BAHCESEHIR UNIVERSITY

CMP 3005 – Analysis of Algorithms

Project Report

2021

Ezgi CUĞ 1729778

Arsen DÜZGÜN 1602980

Berk ÇATALAHMETOĞLU 1736771

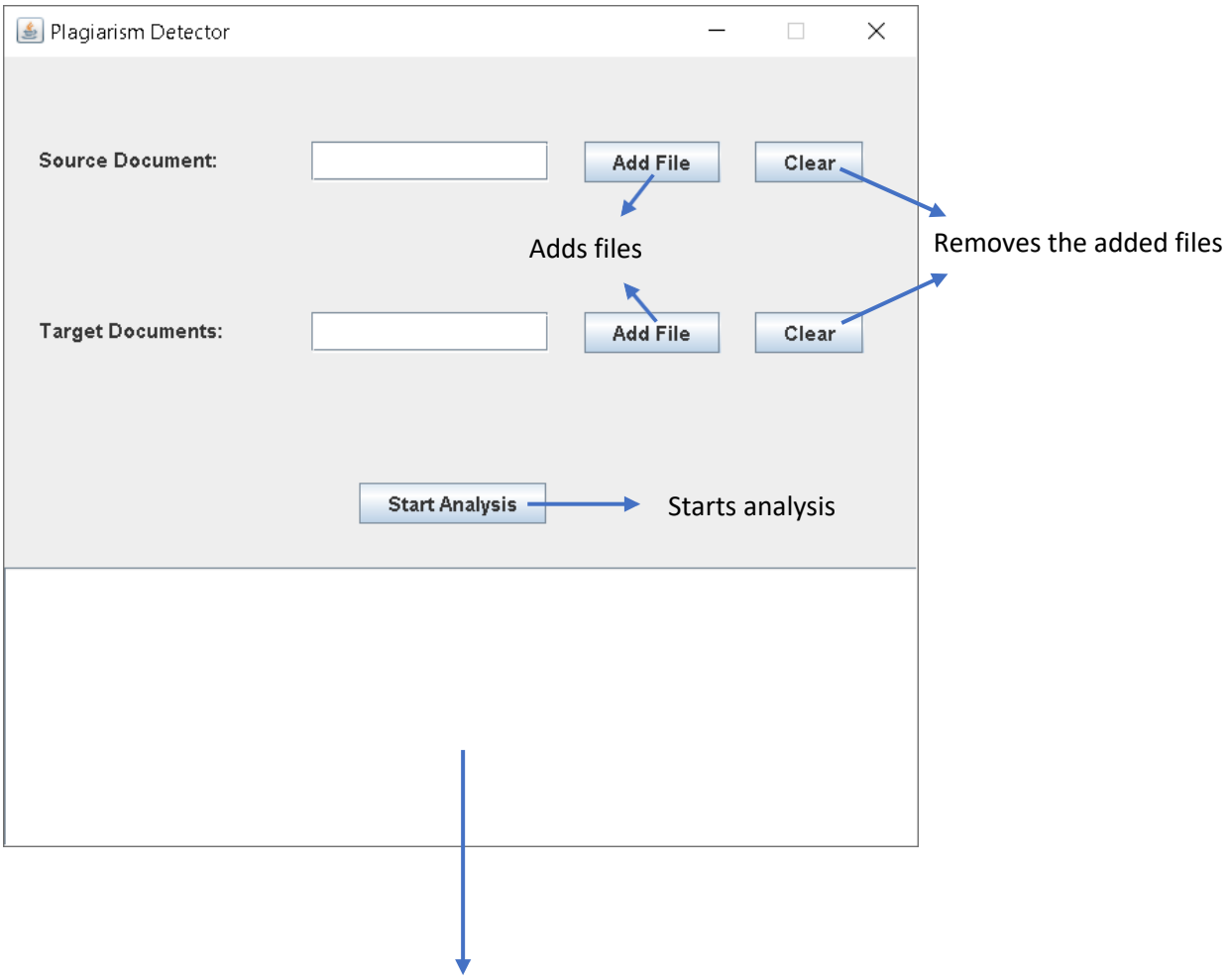
Course Instructor

Assistant Professor Cemal Okan ŞAKAR

TABLE OF CONTENTS

1-) Definition and Overview of the Project

This project is designed to detect plagiarism and calculate the similarity rate between source text document and target text documents.



2-) Process of Steps in the Project

Before analyzing, it is expected from user to add one source text document (thought to be plagiarized) and one or more text documents (thought to be used for plagiarism) to the application via the interface.

After clicking the Start Analysis button, the program matches the source text document separately with all the target text documents added and creates a thread for each match to run concurrently.

The source text document is divided into sentences and then those sentences are divided into sub-sentences with at least 3 words without breaking the word order of the sentence. The target text document or documents to be compared are divided into sentences.

The similarity rate between any sentence of the source text document and any sentence in the target text document is calculated as follows:

All sub-sentences of the source text document sentence are searched within the sentence in the target text document. All matching sub-sentences are collected in a list. After the matching is completed, the size of the plagiarized sub-sentences (in terms of the number of words it contains) and the size of the source text document sentence (in terms of the number of words it contains) are compared, and the similarity rate between the sentence of the source text document and the sentence in the target text document is calculated.

The similarity rate between the source text document and the target text document is calculated as follows:

The total plagiarism size of a source text document sentence is obtained by comparing the sentence of the source text document with all the sentences in the target text document. The total plagiarism size of the source text document is obtained by summing the total plagiarism size values calculated for each sentence of the source text document. Then, the total plagiarism size of the source text document is compared with the total size of the source text document and the similarity rate between the source text document and the target text document is calculated.

3-) String Matching Algorithm used in the Project

Horspool string matching algorithm is used in the project.

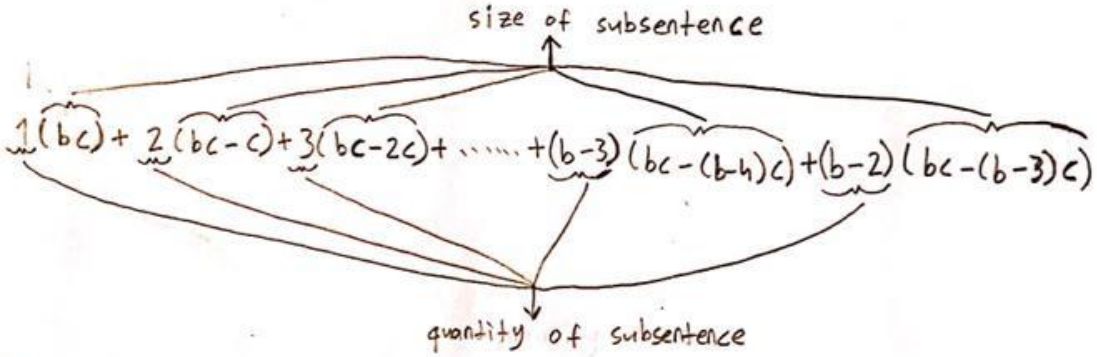
Horspool is an algorithm for finding substrings into strings. This algorithm compares each characters of substring to find a word or the same characters into the string. When characters do not match, the search jumps to the next matching position in the pattern by the value indicated in the shift table. Shift table indicates how many jumps should it move from the current position to the next.

Worst-case time complexity of the horspool algorithm is O(mn), where m is the length of the substring and n is the length of the string, and the average time complexity is O(n).

4-) Worst-Case Time Complexity of a Project

Source document: Has k sentence. Each sentence has b word, and each word has c letter.

Subsentences of source document sentence:



Target document: Has d sentence. Each sentence has e letter.

Worst-case time complexity:

$$kde\left(bc + 2(bc-c) + 3(bc-2c) + \dots + (b-3)(bc-(b-4)c) + (b-2)(bc-(b-3)c)\right)$$

$$(bc + 2bc + 3bc + \dots + (b-3)bc + (b-2)bc) - (1*2c + 2*3c + 3*4c + \dots + (b-4)(b-3)c + (b-3)(b-2)c)$$

$$\frac{bc(b-2)(b-1)}{2} - \frac{(b-3)(b-2)(b-1)c}{3} = \frac{b^3c + 3b^2c - 16bc + 12c}{6}$$

worst-case time complexity is $\longrightarrow O(kdeb^3c)$

*If multiple target documents are compared with source document, the largest target document in terms of number of letters among all target documents is considered

5-) Programming Languages and Frameworks used in the Project

Programming Languages used in the project: Java

Frameworks used in the project:

Java GUI Frameworks:

javax.swing.JButton

javax.swing.JFileChooser

javax.swing.JFrame

javax.swing.JLabel

javax.swing.JOptionPane

javax.swing.JScrollPane

javax.swing.JTextArea

javax.swing.Timer

Java Collections Frameworks:

java.util.ArrayList

java.util.HashMap

java.util.HashSet

java.util.Collections

java.util.Scanner

Java Input/Output Frameworks:

java.io.File

java.io.FileNotFoundException

Java Event Frameworks:

java.awt.event.ActionEvent

java.awt.event.ActionListener

6-) Exemplary Outputs of Analysis

1-) Example 1

Source text document:

The European Values Study is large-scale, time-intensive survey on basic human values. It provides insights into the values, beliefs and preferences of citizens all over Europe. It is a unique research project on how Europeans think about life, family, work, religion, politics and society. The European Values Study was launched in 1981, when a couple of hundred citizens in the European Member States were interviewed using standardized questionnaries. Every nine years, the survey is repeated in an increasing number of countries.

Not all the respondents of the original data sample are included in the analysis. People who did not answer one or more of the questions included, are filtered out of dataset. The final number of respondent has been brought down to a sample analysis of 60077 respondents.

Target text document:

The European Values Study is a large-scale, cross-national, and longitudinal survey research program on basic human values. It provides insights into the ideas, beliefs, preferences, attitudes, values and opinions of citizens all over Europe. It is a unique research project on how Europeans think about life, family, work, religion, politics and society.

The European Values Study started in 1981, when a thousand citizens in the European Member States of that time were interviewed using standardized questionnaires. Every nine years, the survey is repeated in an increasing number of countries. The fourth wave in 2008 covers no less than 47 European countries/regions, from Iceland to Azerbaijan and from Portugal to Norway. In total, about 70,000 people in Europe are interviewed.

Output of analysis:

Similarity Rate between source.txt and target.txt is 53.125%

Most similar sentence is "It is a unique research project on how Europeans think about life, family, work, religion, politics and society." in document named source.txt, in paragraph 1, and sentence 3, and it has 100.0% similarity rate with the sentence "It is a unique research project on how Europeans think about life, family, work, religion, politics and society." in document named target.txt, in paragraph 1, and sentence 3. Plagiarised parts: "It is a unique research project on how Europeans think about life, family, work, religion, politics and society.".

Second most similar sentence is "Every nine years, the survey is repeated in an increasing number of countries." in document named source.txt, in paragraph 1, and sentence 5, and it has 100.0% similarity rate with the sentence "Every nine years, the survey is repeated in an increasing number of countries." in document named target.txt, in paragraph 2, and sentence 2. Plagiarised parts: "Every nine years, the survey is repeated in an increasing number of countries.".

Third most similar sentence is "The European Values Study is large-scale, time-intensive survey on basic human values." in document named source.txt, in paragraph 1, and sentence 1, and it has 75.0% similarity rate with the sentence "The European Values Study is a large-scale, cross-national, and longitudinal survey research program on basic human values." in document named target.txt, in paragraph 1, and sentence 1. Plagiarised parts: "The European Values Study is", "on basic human values.".

Fourth most similar sentence is "The European Values Study was launched in 1981, when a couple of hundred citizens in the European Member States were interviewed using standardized questionnaries." in document named source.txt, in paragraph 1, and sentence 4, and it has 75.0% similarity rate with the sentence "The European Values Study started in 1981, when a thousand citizens in the European Member States of that time were interviewed using standardized questionnaires." in document named target.txt, in paragraph 2, and sentence 1. Plagiarised parts: "citizens in the European Member States", "The European Values Study", "in 1981, when a", "were interviewed using standardized".

Fifth most similar sentence is "It provides insights into the values, beliefs and preferences of citizens all over Europe." in document named source.txt, in paragraph 1, and sentence 2, and it has 71.42857% similarity rate with the sentence "It provides insights into the ideas, beliefs, preferences, attitudes, values and opinions of citizens all over Europe." in document named target.txt, in paragraph 1, and sentence 2. Plagiarised parts: "It provides insights into the", "of citizens all over Europe.".

Elapsed time ≈ 8ms – 12ms

2-) Example 2

Source text document:

ABC College for Women is one of the most prestigious institutions of London with a full time enrollment of about 8000 students. The glorious academic values of this oldest premier post-graduate female institution have been shaped by its institutional history, which is spread over a span of 64 years. In january 2002, the University made all strong decisions for the improvement in Higher Education. Established in May 1962 as an Intermediate residental college and affiliated with the University of the Oxford, it was housed in a building on XYZ Road, with strength of 90 students and then the progress flourished with full shot. And College started programs like Electronics, Environmental Science, Fine Arts, Economics and Mass Communication. Various national industries and linkages with foreign Colleges helped a lot...

Target text document:

Since the establishment of ABC College for Women and in early January 2002, the University has tried its level best for improvement in Higher Education. Government did various national industries and linkages with foreign universities MoU with various national industries and linkages with foreign universities have been established in the field of Pharmacy, Electronics, Environmental Science, Fine Arts, Economics and Mass Communication. This is how they made the glorious academic values of this oldest premier post-graduate female institution very nicely.

Output of analysis:

Similarity Rate between source.txt and target.txt is 30.46875%

Most similar sentence is "And College started programs like Electronics, Environmental Science, Fine Arts, Economics and Mass Communication." in document named source.txt, in paragraph 1, and sentence 5, and it has 64.28571% similarity rate with the sentence "Government did various national industries and linkages with foreign universities MoU with various national industries and linkages with foreign universities have been established in the field of Pharmacy, Electronics, Environmental Science, Fine Arts, Economics and Mass Communication." in document named target.txt, in paragraph 1, and sentence 2. Plagiarised parts: "Electronics, Environmental Science, Fine Arts, Economics and Mass Communication.".

Second most similar sentence is "Various national industries and linkages with foreign Colleges helped a lot..." in document named source.txt, in paragraph 1, and sentence 6, and it has 63.636364% similarity rate with the sentence "Government did various national industries and linkages with foreign universities MoU with various national industries and linkages with foreign universities have been established in the field of Pharmacy, Electronics, Environmental Science, Fine Arts, Economics and Mass Communication." in document named target.txt, in paragraph 1, and sentence 2. Plagiarised parts: "Various national industries and linkages with foreign".

Third most similar sentence is "In january 2002, the University made all strong decisions for the improvement in Higher Education." in document named source.txt, in paragraph 1, and sentence 3, and it has 53.333336% similarity rate with the sentence "Since the establishment of ABC College for Women and in early January 2002, the University has tried its level best for improvement in Higher Education." in document named target.txt, in paragraph 1, and sentence 1. Plagiarised parts: "january 2002, the University", "improvement in Higher Education.".

Fourth most similar sentence is "The glorious academic values of this oldest premier post-graduate female institution have been shaped by its institutional history, which is spread over a span of 64 years." in document named source.txt, in paragraph 1, and sentence 2, and it has 40.74074% similarity rate with the sentence "This is how they made the glorious academic values of this oldest premier post-graduate female institution very nicely." in document named target.txt, in paragraph 1, and sentence 3. Plagiarised parts: "The glorious academic values of this oldest premier post-graduate female institution".

Fifth most similar sentence is "ABC College for Women is one of the most prestigious institutions of London with a full time enrollment of about 8000 students." in document named source.txt, in paragraph 1, and sentence 1, and it has 18.181818% similarity rate with the sentence "Since the establishment of ABC College for Women and in early January 2002, the University has tried its level best for improvement in Higher Education." in document named target.txt, in paragraph 1, and sentence 1. Plagiarised parts: "ABC College for Women".

Elapsed time ≈ 16ms – 22ms