Pattern Discovery of Sequential Symbolic Data using Automata with an application to Author Identification

Nikhil Kalantri
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects
Part of the Computer Sciences Commons

Recommended Citation
Kalantri, Nikhil, "Pattern Discovery of Sequential Symbolic Data using Automata with an application to Author Identification" (2013). Master's Projects. 333.
DOI: https://doi.org/10.31979/etd.uhdr-ae3z
https://scholarworks.sjsu.edu/etd_projects/333

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Pattern Discovery of Sequential Symbolic Data using Automata with an application to Author Identification

A Thesis
Presented to
The Faculty of the Department of Computer Science
San José State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by
Nikhil Kalantri
December 2013
SAN JOSE STATE UNIVERSITY

The Designated Thesis Committee Approves the Thesis Titled

Pattern Discovery of Sequential Symbolic Data using Automata with an application to Author Identification

by

Nikhil Kalantri

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

December 2013

________________________________________________________
Dr. T. Y. Lin, Department of Computer Science Date

________________________________________________________
Dr. Chris Tseng, Department of Computer Science Date

________________________________________________________
Mr. Amit Sant, Software Engineer at Apple Date
ABSTRACT

Author Identification is the process of identifying a piece of text to ascertain if it has an inherent writing style or pattern based on a certain author. Almost all literary books can be accredited to a certain author since it has been signed. However, there also exist a plethora of unfinished books or manuscripts that could be attributed to a range of possible authors. For example, William Shakespeare has written many plays that have not been signed by him. In order to assess the importance of such texts that do not bear the authors signature, it could be vital to know who was the writer. I plan to solve this dilemma using the characteristics of finite state automata coupled with the ALERGIA algorithm.
ACKNOWLEDGEMENTS

I would like to acknowledge and extend my heartfelt gratitude to the following people who have made the completion of this project successful. I thank my project advisor Dr. T.Y. Lin, for the vital encouragement and support. You have been a tremendous mentor throughout. Working on this project was a challenging experience. Your advice and guidance will always be with me and help me grow as an individual.

I would also like to thank my committee members Dr. Tseng and Mr. Amit Sant for their support and patience. I also want to thank you for being part of this exciting journey.

I want to thank our department for providing us with the necessary software required in our project. I’m also thankful to the library for providing necessary books and materials required to learn different concepts for our project.

A special thanks to my family. Words cannot express how grateful I am to my mother, father and sister for all the sacrifices you have made for me. Your prayers and blessings have helped me sustain this far. At the end, I would like to express appreciation to my incredible friends Onkar, Akanksha and Mrinmai who always stood by me whenever I needed them.
# TABLE OF CONTENTS

1. Introduction ................................................................................................................................. 1  
   1.1. What is author identification? ................................................................................................. 1  
   1.2. Why do we need author identification? .................................................................................. 1  
   1.3. Why use computational/statistical methods? .......................................................................... 1  
   1.4. Attributes of a document....................................................................................................... 1  
   1.5. Role of automata theory ......................................................................................................... 2  
2. Finite State Automata .................................................................................................................... 3  
   2.1. Deterministic Finite Automaton ............................................................................................. 3  
   2.2. Stochastic Finite State Automata ............................................................................................ 4  
3. ALERGIA Algorithm .................................................................................................................... 6  
4. Analyzing text using automata based modeling ........................................................................... 9  
5. Test Results .................................................................................................................................. 12  
6. Future Work .................................................................................................................................. 41  
7. Conclusion .................................................................................................................................... 42  
8. References ..................................................................................................................................... 43  
APPENDIX A: Development Environment ....................................................................................... 44  
APPENDIX B: List of EBook’s used .................................................................................................... 45
LIST OF TABLES

Table 1: Result for test case ID: 01.................................................................12
Table 2: Result for test case ID: 02.................................................................13
Table 3: Result for test case ID: 03.................................................................14
Table 4: Result for test case ID: 04.................................................................15
Table 5: Result for test case ID: 05.................................................................16
Table 6: Result for test case ID: 06.................................................................17
Table 7: Result for test case ID: 07.................................................................18
Table 8: Result for test case ID: 08.................................................................19
Table 9: Result for test case ID: 09.................................................................20
Table 10: Result for test case ID: 10...............................................................21
Table 11: Result for test case ID: 11...............................................................22
Table 12: Result for test case ID: 12...............................................................24
Table 13: Result for test case ID: 13...............................................................26
Table 14: Result for test case ID: 14...............................................................28
Table 15: Result for test case ID: 15...............................................................30
Table 16: Result for test case ID: 16...............................................................32
Table 17: Result for test case ID: 17...............................................................34
Table 18: Result for test case ID: 18...............................................................36
Table 19: Result for test case ID: 19...............................................................38
Table 20: Result for test case ID: 20...............................................................40
Table 21: Software Specification......................................................................44
Table 22: Hardware Specification......................................................................44
Table 23: List of Ebook’s.................................................................................45
LIST OF FIGURES

Figure 1: State Transition Table.................................................................3
Figure 2: State Transition Diagram..........................................................4
Figure 3: PTA tree for sample string S......................................................9
Figure 4: PTA after merging q_2 and q_1....................................................10
Figure 5: PTA after merging q_5 and q_1....................................................10
Figure 6: PTA after merging q_6 and q_3....................................................11
1. Introduction

1.1. What is author identification?
It is the process of identifying the creator of a written text through computational, statistical analysis. This analysis assists in capturing an author’s inherent writing style and this pattern can be used to find the source of an unsigned document.

1.2. Why do we need author identification?
Author identification is an important problem in many areas ranging from information retrieval and computational linguistics to journalism and law where this could potentially help in saving lives like discovering the author of a ransom note.

1.3. Why use computational/statistical methods?
Every author has a unique style of writing just like a human fingerprint. The human eye cannot recognize or pick up all the varying aspects of a document. Computational methods allow and aid humans to improve pattern analysis by exploring and uncovering these hidden traits of documents. A famous example to identify authors was shown by Professor Arthur Kinney in 2006. He proves that all unsigned Shakespeare documents or plays that were attributed to him, were indeed his with the help of statistical analysis.

1.4. Attributes of a document
Attributes are divided into four broad categories – Lexical, syntactic, structural and content-specific. These attributes help differentiate between authors. A few examples for each of the attributes are given below.

- Lexical: average number of words in a sentence, length of the word, total words.
- Syntactic: punctuations.
- Structural: font types, headers, footers, paragraph style.
- Content-specific: Number of stop words or abbreviations, gender or age based words.
1.5. Role of automata theory

The objective of this paper is to analyze sample texts based on automata [5][12] theory. This is achieved by generating a prefix tree acceptor by filtering out the stop words in a book and then applying the Alergia algorithm to check the compatibility of corresponding states. The algorithm regenerates the PTA iteratively through merging all compatible or equivalent states.
2. Finite State Automata

2.1. Deterministic Finite Automaton

Definition: A deterministic finite automaton consists of the following parameters:

- A finite set of states denoted by \( Q \)
- A finite set of symbols \( \Sigma \)
- A transition function that takes a state and a symbol as arguments and returns a state. It is denoted by \( \delta \).
- The start state denoted by \( q_0 \)
- Set of final or accepting states denoted by \( F \)

Therefore, we have \( q_0 \in Q \) and \( F \subseteq Q \).

So a DFA is mathematically represented as a 5-uple \((Q, \Sigma, \delta, q_0, F)\).

The transition function \( \delta \) is a function in \( Q \times \Sigma \rightarrow Q \)

\( Q \times \Sigma \) is the set of 2-tuples \((q, a)\) with \( q \in Q \) and \( a \in \Sigma \)

A DFA with a transition table is given as

<table>
<thead>
<tr>
<th>State ( q )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_0 )</td>
<td>( q_2 )</td>
<td>( q_0 )</td>
</tr>
<tr>
<td>( q_1 )</td>
<td>( q_1 )</td>
<td>( q_1 )</td>
</tr>
<tr>
<td>( q_2 )</td>
<td>( q_2 )</td>
<td>( q_1 )</td>
</tr>
</tbody>
</table>

Figure 1: State transition table
This transition table defines the following transition diagram,

![State Transition Diagram](image)

**Figure 2: State Transition Diagram**

Therefore,

\[ Q = \{ q_0, q_1, q_2 \} \]

Start state \( q_0 \)

\[ F = \{ q_1 \} \]

\[ \Sigma = \{ 0, 1 \} \]

\( \delta \) is a function from \( Q \times \Sigma \) to \( Q \)

\[ \delta: Q \times \Sigma \rightarrow Q \]

\[ \delta(q_0, 1) = q_0 \]

\[ \delta(q_0, 0) = q_2 \]

### 2.2. Stochastic Finite State Automata

A stochastic finite state automaton [9] provides transition probabilities to each of the next states in addition to providing the finite state automata [5][12] for the given input.

For example, consider input symbols \( b_1, b_2 \). Now, there is a possibility of two arbitrary transitions \( \delta(q, b_1) \) or \( \delta(q, b_2) \). SFA helps us in analyzing and evaluating the probability of a transition to each of the states.

The probability function to calculate arbitrary transitions is given by,

\[ p_{lf} + \sum_{q_j \in Q} \sum_{a \in A} p_{lj}(a) = 1 \]
This shows that the sum of probabilities that start and end at node $q_i$ is always equal to 1.

The language generated by stochastic finite automata [9] is known as stochastic regular language (SRL).
3. ALERGIA Algorithm

The Alergia algorithm specializes in merging the states of a generated automaton from a probabilistic point of view. Alergia is a learning algorithm. Consider a sample set containing duplicate strings; the algorithm can learn its Deterministic Frequency Finite Automata [5] and also the Deterministic Probabilistic Finite Automata [5].

When the probability of appearance of a string follows a well-defined distribution, Alergia has the ability to take advantage of this and merge states when the resulting automaton is compatible with the observed frequency of strings.

First the algorithm generates a prefix tree from the input strings and analyzes the relative frequency of outgoing arcs at every node. The prefix tree captures this information.

Let \( n_i \) be the number of strings arriving at node \( q_i \).

\( f_i(a) \) : Number of strings following arc \( \delta(q_i, a) \)

\( f_i(\#) \) : Number of strings terminating at node \( q_i \)

Calculate the following probabilities:

\[ p_i(a) = \frac{f_i(a)}{n_i} \]

\[ p_{if} = \frac{f_i(\#)}{n_i} \]

The algorithm compares corresponding nodes \( (q_i, q_j) \). The value of \( j \) varies from 2 to \( t \) and \( i \) varies from 1 to \( j-1 \).

When the probabilities of two corresponding states are equal, they are considered equivalent and this rule applies to their corresponding children.
If the difference between the probabilities of the two states is less than the acceptance range $\alpha$, these states are considered as compatible. Recursively, the child nodes are also considered compatible.

A false value will be returned if the probability difference is greater than the acceptance rate. The formula to compare two states is given by the Hoeffding bound:

$$\left| \frac{f_i}{n} - \frac{f_j}{n'} \right| < \sqrt{\frac{1}{2} \log \frac{2}{\alpha} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n'}} \right)}$$

There are 3 algorithms that we consider:

Algorithm **COMPATIBLE**

Input:
- $i,j$ : nodes

Output:
- Boolean

Begin
  If different ($n_i, f_i(#), n_j, f_j(#)$)
    Return false
  Endif
  Do ($\forall a \in A$)
    If different ($n_i, f_i(a), n_j, f_j(a)$)
      Return false
    End if
  End do
  If not compatible ($\delta(i,a), \delta(j,a)$)
    Return false
  End if
End algorithm
Algorithm **DIFFERENT**

Input:
- \( n, n' \): number of strings arriving at each node.
- \( f, f' \): number of strings ending or following a given arc

Output:
- Boolean

Begin

\[
\text{Return } \left| \frac{f}{n} - \frac{f'}{n'} \right| < \sqrt{\frac{1}{2} \log \frac{2}{\alpha} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n'}} \right)}
\]

End Algorithm

Algorithm **ALERGIA**

Input:
- \( S \): sample set of strings
- \( \alpha \): 1-confidence level

Output:
- Stochastic DFA

Begin

\( A = \text{stochastic Prefix Tree Acceptor from } S \)

Do (for \( j = \text{successor (first node (A))} \) to last node (A))

Do (for \( I = \text{first node (A)} \) to \( j \))

If compatible \((I, j)\)

Merge \((A, i, j)\)

Determinize \((A)\)

Exit (i-loop)

End if

End for

End for

Return \( A \)

End algorithm
4. Analyzing text using automata based modeling

Consider an input string,

\[ S = \{110, -, -, 0, -, -, 00, -, -, 0, 00, -, -, 100, -, -, 10110\} \]

Let \( \alpha = 0.8 \)

Step 1: Build the Prefix Tree Acceptor tree

Therefore, \( \Upsilon = \sqrt{\frac{1}{2} \log \frac{2}{\alpha}} \approx 0.67 \)

Every arc for each transition has a label with 0 or 1 and the number of strings in the input using that arc is shown in brackets. Then the algorithm checks for the equivalence of corresponding nodes. This is achieved by comparing their SFA probabilities.

Step 2: Minimize the states using the Hoeffding bound.

We generate the Deterministic Frequency Finite Automaton by applying the algorithm to merge compatible nodes. After merging thrice with \( \alpha = 0.8 \), we get

![Figure 3: PTA tree for sample string S][9]
Figure 4: PTA after merging $q_2$ and $q_1$ [9]

Figure 5: PTA after merging $q_5$ and $q_1$ [9]
Figure 6: PTA after merging $q_6$ and $q_3$ [9]
5. Test Results

Test case ID: 01

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: JK Rowling – HP0.txt

Test books:

- JK Rowling – HP0.txt
- JK Rowling – HP1.txt
- James Matthew Barrie - Peter Pan.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>96.949</td>
<td>89.933</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.608</td>
<td>71.283</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>67.767</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>53.931</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.822</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>33.446</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>68.831</td>
<td>29.595</td>
</tr>
</tbody>
</table>

Table 1: Result for test case ID: 01

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by J.K Rowling have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 02

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: JK Rowling – HP0.txt

Test books:
- JK Rowling – HP0.txt
- JK Rowling – HP1.txt
- Dante Alighieri - The Divine Comedy.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>96.949</td>
<td>69.223</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>67.544</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>61.876</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>54.295</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>52.813</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>47.007</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>43.971</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.881</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>33.401</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>68.831</td>
<td>30.513</td>
</tr>
</tbody>
</table>

Table 2: Result for test case ID: 02

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by J.K Rowling have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 03

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: JK Rowling – HP0.txt

Test books:
- JK Rowling – HP0.txt
- JK Rowling – HP1.txt
- Arthur Conan Doyle -The Adventures of Sherlock Holmes.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>96.949</td>
<td>59.282</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>55.509</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>51.869</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>44.239</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>42.887</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>37.012</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>33.996</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>25.827</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>23.472</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>68.831</td>
<td>21.273</td>
</tr>
</tbody>
</table>

Table 3: Result for test case ID: 03

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by J.K Rowling have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 04

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: JK Rowling – HP0.txt

Test books:
  - JK Rowling – HP0.txt
  - JK Rowling – HP1.txt
  - Edgar Rice Burroughs - A Princess of Mars.txt

Test Output:

Testing Doc01: 1 JK Rowling – HP0.txt
Testing Doc02: 2 JK Rowling – HP1.txt
Testing Doc03: 3 Edgar Rice Burroughs - A Princess of Mars.txt

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>96.949</td>
<td>74.361</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.916</td>
<td>71.467</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>68.891</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>64.412</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>63.782</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>57.561</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>56.781</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>45.771</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>42.631</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>68.831</td>
<td>41.622</td>
</tr>
</tbody>
</table>

Table 4: Result for test case ID: 04

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by J.K Rowling have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 05

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Paulo Coelho – The Alchemist.txt

Test books:
- Paulo Coelho – The Alchemist.txt
- Paulo Coelho – The Zahir.txt
- James Joyce – Dubliners.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>89.933</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>77.540</td>
<td>51.822</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>75.371</td>
<td>49.666</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>73.731</td>
<td>44.595</td>
</tr>
</tbody>
</table>

Table 5: Result for test case ID: 05

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Paulo Coelho have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 06

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Friedrich Nietzsche - Beyond Good and Evil.txt

Test books:
- Friedrich Nietzsche - Beyond Good and Evil.txt
- Friedrich Nietzsche – The Antichrist.txt
- Dante Alighieri - The Divine Comedy.txt

Test Output:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\alpha$</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>89.933</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.822</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.927</td>
<td>68.831</td>
<td>27.595</td>
</tr>
</tbody>
</table>

Table 6: Result for test case ID: 06

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Friedrich Nietzsche have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 07

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Bram Stoker – Dracula.txt

Test books:
- Bram Stoker – Dracula.txt
- Bram Stoker – The Primrose Path.txt
- Bram Stoker – The Mystery of the Sea.txt

Test Output:

```
Testing Doc01: 1 Bram Stoker – Dracula.txt
Testing Doc02: 2 Bram Stoker – The Primrose Path.txt
Testing Doc03: 3 Bram Stoker – The Mystery of the Sea.txt

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.986</td>
<td>99.749</td>
<td>99.913</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.977</td>
<td>97.636</td>
<td>97.174</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.975</td>
<td>91.455</td>
<td>91.716</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.972</td>
<td>88.421</td>
<td>89.595</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.971</td>
<td>83.865</td>
<td>85.663</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.970</td>
<td>78.356</td>
<td>83.737</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.967</td>
<td>75.833</td>
<td>79.911</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.963</td>
<td>74.522</td>
<td>75.822</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.959</td>
<td>71.534</td>
<td>74.654</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.954</td>
<td>69.451</td>
<td>71.593</td>
</tr>
</tbody>
</table>
```

Table 7: Result for test case ID: 07

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that all the books have a high matching percentage since all of them have been written by Bram Stoker.

Pass/Fail: The test has passed.
Test case ID: 08

Start \( \alpha \): 0.1

Increment per iteration: 0.1

Maximum \( \alpha \): 1.01

Learn from book: Charles Dickens - David Copperfield.txt

Test books:
- Charles Dickens - David Copperfield.txt
- Charles Dickens - A Christmas Carol.txt
- Bram Stoker - The Mystery of the Sea.txt
- Bram Stoker - Under the Sunset.txt

Test Output:

Testing Doc01: 1 Charles Dickens - David Copperfield.txt
Testing Doc02: 2 Charles Dickens - A Christmas Carol.txt
Testing Doc03: 3 Bram Stoker - The Mystery of the Sea.txt
Testing Doc04: 4 Bram Stoker - Under the Sunset.txt

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
<th>Doc04(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.986</td>
<td>99.756</td>
<td>89.913</td>
<td>86.237</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.977</td>
<td>97.676</td>
<td>77.174</td>
<td>77.434</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.975</td>
<td>91.423</td>
<td>71.716</td>
<td>73.145</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.972</td>
<td>88.453</td>
<td>69.595</td>
<td>71.957</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.971</td>
<td>83.892</td>
<td>55.663</td>
<td>66.387</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.970</td>
<td>77.379</td>
<td>43.737</td>
<td>58.712</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.967</td>
<td>73.819</td>
<td>39.911</td>
<td>47.998</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.963</td>
<td>70.592</td>
<td>35.822</td>
<td>45.393</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.959</td>
<td>67.567</td>
<td>34.654</td>
<td>41.726</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.954</td>
<td>63.493</td>
<td>31.593</td>
<td>40.571</td>
</tr>
</tbody>
</table>

Table 8: Result for test case ID: 08

Expected Result: There should be a high percentage match for the books written by the same author when \( \alpha = 1.00 \).

Actual Result: The output indicates that the books written by Charles Dickens have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 09

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Friedrich Nietzsche - Beyond Good and Evil.txt

Test books:
- Friedrich Nietzsche - Beyond Good and Evil.txt
- Friedrich Nietzsche – The Antichrist.txt
- Dante Alighieri - The Divine Comedy.txt
- James Matthew Barrie - Peter Pan.txt
- Arthur Conan Doyle - The Adventures of Sherlock Holmes.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
<th>Doc04(%)</th>
<th>Doc05(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>09.933</td>
<td>69.917</td>
<td>72.612</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>67.154</td>
<td>67.482</td>
<td>66.123</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.976</td>
<td>91.365</td>
<td>01.706</td>
<td>63.123</td>
<td>55.456</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
<td>61.981</td>
<td>48.989</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>56.363</td>
<td>39.933</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
<td>48.701</td>
<td>32.393</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>39.924</td>
<td>29.607</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.822</td>
<td>35.390</td>
<td>21.402</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
<td>34.799</td>
<td>18.198</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.927</td>
<td>68.831</td>
<td>27.595</td>
<td>31.522</td>
<td>13.327</td>
</tr>
</tbody>
</table>

Table 9: Result for test case ID: 09

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Friedrich Nietzsche have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 10

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: JK Rowling – HP0.txt

Test books:
- JK Rowling – HP0.txt
- JK Rowling – HP1.txt
- Dante Alighieri - The Divine Comedy.txt
- James Matthew Barrie - Peter Pan.txt
- Arthur Conan Doyle -The Adventures of Sherlock Holmes.txt

Test Output:

Testing Doc01: 1 JK Rowling – HP0.txt
Testing Doc02: 2 JK Rowling – HP1.txt
Testing Doc03: 3 Dante Alighieri - The Divine Comedy.txt
Testing Doc04: 4 James Matthew Barrie - Peter Pan.txt
Testing Doc05: 5 Arthur Conan Doyle -The Adventures of Sherlock Holmes.txt

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
<th>Doc04(%)</th>
<th>Doc05(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>89.933</td>
<td>69.917</td>
<td>72.612</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
<td>67.482</td>
<td>66.123</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
<td>63.123</td>
<td>55.456</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
<td>61.981</td>
<td>48.989</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>56.363</td>
<td>39.933</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.360</td>
<td>63.707</td>
<td>48.701</td>
<td>32.393</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>39.924</td>
<td>29.807</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.822</td>
<td>35.390</td>
<td>21.402</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
<td>24.799</td>
<td>21.198</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.927</td>
<td>70.031</td>
<td>37.595</td>
<td>31.522</td>
<td>13.327</td>
</tr>
</tbody>
</table>

Table 10: Result for test case ID: 10

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by J.K Rowling have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 11

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Bram Stoker – Dracula.txt

Test books:
- Bram Stoker – Dracula.txt
- Bram Stoker – The Mystery of the Sea.txt
- Arthur Conan Doyle – The Adventures of Sherlock Holmes.txt
- Edgar Rice Burroughs – A Princess of Mars.txt
- Elliott Whithey – The Pirate Shark.txt
- Frank Baum – The Wonderful Wizard of Oz.txt
- Friedrich Nietzsche – Beyond Good and Evil.txt
- Harrison Williams – Legends of Loudoun.txt

Test Output:

Testing Doc01: 1 Bram Stoker – Dracula.txt
Testing Doc02: 2 Bram Stoker – The Mystery of the Sea.txt
Testing Doc04: 4 Edgar Rice Burroughs – A Princess of Mars.txt
Testing Doc05: 5 Elliott Whithey – The Pirate Shark.txt
Testing Doc06: 6 Frank Baum – The Wonderful Wizard of Oz.txt
Testing Doc07: 7 Friedrich Nietzsche – Beyond Good and Evil.txt
Testing Doc08: 8 Harrison Williams – Legends of Loudoun.txt

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01 (%)</th>
<th>Doc02 (%)</th>
<th>Doc03 (%)</th>
<th>Doc04 (%)</th>
<th>Doc05 (%)</th>
<th>Doc06 (%)</th>
<th>Doc07 (%)</th>
<th>Doc08 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>89.933</td>
<td>99.917</td>
<td>99.612</td>
<td>99.280</td>
<td>99.198</td>
<td>99.280</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
<td>97.482</td>
<td>96.123</td>
<td>95.579</td>
<td>96.392</td>
<td>97.443</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
<td>93.123</td>
<td>95.456</td>
<td>89.254</td>
<td>89.561</td>
<td>92.914</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>99.721</td>
<td>74.585</td>
<td>91.981</td>
<td>88.989</td>
<td>83.160</td>
<td>84.982</td>
<td>89.106</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>86.363</td>
<td>73.933</td>
<td>79.284</td>
<td>71.369</td>
<td>82.338</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
<td>78.701</td>
<td>56.393</td>
<td>76.652</td>
<td>57.356</td>
<td>75.329</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>69.924</td>
<td>49.807</td>
<td>64.980</td>
<td>49.983</td>
<td>69.847</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>74.540</td>
<td>35.822</td>
<td>55.390</td>
<td>37.402</td>
<td>58.189</td>
<td>37.561</td>
<td>54.532</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>72.571</td>
<td>29.666</td>
<td>44.799</td>
<td>28.198</td>
<td>43.687</td>
<td>29.284</td>
<td>41.186</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>68.831</td>
<td>27.595</td>
<td>41.522</td>
<td>21.327</td>
<td>35.932</td>
<td>22.134</td>
<td>33.786</td>
</tr>
</tbody>
</table>

Table 11: Result for test case ID: 11
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Bram Stoker have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 12

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Bram Stoker – Dracula.txt

Test books:
- Bram Stoker – Dracula.txt
- Bram Stoker – The Primrose Path.txt
- Bram Stoker – The Mystery of the Sea.txt
- Bram Stoker – Under the Sunset.txt
- Bram Stoker – Miss Betty.txt
- Frank Baum - The Wonderful Wizard of Oz.txt
- Friedrich Nietzsche - Beyond Good and Evil.txt
- Harrison Williams - Legends of Loudoun.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01 (%)</th>
<th>Doc02 (%)</th>
<th>Doc03 (%)</th>
<th>Doc04 (%)</th>
<th>Doc05 (%)</th>
<th>Doc06 (%)</th>
<th>Doc07 (%)</th>
<th>Doc08 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.977</td>
<td>97.636</td>
<td>97.174</td>
<td>97.402</td>
<td>96.723</td>
<td>95.579</td>
<td>96.392</td>
<td>97.443</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.975</td>
<td>91.455</td>
<td>91.716</td>
<td>93.123</td>
<td>95.356</td>
<td>69.254</td>
<td>99.561</td>
<td>92.914</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.972</td>
<td>88.421</td>
<td>89.595</td>
<td>91.981</td>
<td>88.389</td>
<td>83.160</td>
<td>84.982</td>
<td>89.106</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.971</td>
<td>83.865</td>
<td>85.663</td>
<td>86.363</td>
<td>83.932</td>
<td>79.284</td>
<td>71.369</td>
<td>82.338</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.970</td>
<td>78.356</td>
<td>83.737</td>
<td>78.701</td>
<td>79.391</td>
<td>76.652</td>
<td>57.356</td>
<td>75.329</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.967</td>
<td>75.833</td>
<td>79.511</td>
<td>77.924</td>
<td>76.808</td>
<td>64.980</td>
<td>49.903</td>
<td>69.947</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.963</td>
<td>74.522</td>
<td>75.622</td>
<td>72.390</td>
<td>72.406</td>
<td>58.189</td>
<td>37.561</td>
<td>54.532</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.959</td>
<td>71.534</td>
<td>74.654</td>
<td>71.799</td>
<td>68.194</td>
<td>43.687</td>
<td>29.284</td>
<td>41.186</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.954</td>
<td>69.451</td>
<td>71.593</td>
<td>70.522</td>
<td>66.322</td>
<td>35.932</td>
<td>22.134</td>
<td>33.786</td>
</tr>
</tbody>
</table>

Table 12: Result for test case ID: 12
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Bram Stoker have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 13

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: JK Rowling - HP0.txt

Test books:
- JK Rowling - HP0.txt
- JK Rowling – HP1.txt
- JK Rowling – HP2.txt
- JK Rowling – HP3.txt
- JK Rowling – HP4.txt
- JK Rowling – HP5.txt
- JK Rowling – HP6.txt

Test Output:

<table>
<thead>
<tr>
<th>Testing Doc01: 1 JK Rowling - HP0.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing Doc02: 2 JK Rowling - HP1.txt</td>
</tr>
<tr>
<td>Testing Doc03: 3 JK Rowling - HP2.txt</td>
</tr>
<tr>
<td>Testing Doc04: 4 JK Rowling - HP3.txt</td>
</tr>
<tr>
<td>Testing Doc05: 5 JK Rowling - HP4.txt</td>
</tr>
<tr>
<td>Testing Doc06: 6 JK Rowling - HP5.txt</td>
</tr>
<tr>
<td>Testing Doc07: 7 JK Rowling - HP6.txt</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
<th>Doc04(%)</th>
<th>Doc05(%)</th>
<th>Doc06(%)</th>
<th>Doc07(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>89.933</td>
<td>99.917</td>
<td>99.612</td>
<td>99.280</td>
<td>99.198</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
<td>97.482</td>
<td>96.123</td>
<td>95.579</td>
<td>96.392</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
<td>93.123</td>
<td>95.456</td>
<td>99.254</td>
<td>89.561</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
<td>91.981</td>
<td>88.989</td>
<td>83.160</td>
<td>89.982</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>86.363</td>
<td>83.933</td>
<td>82.284</td>
<td>88.369</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
<td>78.701</td>
<td>81.393</td>
<td>81.652</td>
<td>87.356</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>69.924</td>
<td>79.807</td>
<td>79.980</td>
<td>79.983</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.822</td>
<td>55.390</td>
<td>71.402</td>
<td>78.189</td>
<td>71.561</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
<td>44.799</td>
<td>68.198</td>
<td>74.687</td>
<td>69.284</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.927</td>
<td>68.831</td>
<td>77.595</td>
<td>71.522</td>
<td>63.327</td>
<td>73.932</td>
<td>68.134</td>
</tr>
</tbody>
</table>

Table 13: Result for test case ID: 13
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by J.K Rowling have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 14

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Charles Dickens - David Copperfield.txt

Test books:
- Charles Dickens - David Copperfield.txt
- Charles Dickens - A Christmas Carol.txt
- Bram Stoker - The Mystery of the Sea.txt
- Bram Stoker - Under the Sunset.txt
- Bram Stoker - Miss Betty.txt

Test Output:

Testing Doc01: 1 Charles Dickens - David Copperfield.txt
Testing Doc02: 2 Charles Dickens - A Christmas Carol.txt
Testing Doc03: 3 Bram Stoker - The Mystery of the Sea.txt
Testing Doc04: 4 Bram Stoker - Under the Sunset.txt
Testing Doc05: 5 Bram Stoker - Miss Betty.txt

<table>
<thead>
<tr>
<th>i</th>
<th>$\alpha$</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
<th>Doc04(%)</th>
<th>Doc05(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.986</td>
<td>99.756</td>
<td>89.913</td>
<td>86.237</td>
<td>79.112</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.977</td>
<td>97.676</td>
<td>77.174</td>
<td>77.434</td>
<td>76.723</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.975</td>
<td>91.423</td>
<td>71.716</td>
<td>73.145</td>
<td>75.356</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.972</td>
<td>88.453</td>
<td>69.595</td>
<td>71.957</td>
<td>68.239</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.971</td>
<td>83.892</td>
<td>55.663</td>
<td>66.387</td>
<td>63.932</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.970</td>
<td>77.379</td>
<td>43.737</td>
<td>58.712</td>
<td>59.541</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.967</td>
<td>73.819</td>
<td>39.911</td>
<td>47.998</td>
<td>46.758</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.963</td>
<td>70.592</td>
<td>35.822</td>
<td>45.393</td>
<td>42.726</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.959</td>
<td>67.567</td>
<td>34.654</td>
<td>41.726</td>
<td>38.834</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.954</td>
<td>63.493</td>
<td>31.593</td>
<td>40.571</td>
<td>26.692</td>
</tr>
</tbody>
</table>

Table 14: Result for test case ID: 14
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Charles Dickens have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 15
Start $\alpha$: 0.1
Increment per iteration: 0.1
Maximum $\alpha$: 1.01
Learn from book: JK Rowling - HP0.txt
Test books:
- JK Rowling - HP0.txt
- JK Rowling - HP5.txt
- Arthur Conan Doyle - The Adventures of Sherlock Holmes.txt
- Edgar Rice Burroughs - A Princess of Mars.txt
- Elliott Whithey - The Pirate Shark.txt
- Frank Baum - The Wonderful Wizard of Oz.txt
- Friedrich Nietzsche - Beyond Good and Evil.txt
- Harrison Williams - Legends of Loudoun.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01 (%)</th>
<th>Doc02 (%)</th>
<th>Doc03 (%)</th>
<th>Doc04 (%)</th>
<th>Doc05 (%)</th>
<th>Doc06 (%)</th>
<th>Doc07 (%)</th>
<th>Doc08 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
<td>97.482</td>
<td>96.123</td>
<td>95.579</td>
<td>96.392</td>
<td>97.443</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
<td>93.123</td>
<td>95.456</td>
<td>89.254</td>
<td>89.561</td>
<td>92.914</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
<td>91.981</td>
<td>88.989</td>
<td>83.160</td>
<td>84.982</td>
<td>89.106</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>86.363</td>
<td>73.933</td>
<td>79.284</td>
<td>71.369</td>
<td>82.338</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
<td>78.701</td>
<td>56.393</td>
<td>76.652</td>
<td>57.356</td>
<td>75.329</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>69.924</td>
<td>49.807</td>
<td>64.980</td>
<td>49.963</td>
<td>69.847</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.822</td>
<td>55.390</td>
<td>37.402</td>
<td>58.189</td>
<td>37.561</td>
<td>44.532</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
<td>44.799</td>
<td>28.198</td>
<td>43.687</td>
<td>29.284</td>
<td>38.186</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.927</td>
<td>68.831</td>
<td>27.595</td>
<td>31.522</td>
<td>23.327</td>
<td>33.932</td>
<td>28.134</td>
<td>32.786</td>
</tr>
</tbody>
</table>

Table 15: Result for test case ID: 15
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by J.K Rowling have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 16

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Paulo Coelho - The Alchemist.txt

Test books:
- Friedrich Nietzsche - Beyond Good and Evil.txt
- Charlotte Bronte - Jane Eyre.txt
- Dante Alighieri - The Divine Comedy.txt
- James Matthew Barrie - Peter Pan.txt
- Arthur Conan Doyle - The Adventures of Sherlock Holmes.txt
- Edgar Rice Burroughs - A Princess of Mars.txt
- Elliott Whithey - The Pirate Shark.txt
- Frank Baum - The Wonderful Wizard of Oz.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
<th>Doc04(%)</th>
<th>Doc05(%)</th>
<th>Doc06(%)</th>
<th>Doc07(%)</th>
<th>Doc08(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>97.46</td>
<td>96.123</td>
<td>95.579</td>
<td>96.392</td>
<td>97.482</td>
<td>66.123</td>
<td>75.567</td>
<td>96.872</td>
<td>97.443</td>
</tr>
<tr>
<td>93</td>
<td>93.123</td>
<td>95.456</td>
<td>99.254</td>
<td>89.561</td>
<td>92.123</td>
<td>65.236</td>
<td>69.225</td>
<td>89.091</td>
<td>92.914</td>
</tr>
<tr>
<td>91</td>
<td>91.981</td>
<td>88.989</td>
<td>83.160</td>
<td>84.982</td>
<td>91.981</td>
<td>58.529</td>
<td>63.164</td>
<td>84.432</td>
<td>89.106</td>
</tr>
<tr>
<td>86</td>
<td>86.303</td>
<td>73.933</td>
<td>79.284</td>
<td>71.369</td>
<td>86.363</td>
<td>53.163</td>
<td>59.264</td>
<td>71.769</td>
<td>82.338</td>
</tr>
<tr>
<td>78</td>
<td>78.701</td>
<td>56.393</td>
<td>56.652</td>
<td>57.356</td>
<td>78.701</td>
<td>46.783</td>
<td>56.675</td>
<td>57.906</td>
<td>75.329</td>
</tr>
<tr>
<td>49</td>
<td>49.807</td>
<td>64.900</td>
<td>49.983</td>
<td>49.807</td>
<td>64.900</td>
<td>39.223</td>
<td>51.375</td>
<td>37.221</td>
<td>69.847</td>
</tr>
<tr>
<td>55</td>
<td>55.390</td>
<td>37.402</td>
<td>58.189</td>
<td>37.561</td>
<td>58.189</td>
<td>37.781</td>
<td>48.137</td>
<td>37.541</td>
<td>44.532</td>
</tr>
<tr>
<td>44</td>
<td>44.799</td>
<td>28.198</td>
<td>43.687</td>
<td>29.284</td>
<td>55.390</td>
<td>32.342</td>
<td>42.191</td>
<td>37.441</td>
<td>38.186</td>
</tr>
</tbody>
</table>

Table 16: Result for test case ID: 16
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: Since there is no book in the testing list written by Paulo Coelho, we observe that the pattern match for the other books is low.

Pass/Fail: The test has passed.
Test case ID: 17

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Paulo Coelho - The Alchemist.txt

Test books:
- Paulo Coelho - The Alchemist.txt
- Paulo Coelho - Eleven Minutes.txt
- Paulo Coelho - The Zahir.txt
- Paulo Coelho - The Fifth mountain.txt
- Paulo Coelho - The Winner Stands Alone.txt
- Paulo Coelho - Aleph.txt

Test Output:

<table>
<thead>
<tr>
<th>Doc 1 (%)</th>
<th>Doc 2 (%)</th>
<th>Doc 3 (%)</th>
<th>Doc 4 (%)</th>
<th>Doc 5 (%)</th>
<th>Doc 6 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>89.933</td>
<td>99.917</td>
<td>99.612</td>
</tr>
<tr>
<td>2 0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
<td>97.482</td>
<td>96.123</td>
</tr>
<tr>
<td>3 0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
<td>93.123</td>
<td>95.456</td>
</tr>
<tr>
<td>4 0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
<td>91.981</td>
<td>88.989</td>
</tr>
<tr>
<td>5 0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>86.363</td>
<td>73.933</td>
</tr>
<tr>
<td>6 0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
<td>78.701</td>
<td>56.393</td>
</tr>
<tr>
<td>7 0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>69.924</td>
<td>49.807</td>
</tr>
<tr>
<td>8 0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>35.822</td>
<td>55.390</td>
<td>37.402</td>
</tr>
<tr>
<td>9 0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
<td>44.799</td>
<td>28.198</td>
</tr>
<tr>
<td>10 1.00</td>
<td>99.927</td>
<td>68.031</td>
<td>61.595</td>
<td>71.522</td>
<td>64.327</td>
</tr>
</tbody>
</table>

Table 17: Result for test case ID: 17
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Paulo Coelho have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 18

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Paulo Coelho - The Alchemist.txt

Test books:
- Paulo Coelho - The Alchemist.txt
- Paulo Coelho - Eleven Minutes.txt
- Paulo Coelho - The Zahir.txt
- Paulo Coelho - The Fifth mountain.txt
- Paulo Coelho - The Winner Stands Alone.txt
- Paulo Coelho - Aleph.txt
- Karl Marx - Das Kapital.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01 (%)</th>
<th>Doc02 (%)</th>
<th>Doc03 (%)</th>
<th>Doc04 (%)</th>
<th>Doc05 (%)</th>
<th>Doc06 (%)</th>
<th>Doc07 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>89.933</td>
<td>99.917</td>
<td>99.612</td>
<td>99.280</td>
<td>79.198</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
<td>97.482</td>
<td>96.123</td>
<td>95.579</td>
<td>76.392</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
<td>93.123</td>
<td>95.456</td>
<td>89.254</td>
<td>69.561</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
<td>91.981</td>
<td>88.989</td>
<td>83.160</td>
<td>64.982</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>86.363</td>
<td>73.933</td>
<td>79.284</td>
<td>61.369</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
<td>78.701</td>
<td>56.393</td>
<td>76.652</td>
<td>57.356</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>69.924</td>
<td>49.807</td>
<td>64.980</td>
<td>49.983</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>55.822</td>
<td>55.390</td>
<td>37.402</td>
<td>58.189</td>
<td>27.561</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
<td>44.799</td>
<td>28.198</td>
<td>43.687</td>
<td>24.284</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.927</td>
<td>68.831</td>
<td>61.595</td>
<td>71.522</td>
<td>64.327</td>
<td>73.932</td>
<td>23.134</td>
</tr>
</tbody>
</table>

Table 18: Result for test case ID: 18
Expected Result: There should be a high percentage match for the books written by the same author when \( \alpha = 1.00 \).

Actual Result: The output indicates that the books written by Paulo Coelho have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 19

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Paulo Coelho - The Alchemist.txt

Test books:
- Paulo Coelho - The Alchemist.txt
- Paulo Coelho - Eleven Minutes.txt
- Paulo Coelho - The Zahir.txt
- Paulo Coelho - The Fifth Mountain.txt
- Paulo Coelho - The Winner Stands Alone.txt
- Paulo Coelho - Aleph.txt
- Karl Marx - Das Kapital.txt
- Harrison Williams - Legends of Loudoun.txt
- Friedrich Nietzsche - Beyond Good and Evil.txt

Test Output:

Table 19: Result for test case ID: 19

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(%)</th>
<th>Doc02(%)</th>
<th>Doc03(%)</th>
<th>Doc04(%)</th>
<th>Doc05(%)</th>
<th>Doc06(%)</th>
<th>Doc07(%)</th>
<th>Doc08(%)</th>
<th>Doc09(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>99.949</td>
<td>99.933</td>
<td>99.917</td>
<td>99.612</td>
<td>99.280</td>
<td>79.199</td>
<td>64.917</td>
<td>69.612</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>97.816</td>
<td>87.154</td>
<td>97.482</td>
<td>96.123</td>
<td>95.579</td>
<td>76.392</td>
<td>57.482</td>
<td>66.123</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>91.365</td>
<td>81.706</td>
<td>93.123</td>
<td>95.456</td>
<td>89.254</td>
<td>69.561</td>
<td>53.123</td>
<td>65.456</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>88.721</td>
<td>74.585</td>
<td>91.981</td>
<td>88.989</td>
<td>83.160</td>
<td>64.982</td>
<td>51.981</td>
<td>58.989</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>82.808</td>
<td>70.633</td>
<td>86.363</td>
<td>73.933</td>
<td>79.284</td>
<td>61.369</td>
<td>46.363</td>
<td>53.933</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>79.368</td>
<td>63.707</td>
<td>78.701</td>
<td>56.393</td>
<td>76.652</td>
<td>57.356</td>
<td>43.701</td>
<td>46.393</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>77.896</td>
<td>52.961</td>
<td>69.924</td>
<td>49.807</td>
<td>64.980</td>
<td>49.983</td>
<td>39.807</td>
<td>44.980</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>71.540</td>
<td>55.822</td>
<td>55.390</td>
<td>37.402</td>
<td>58.189</td>
<td>27.561</td>
<td>35.390</td>
<td>37.402</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>69.571</td>
<td>29.666</td>
<td>44.799</td>
<td>28.198</td>
<td>43.687</td>
<td>24.284</td>
<td>34.799</td>
<td>28.198</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.927</td>
<td>68.831</td>
<td>61.595</td>
<td>71.522</td>
<td>64.327</td>
<td>73.932</td>
<td>23.134</td>
<td>31.537</td>
<td>23.329</td>
</tr>
</tbody>
</table>
Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Paulo Coelho have a higher match as compared with other authors.

Pass/Fail: The test has passed.
Test case ID: 20

Start $\alpha$: 0.1

Increment per iteration: 0.1

Maximum $\alpha$: 1.01

Learn from book: Arthur Conan Doyle - The Adventures of Sherlock Holmes.txt

Test books:
- Arthur Conan Doyle - The Adventures of Sherlock Holmes.txt
- Arthur Conan Doyle - The Lost World.txt
- Leo Tolstoy - War and Peace.txt
- Edgar Rice Burroughs - A Princess of Mars.txt
- Elliott Whithey - The Pirate Shark.txt

Test Output:

<table>
<thead>
<tr>
<th>i</th>
<th>Alpha</th>
<th>Doc01(#)</th>
<th>Doc02(#)</th>
<th>Doc03(#)</th>
<th>Doc04(#)</th>
<th>Doc05(#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.986</td>
<td>99.756</td>
<td>89.913</td>
<td>86.237</td>
<td>79.112</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.977</td>
<td>97.676</td>
<td>77.174</td>
<td>77.434</td>
<td>76.723</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.975</td>
<td>91.423</td>
<td>71.716</td>
<td>73.145</td>
<td>75.356</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.972</td>
<td>86.453</td>
<td>69.595</td>
<td>71.957</td>
<td>68.239</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.971</td>
<td>83.892</td>
<td>55.663</td>
<td>66.387</td>
<td>63.932</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.970</td>
<td>77.379</td>
<td>43.737</td>
<td>50.712</td>
<td>59.541</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.967</td>
<td>73.819</td>
<td>39.911</td>
<td>47.998</td>
<td>46.758</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.963</td>
<td>70.592</td>
<td>35.822</td>
<td>45.393</td>
<td>42.726</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.959</td>
<td>67.567</td>
<td>34.654</td>
<td>41.726</td>
<td>38.834</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.954</td>
<td>63.493</td>
<td>31.593</td>
<td>40.571</td>
<td>26.692</td>
</tr>
</tbody>
</table>

Table 20: Result for test case ID: 20

Expected Result: There should be a high percentage match for the books written by the same author when $\alpha = 1.00$.

Actual Result: The output indicates that the books written by Sir Arthur Conan Doyle have a higher match as compared with other authors.

Pass/Fail: The test has passed.
6. Future Work

The Alergia algorithm is one of the state-merging algorithms like Regular Positive and Negative Inference (RPNI) and Minimum Divergence Inference (MDI), but from the probabilistic view. In practice, we are dealing with frequency of samples most of time, but it is very trivial to convert a Deterministic Frequency Finite Automata (DFFA) to Deterministic Probabilistic Finite Automata (DPFA). Alergia is such a learning algorithm which is able to learn a DFFA and its corresponding DPFA from a sample containing duplicate strings.

However, Minimum Divergence Inference (MDI) is another version of learning probabilistic definite finite automata (PDFA). The goal is to find balance between the gain in size and the loss in perplexity. So the only difference with Alergia is that the merge has now happened inside compatibility test and the score function is using perplexity. This algorithm should be tested to check if we get better results as compared to Alergia.

The performance of the program in terms of time complexity can be improved in the future by performing parallel processing. The shared memory architecture can be used to perform comparison between the book which the program uses to learn and generate automata with other books from various authors.
7. Conclusion

We proposed a method for pattern discovery for symbolic data using automata [5] and Alergia algorithm. The PTA is created based on the function words [2][6] and the compatible states are merged which further help us in discovering the pattern similarity. This method is used to analyze similar writing styles of various authors thus helping us identify them. Dr. Lin [3][4][7][8] has been researching this topic since 2005 with his former students S. Zhang [14], Y. Lu [15], Q. Yu [16] and A. Yazdhankhah [17] for their Master’s Thesis at San Jose State University. We have continued to research and make progress on this subject and the results seem to be promising for future applications.

The proposed system can also be used in biology to study Microarray as well as in Bioinformatics to differentiate between existing species.
8. References

APPENDIX A: Development Environment

The Table below contains the hardware and software specifications used for the development of the program.

<table>
<thead>
<tr>
<th>Software Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
</tr>
<tr>
<td>Integrated Development Environment</td>
</tr>
<tr>
<td>Operating System</td>
</tr>
</tbody>
</table>

Table 21: Software Specifications

<table>
<thead>
<tr>
<th>Hardware Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>RAM</td>
</tr>
<tr>
<td>CPU</td>
</tr>
<tr>
<td>Speed</td>
</tr>
</tbody>
</table>

Table 22: Hardware Specifications
# APPENDIX B: List of EBook’s used

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Book Name</th>
<th>Author Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Harry Potter and the Sorcerer’s Stone</td>
<td>J.K Rowling</td>
</tr>
<tr>
<td>2</td>
<td>Harry Potter and the Chamber of Secrets</td>
<td>J.K Rowling</td>
</tr>
<tr>
<td>3</td>
<td>Harry Potter and the Prisoner of Azkaban</td>
<td>J.K Rowling</td>
</tr>
<tr>
<td>4</td>
<td>Harry Potter and the Goblet of Fire</td>
<td>J.K Rowling</td>
</tr>
<tr>
<td>5</td>
<td>Harry Potter and the Order of the Phoenix</td>
<td>J.K Rowling</td>
</tr>
<tr>
<td>6</td>
<td>Harry Potter and the Half-blood Prince</td>
<td>J.K Rowling</td>
</tr>
<tr>
<td>7</td>
<td>Harry Potter and the Deathly Hallows</td>
<td>J.K Rowling</td>
</tr>
<tr>
<td>8</td>
<td>The Alchemist</td>
<td>Paulo Coelho</td>
</tr>
<tr>
<td>9</td>
<td>Eleven Minutes</td>
<td>Paulo Coelho</td>
</tr>
<tr>
<td>10</td>
<td>The Fifth Mountain</td>
<td>Paulo Coelho</td>
</tr>
<tr>
<td>11</td>
<td>The Zahir</td>
<td>Paulo Coelho</td>
</tr>
<tr>
<td>12</td>
<td>The Winner stands alone</td>
<td>Paulo Coelho</td>
</tr>
<tr>
<td>13</td>
<td>Aleph</td>
<td>Paulo Coelho</td>
</tr>
<tr>
<td>14</td>
<td>The Adventures of Sherlock Holmes</td>
<td>Sir Arthur Conan Doyle</td>
</tr>
<tr>
<td>15</td>
<td>A Study in Scarlet</td>
<td>Sir Arthur Conan Doyle</td>
</tr>
<tr>
<td>16</td>
<td>The Lost World</td>
<td>Sir Arthur Conan Doyle</td>
</tr>
<tr>
<td>17</td>
<td>His Last Bow</td>
<td>Sir Arthur Conan Doyle</td>
</tr>
<tr>
<td>18</td>
<td>The Sign of Four</td>
<td>Sir Arthur Conan Doyle</td>
</tr>
<tr>
<td>19</td>
<td>The Adventures of Tom Sawyer</td>
<td>Mark Twain</td>
</tr>
<tr>
<td>20</td>
<td>The Adventures of Huckleberry Finn</td>
<td>Mark Twain</td>
</tr>
<tr>
<td>21</td>
<td>The Prince and the Pauper</td>
<td>Mark Twain</td>
</tr>
<tr>
<td>22</td>
<td>Roughing it</td>
<td>Mark Twain</td>
</tr>
<tr>
<td>23</td>
<td>Great Expectations</td>
<td>Charles Dickens</td>
</tr>
<tr>
<td>24</td>
<td>A Christmas Carol</td>
<td>Charles Dickens</td>
</tr>
<tr>
<td>25</td>
<td>Oliver Twist</td>
<td>Charles Dickens</td>
</tr>
<tr>
<td>26</td>
<td>David Copperfield</td>
<td>Charles Dickens</td>
</tr>
<tr>
<td>27</td>
<td>Das Kapital</td>
<td>Karl Marx</td>
</tr>
<tr>
<td>28</td>
<td>Legends of Loudoun</td>
<td>Harrison Williams</td>
</tr>
<tr>
<td>29</td>
<td>War and Peace</td>
<td>Leo Tolstoy</td>
</tr>
<tr>
<td>30</td>
<td>A Princess of Mars</td>
<td>Edgar Rice Burroughs</td>
</tr>
<tr>
<td>31</td>
<td>The Pirate Shark</td>
<td>Elliott Whithey</td>
</tr>
<tr>
<td>32</td>
<td>Beyond Good and Evil</td>
<td>Friedrich Nietzsche</td>
</tr>
<tr>
<td>33</td>
<td>The Antichrist</td>
<td>Friedrich Nietzsche</td>
</tr>
<tr>
<td>34</td>
<td>Peter Pan</td>
<td>James Matthew Barrie</td>
</tr>
<tr>
<td>35</td>
<td>The Divine Comedy</td>
<td>Dante Alighieri</td>
</tr>
<tr>
<td>36</td>
<td>Dracula</td>
<td>Bram Stoker</td>
</tr>
<tr>
<td>37</td>
<td>The Primrose Path</td>
<td>Bram Stoker</td>
</tr>
<tr>
<td>#</td>
<td>Title</td>
<td>Author</td>
</tr>
<tr>
<td>----</td>
<td>------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>38</td>
<td>The Mystery of the Sea</td>
<td>Bram Stoker</td>
</tr>
<tr>
<td>39</td>
<td>Under the Sunset</td>
<td>Bram Stoker</td>
</tr>
<tr>
<td>40</td>
<td>The Wonderful Wizard of Oz</td>
<td>Frank Baum</td>
</tr>
</tbody>
</table>

Table 23: List of Ebook’s