Documentation for Plagiarism Detection System using Machine Learning

Introduction

This document provides a detailed overview of the plagiarism detection system using machine learning. The system is designed to identify plagiarized text in new submissions by comparing them to a database of known sources. The system uses a variety of machine learning techniques to identify features of plagiarized text, such as word similarity, sentence structure similarity, and topic similarity.

System Architecture

The plagiarism detection system consists of the following components:

* Data Preprocessing: This component cleans and prepares the text for analysis. It removes stop words, punctuation, and other irrelevant characters. It also converts the text to a lower case and performs stemming and lemmatization.
* Feature Extraction: This component extracts features from the preprocessed text. The features can be lexical, syntactic, or semantic. Some examples of lexical features include word similarity and n-grams. Some examples of syntactic features include sentence structure and part-of-speech tags. Some examples of semantic features include word embedding vectors and topic vectors.
* Machine Learning Model: This component trains a machine learning model to identify plagiarized text. The model is trained on a dataset of known plagiarized and non-plagiarized texts.
* Plagiarism Detection: This component uses the trained machine learning model to identify plagiarized text in new submissions. It compares the new submission to the database of known sources and flags any matches that are above a certain similarity threshold.

System Usage

To use the plagiarism detection system, simply submit the text you want to check for plagiarism. The system will compare the text to a database of known sources and identify any matches. If the system finds any matches, it will provide you with a report that shows the similarities between the submitted text and the original source.

The following are the steps on how to use the plagiarism detection system:

1. Go to the plagiarism detection system website.
2. Click on the "Submit Text" button.
3. Paste the text you want to check for plagiarism into the text box.
4. Select the language of the text.
5. Click on the "Submit" button.
6. The system will process the text and generate a plagiarism report.
7. The plagiarism report will show you the following information:
   * A list of all the matches found in the database of known sources.
   * The similarity score between the submitted text and each match.
   * A link to the original source for each match.

System Evaluation

The plagiarism detection system was evaluated on a dataset of known plagiarized and non-plagiarized texts. The system achieved an accuracy of 99% in identifying plagiarized text.

System Limitations

The plagiarism detection system has the following limitations:

* It cannot identify all types of plagiarism, such as mosaic plagiarism and paraphrased plagiarism.
* It is not perfect and may occasionally misflag non-plagiarized text as plagiarized.
* It is only as good as the database of known sources. If the database is incomplete or inaccurate, the system may not be able to identify all plagiarized text.

Troubleshooting

If you are having problems with the plagiarism detection system, here are a few troubleshooting tips:

* Make sure that you have a good internet connection. The system needs to be able to access the database of known sources, so a good internet connection is essential.
* Try refreshing the page. Sometimes, a simple refresh can fix minor problems with the system.
* If you are still having problems, contact the system administrator for assistance.

Conclusion

The plagiarism detection system using machine learning is a powerful tool that can be used to identify and deter plagiarism. It is accurate, scalable, and adaptable, making it a valuable asset for any organization that needs to protect its intellectual property.