploading a CSV file, the backend intelligently scans the column headers for common keywords like 'Name', 'Title', or 'Publisher'. This makes the tool highly versatile and user-friendly, as it doesn't require manual configuration.
2. **Sophisticated Scoring**: The core of the deduplication process uses thefuzz.token\_sort\_ratio. This algorithm is particularly effective because it's insensitive to word order and minor punctuation differences. It processes "Acme Inc." and "Inc., Acme" by sorting their words alphabetically before comparing them, resulting in a very high similarity score.
3. **Grouped Output for Review**: The final result is a collection of tables, where each table represents a group of potential duplicates. This format allows a user to quickly and manually review the flagged entries, making informed decisions on how to clean and consolidate their data.

**Key Takeaway** This application transforms a tedious, manual data cleaning task into an automated and accessible process, providing a clear path to improving data quality and ensuring the reliability of downstream analysis.

**Github-Repo:**[**https://github.com/Kuldeep-Tapodhan/fuzzy-string-deduplication-django**](https://github.com/Kuldeep-Tapodhan/fuzzy-string-deduplication-django)  
  
**Output:**

****

