Plagiarism detection Using NLP (Natural Language Processing) with Smith-waterman algorithm

Prof.T.K. Thivakaran
Department of Information Technology, SRM Institute of Science and Technology, Ramapuram, Chennai.

Abhishek Uniyal | Somnath | Anah Veronica I
Department of Information Technology, SRM Institute of Science and Technology, Ramapuram, Chennai.

ABSTRACT:

The main objective of the project is to improve the accuracy of plagiarism detection by using the power of Natural Language Processing (NLP). The Smith-Waterman algorithm used in this project performs local sequence alignment to determine similar regions between the two strings taken into account. NLP can be used to analyse the text and compare if two articles mean the same as a whole. This method surpasses its contemporary plagiarism detection tool options which compares the texts of the article in question to the texts of the sources available.

Keywords: Artificial intelligence, NLP, Natural Language Processing, Plagiarism, Smith-Waterman algorithm, algorithm.

INTRODUCTION:

Plagiarism is a major problem that is considered as one form of intellectual property theft where the research and ideas in the original paper by an author are taken and misrepresented by another author as their own work. However, Plagiarism exists in various fields, usually in the form on text. The existing text-plagiarism detection tools use one of the two general types, external and intrinsic softwares. External plagiarism detection softwares work by analysing the texts in the strings of the given article to the texts in the strings of the sources available to detect the percentage of the identical words used. The approach however has its limitations. While it can easily detect copy & paste plagiarism, this can easily be overcome by word switch method which replaces the words with their synonyms. Powerful softwares that help in achieving this with the help of artificial intelligence exists. Plagiarism combined with the power of artificial intelligence makes it impossible to be detected by the existing systems. Intrinsic plagiarism detection tool does not compare the article to external documents. It analyses the article to detect any irregularity in the writing style of the author as a potential indicator of plagiarism. The project aims to solve this issue by using the smith-waterman algorithm, which performs local sequence alignment to determine regions of similarity between the two strings analysed and NLP is used to then determine if they mean the same as a whole which will help detect idea plagiarism.

LITERATURE REVIEW:

A detailed literature survey was conducted to know in detail about the existing systems and on-going research in the field of plagiarism detection to understand its approach, its advantages and its limitations.

I. PATTERN BASED SYSTEMS:

The paper “Overview and Comparison of Plagiarism tools” [1] studies and compares the existing systems and approaches in plagiarism detection softwares. The methods discussed in the paper to detect textual plagiarism are:

- Grammar based: Analyses the grammatical structure of the document
- Semantic based: Detects similarities using a vector space model

- Grammar semantics hybrid: Incorporates both grammar and semantic based approach
- External: Compares against a database of documents
- Clustering method: This method is used to cluster the documents to reduce search time.

Some of the tools discussed are plagAware and plagScan, which work with classic search engines for detecting and scanning to compare the texts of a given article against a database of sources considered as genuine. CheckforPlagiasrim,org works by analysing the document and creating the fingerprint for it. The document is given numerical attributes based on its structure and style of writing. This fingerprint of the document can then be run against other recorded fingerprints to identify similarities.

The paper "AntiPlag: Plagiarism detection on electronic submissions of text-based assignments" documents the research conducted by using tri-gram sequence matching technique. N-gram is a technique in computational linguistics where a contiguous sequence of n items of a given sample text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the need. This project uses trigram, where three sets of text-based items were assigned.

II. NLP BASED SYSTEMS:

The paper "Using Natural Language Processing for Automatic Detection of Plagiarism" uses NLP technique in external plagiarism detection in not only to analyse text but also their structure against the structure of its sources. The methodology used in this paper include segmentation of sentences, tokenization of punctuation symbols and number replacement to other symbols to generalise the comparison. Lower case conversion and parts of speech tagging to analyse structure. The text is also transformed to their stems (stemming) and dictionary base forms (lemmatization) to generalise the comparison analysis.

LITERATURE CONCLUSION:

The above papers are based of analysing and comparing documents based of text string and structure or documents. Our project aims to analyse and compare the two documents to determine to what degree the two documents mean the same.

IMPLEMENTATION:

This project uses Smith-waterman algorithm in its implementation. The algorithm works by performing local sequence alignment to determine similar regions in the nucleic acid sequence of the DNA. The similar technique is used in this project to determine similar regions in texts as indicators of plagiarism. The algorithm works in the following steps:

1. **Substitution matrix:** In this step a substitution matrix is made with the given text to analyse for match and mismatch. Matches get positive score and mismatches get negative score. This comes with a gap penalty function to determine score cost for opening or extending gaps.
2. **Initialisation of scoring matrix:** The scoring matrix is initialised with its dimensions set to 1 + length of each sequence. The value is the set to 0.
3. **Scoring:** The elements are scored in left to right, top to bottom manner based on the outcomes of the substitutions. If no element matches the score is set to 0, else the highest number is recorded.
4. **Traceback:** A traceback is done starting from the highest element until a zero is found. The highest score is then recorded. The same procedure is done again to record the second highest element and so on.

The project consists of three modules. The first module segments the text and the elements are used as identifiers. The error modules to used to deal with any unrecognised characters. The third module consists of source documents that the texts from the document in question will be compared against.

FUTURE ENHANCEMENTS:

To develop a Machine Learning model that can get long information in research papers and delivers an exact summary in compressed form. It will help to read and compare multiple research papers at a time. Google Tensorflow uses a deep learning technique called “Sequence to sequence learning” to summarise shorter texts. A learning model can be built over it for summarising longer paragraphs.

CONCLUSION:

In this project we attempted to increase the accuracy in the detection of text-based plagiarism. After researching about the existing systems, methodologies, using smith-waterman algorithm and NLP. The techniques and procedures used in making the system helped prove it to be a better option compared to the existing solution.

REFERENCES:

[1] Overview and Comparison of Plagiarism Detection Tools by Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and V´aclav, Department of Computer Science, VSB-Technical University of Ostrava

[2] Using Natural Language Processing for Automatic Detection of Plagiarism by Miranda Chong, Lucia Specia, Ruslan Mitkov

[3] https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm

[4] https://en.wikipedia.org/wiki/Plagiarism