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 the 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

LIST OF TABLES

LIST OF FIGURES

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 discover 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 Σ
- A transition function that takes a state and a symbol as arguments and returns a state. It is denoted by δ.
- The start state denoted by q_0
- Set of final or accepting states denoted by *F*

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

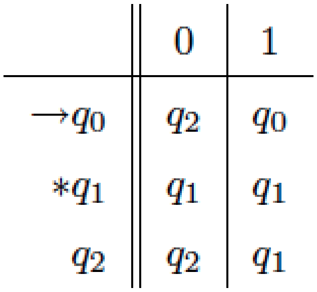
So a DFA is mathematically represented as a 5-uple (Q, Σ, δ, q_0, F).

The transition function δ is a function in

Q x Σ → Q

Q x Σ is the set of 2-tuples (q, a) with q ∈ Q and a ∈ Σ

A DFA with a transition table is given as



	0	1
$\rightarrow q_0$	q_2	q_0
$*q_1$	q_1	q_1
q_2	q_2	q_1

Figure 1: State transition table

This transition table defines the following transition diagram,

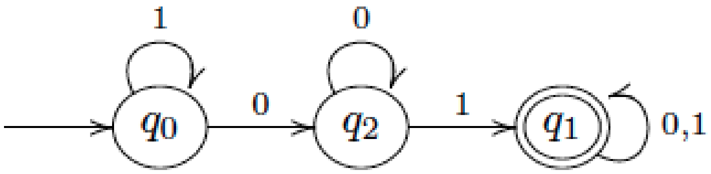


Figure 2: State Transition Diagram

Therefore,

Q = {q_0, q_1, q_2}

Start state q_0

F = {q_1}

Σ = {0, 1}

δ is a function from Q x Σ to Q

δ: Q x Σ ➔ Q

δ(q_0, 1) = q_0

δ(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 δ(q, b_1) or δ(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,

$$p_{if} + \sum_{q_j \in Q} \sum_{a \in A} p_{ij}(a) = 1$$

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 $\delta_i(a)$

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

Calculate the following probabilities:

$p_i(a) = f_i(a)/n_i$

$p_{if} = f_i(\#)/n_i$

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 α, 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:

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

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
        If not compatible (δ(i,a), δ(j,a) )
            Return false
        End if
    End do
    Return true
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
    Return $\left|\frac{f}{n} - \frac{f'}{n'}\right| < \sqrt{\frac{1}{2} log \frac{2}{\alpha}} (\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n'}})$
End Algorithm
```

```
Algorithm ALERGIA
Input:
    S : sample set of strings
    $\alpha$ : 1-confidence level
Output:
    Stochastic DFA
Begin
    A = stochastic Prefix Tree Acceptor from S
    Do (for j = successor (first node (A)) to last node (A))
        Do (for I = 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, -, 00, -, -, 100, -, -, 10110}

Let α = 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.

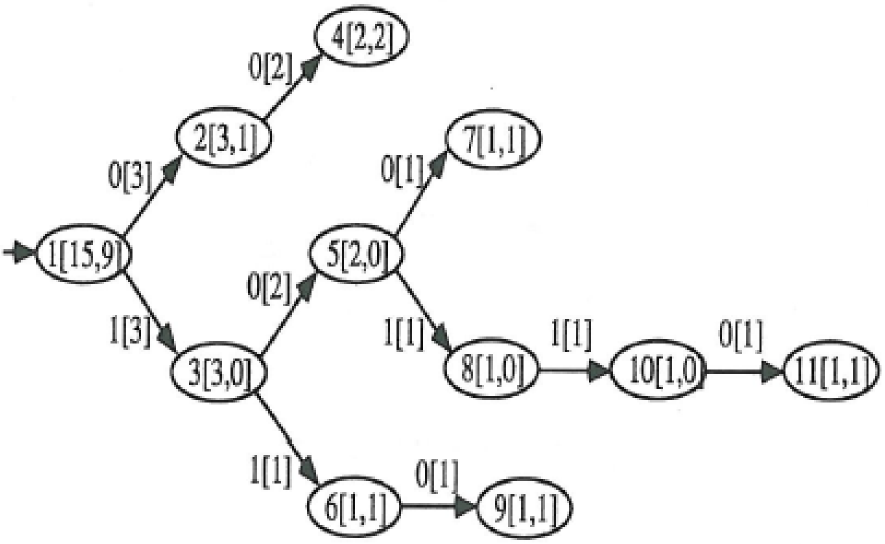


Figure 3: PTA tree for sample string *S* [9]

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 α = 0.8, we get

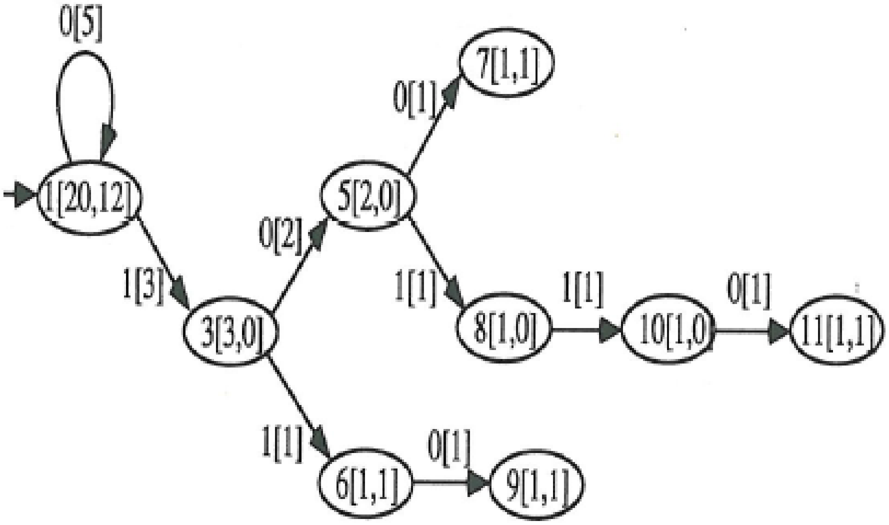


Figure 4: PTA after merging q_2 and q_1 [9]

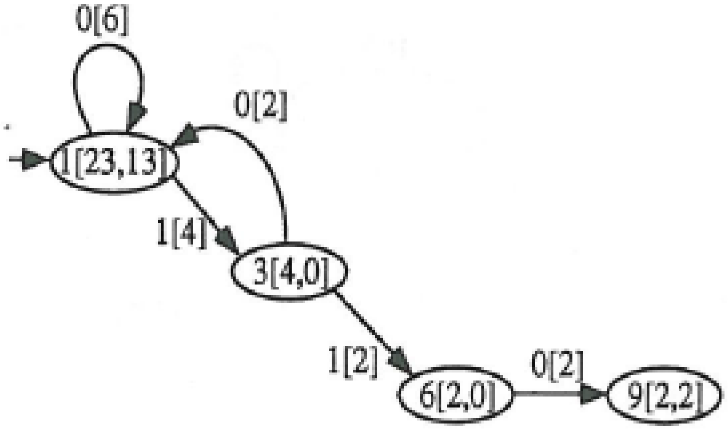


Figure 5: PTA after merging q_5 and q_1 [9]

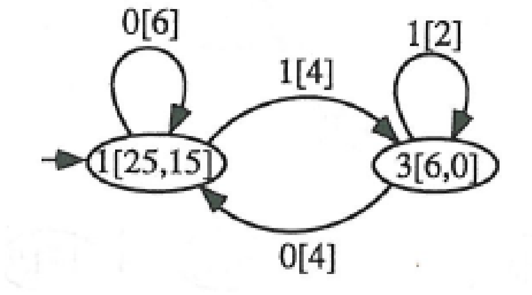


Figure 6: PTA after merging q_6 and q_3 [9]

5. Test Results

Test case ID: 01

Start α: 0.1

Increment per iteration: 0.1

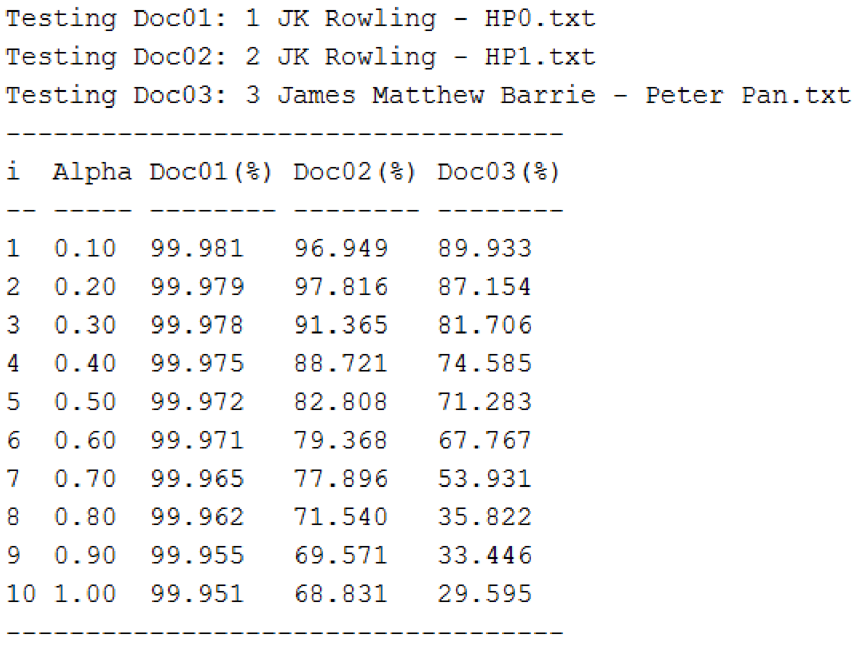
Maximum α: 1.01

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:



```
Testing Doc01: 1 JK Rowling - HP0.txt
Testing Doc02: 2 JK Rowling - HP1.txt
Testing Doc03: 3 James Matthew Barrie - Peter Pan.txt
-----------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%)
-- ----- -------- -------- --------
1  0.10  99.981   96.949   89.933
2  0.20  99.979   97.816   87.154
3  0.30  99.978   91.365   81.706
4  0.40  99.975   88.721   74.585
5  0.50  99.972   82.808   71.283
6  0.60  99.971   79.368   67.767
7  0.70  99.965   77.896   53.931
8  0.80  99.962   71.540   35.822
9  0.90  99.955   69.571   33.446
10 1.00  99.951   68.831   29.595
-----------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

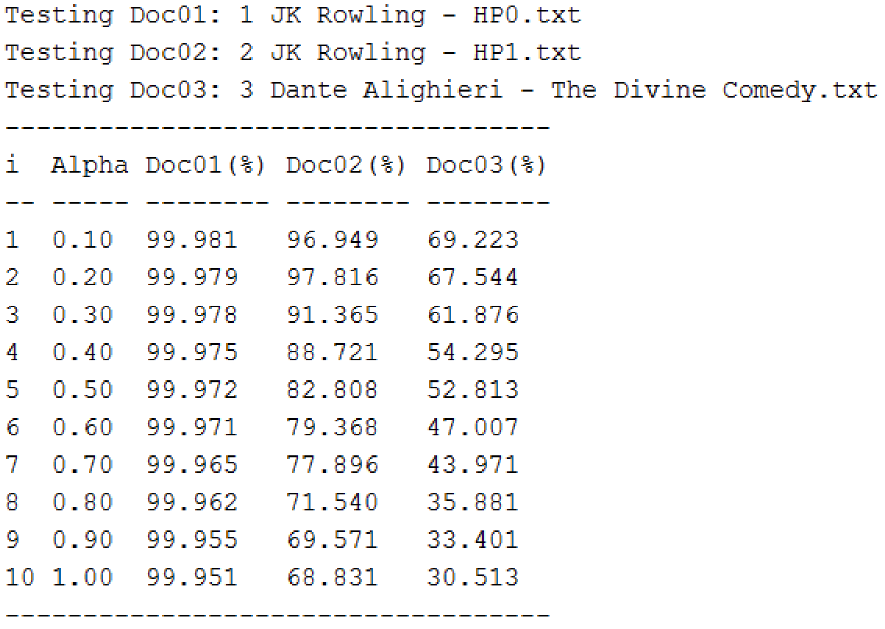
Maximum α: 1.01

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:



```
Testing Doc01: 1 JK Rowling - HP0.txt
Testing Doc02: 2 JK Rowling - HP1.txt
Testing Doc03: 3 Dante Alighieri - The Divine Comedy.txt
-----------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%)
-- ----- -------- -------- --------
1  0.10  99.981   96.949   69.223
2  0.20  99.979   97.816   67.544
3  0.30  99.978   91.365   61.876
4  0.40  99.975   88.721   54.295
5  0.50  99.972   82.808   52.813
6  0.60  99.971   79.368   47.007
7  0.70  99.965   77.896   43.971
8  0.80  99.962   71.540   35.881
9  0.90  99.955   69.571   33.401
10 1.00  99.951   68.831   30.513
-----------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

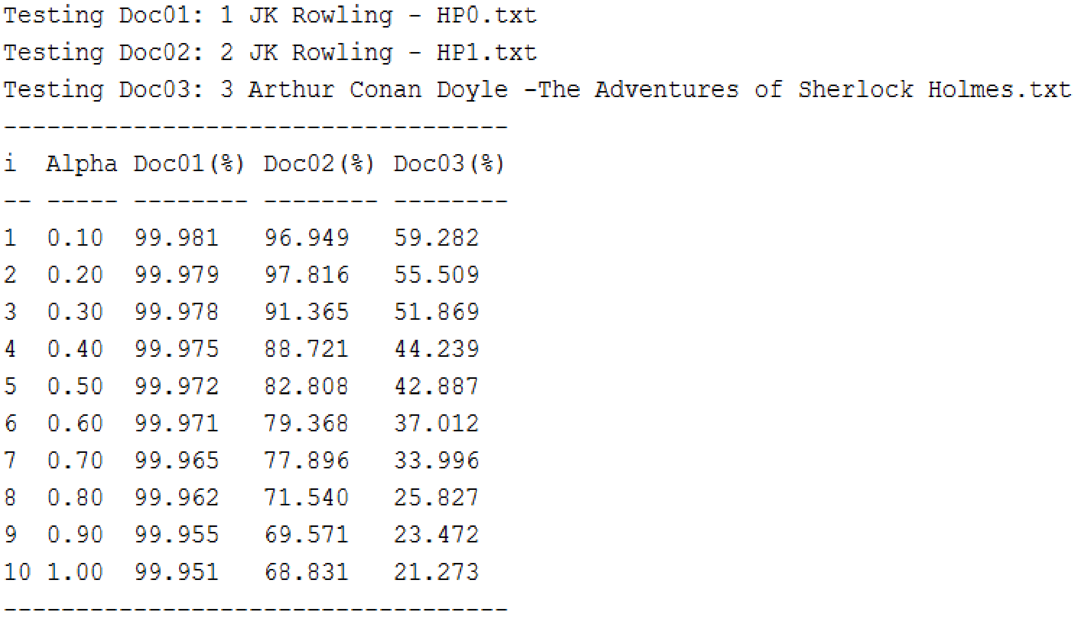
Maximum α: 1.01

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:



```
Testing Doc01: 1 JK Rowling - HP0.txt
Testing Doc02: 2 JK Rowling - HP1.txt
Testing Doc03: 3 Arthur Conan Doyle -The Adventures of Sherlock Holmes.txt
-----------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%)
-- ----- -------- -------- --------
1  0.10  99.981   96.949   59.282
2  0.20  99.979   97.816   55.509
3  0.30  99.978   91.365   51.869
4  0.40  99.975   88.721   44.239
5  0.50  99.972   82.808   42.887
6  0.60  99.971   79.368   37.012
7  0.70  99.965   77.896   33.996
8  0.80  99.962   71.540   25.827
9  0.90  99.955   69.571   23.472
10 1.00  99.951   68.831   21.273
-----------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

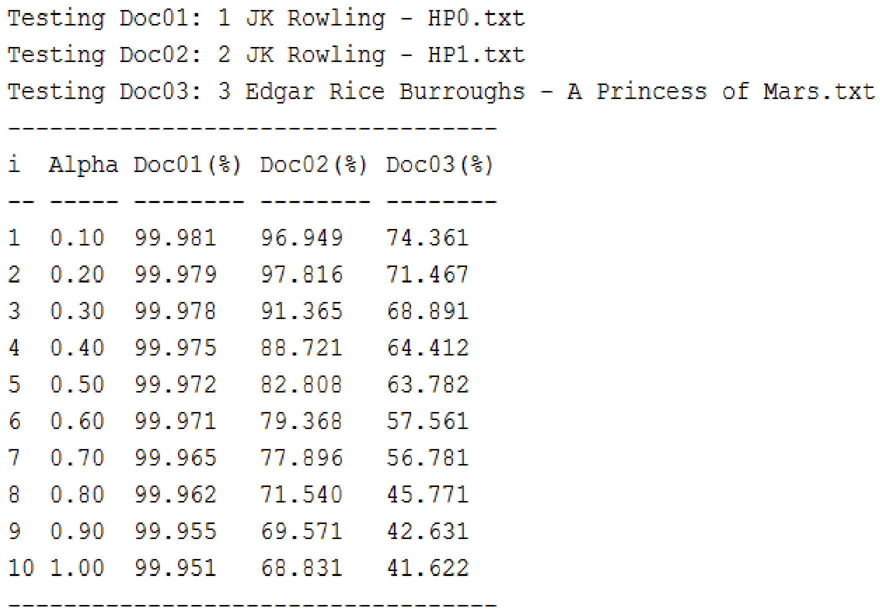
Maximum α: 1.01

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
-----------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%)
-- ----- -------- -------- --------
1  0.10  99.981   96.949   74.361
2  0.20  99.979   97.816   71.467
3  0.30  99.978   91.365   68.891
4  0.40  99.975   88.721   64.412
5  0.50  99.972   82.808   63.782
6  0.60  99.971   79.368   57.561
7  0.70  99.965   77.896   56.781
8  0.80  99.962   71.540   45.771
9  0.90  99.955   69.571   42.631
10 1.00  99.951   68.831   41.622
-----------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

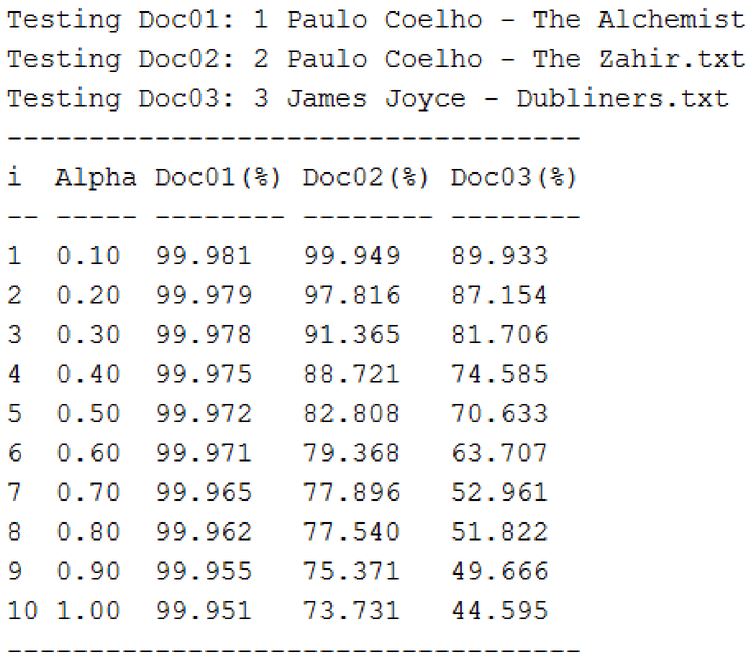
Maximum α: 1.01

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:



```
Testing Doc01: 1 Paulo Coelho - The Alchemist
Testing Doc02: 2 Paulo Coelho - The Zahir.txt
Testing Doc03: 3 James Joyce - Dubliners.txt
-----------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%)
-- ----- -------- -------- --------
1  0.10  99.981   99.949   89.933
2  0.20  99.979   97.816   87.154
3  0.30  99.978   91.365   81.706
4  0.40  99.975   88.721   74.585
5  0.50  99.972   82.808   70.633
6  0.60  99.971   79.368   63.707
7  0.70  99.965   77.896   52.961
8  0.80  99.962   77.540   51.822
9  0.90  99.955   75.371   49.666
10 1.00  99.951   73.731   44.595
-----------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

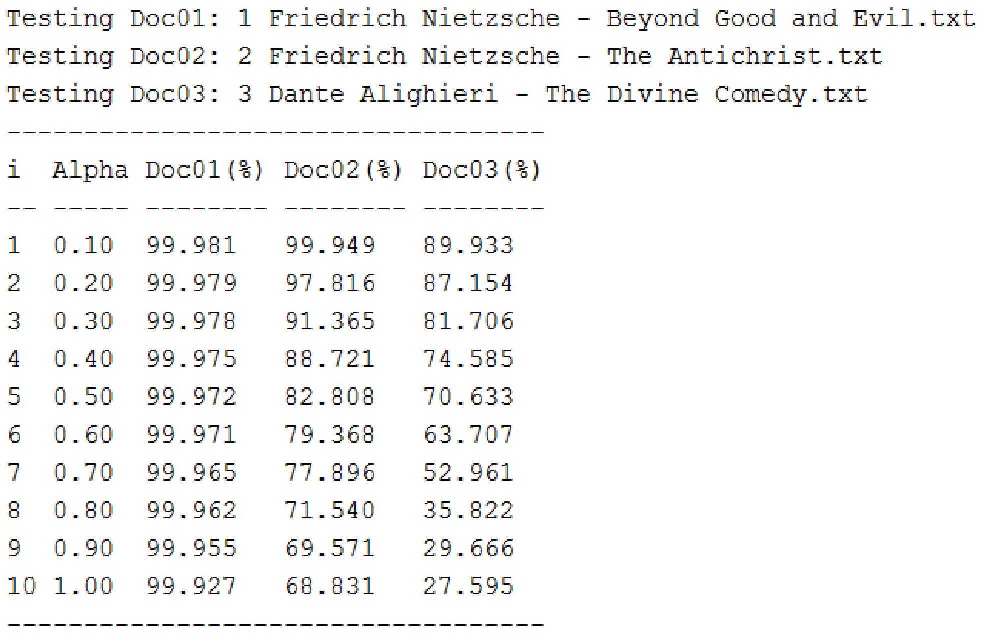
Maximum α: 1.01

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:



```
Testing Doc01: 1 Friedrich Nietzsche - Beyond Good and Evil.txt
Testing Doc02: 2 Friedrich Nietzsche - The Antichrist.txt
Testing Doc03: 3 Dante Alighieri - The Divine Comedy.txt
-----------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%)
-- ----- -------- -------- --------
1  0.10  99.981   99.949   89.933
2  0.20  99.979   97.816   87.154
3  0.30  99.978   91.365   81.706
4  0.40  99.975   88.721   74.585
5  0.50  99.972   82.808   70.633
6  0.60  99.971   79.368   63.707
7  0.70  99.965   77.896   52.961
8  0.80  99.962   71.540   35.822
9  0.90  99.955   69.571   29.666
10 1.00  99.927   68.831   27.595
-----------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

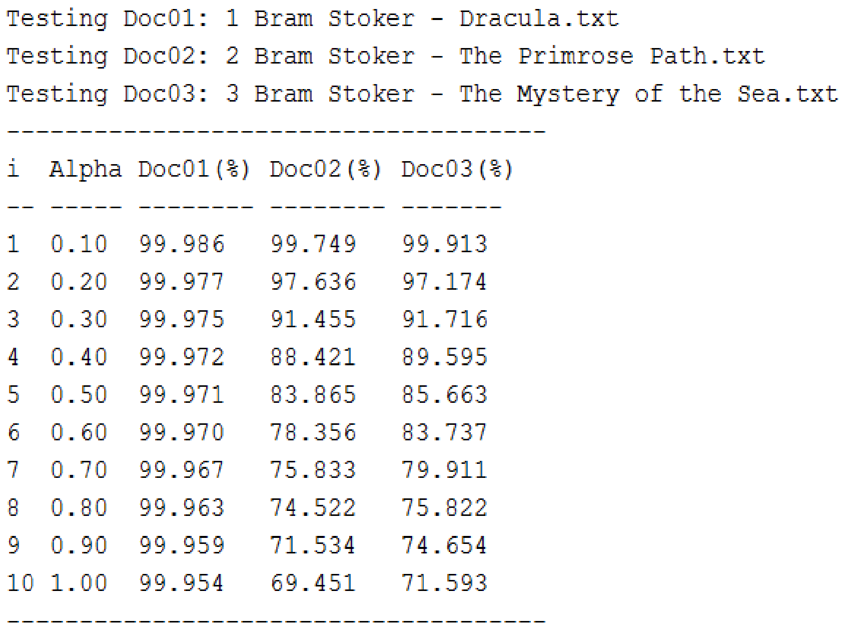
Maximum α: 1.01

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

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 α = 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 α: 0.1

Increment per iteration: 0.1

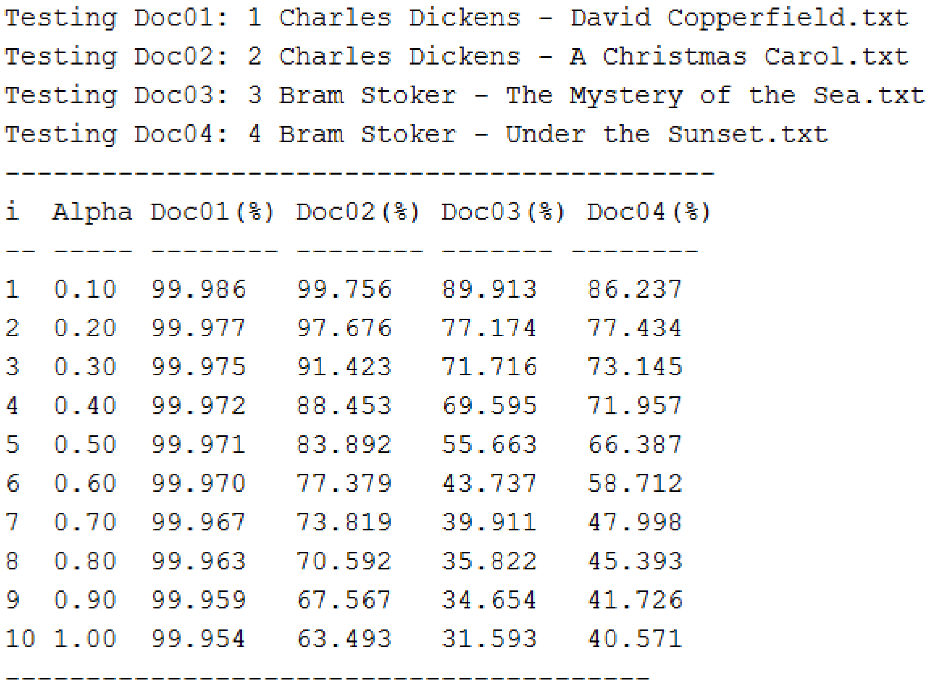
Maximum α: 1.01

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

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 α = 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 α: 0.1

Increment per iteration: 0.1

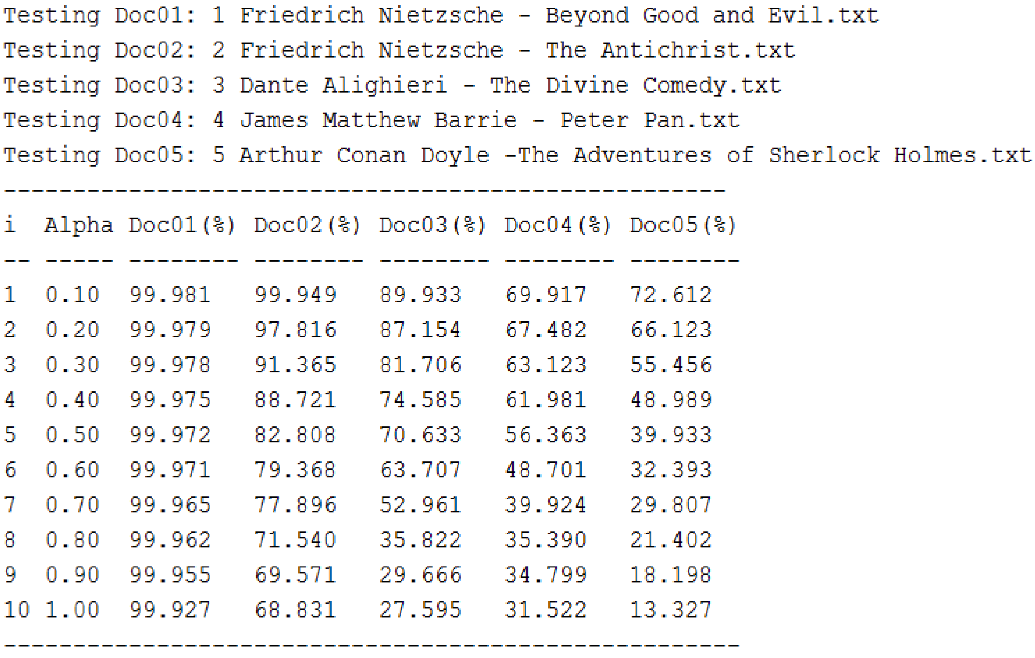
Maximum α: 1.01

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:



```
Testing Doc01: 1 Friedrich Nietzsche - Beyond Good and Evil.txt
Testing Doc02: 2 Friedrich Nietzsche - The Antichrist.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
---------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%)
-- ----- -------- -------- -------- -------- --------
1  0.10  99.981   99.949   89.933   69.917   72.612
2  0.20  99.979   97.816   87.154   67.482   66.123
3  0.30  99.978   91.365   81.706   63.123   55.456
4  0.40  99.975   88.721   74.585   61.981   48.989
5  0.50  99.972   82.808   70.633   56.363   39.933
6  0.60  99.971   79.368   63.707   48.701   32.393
7  0.70  99.965   77.896   52.961   39.924   29.807
8  0.80  99.962   71.540   35.822   35.390   21.402
9  0.90  99.955   69.571   29.666   34.799   18.198
10 1.00  99.927   68.831   27.595   31.522   13.327
----------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

Maximum α: 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
----------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%)
-- ----- -------- -------- -------- -------- --------
1  0.10  99.981   99.949   89.933   69.917   72.612
2  0.20  99.979   97.816   87.154   67.482   66.123
3  0.30  99.978   91.365   81.706   63.123   55.456
4  0.40  99.975   88.721   74.585   61.981   48.989
5  0.50  99.972   82.808   70.633   56.363   39.933
6  0.60  99.971   79.368   63.707   48.701   32.393
7  0.70  99.965   77.896   52.961   39.924   29.807
8  0.80  99.962   71.540   35.822   35.390   21.402
9  0.90  99.955   69.571   29.666   34.799   18.198
10 1.00  99.927   78.831   37.595   31.522   13.327
-----------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

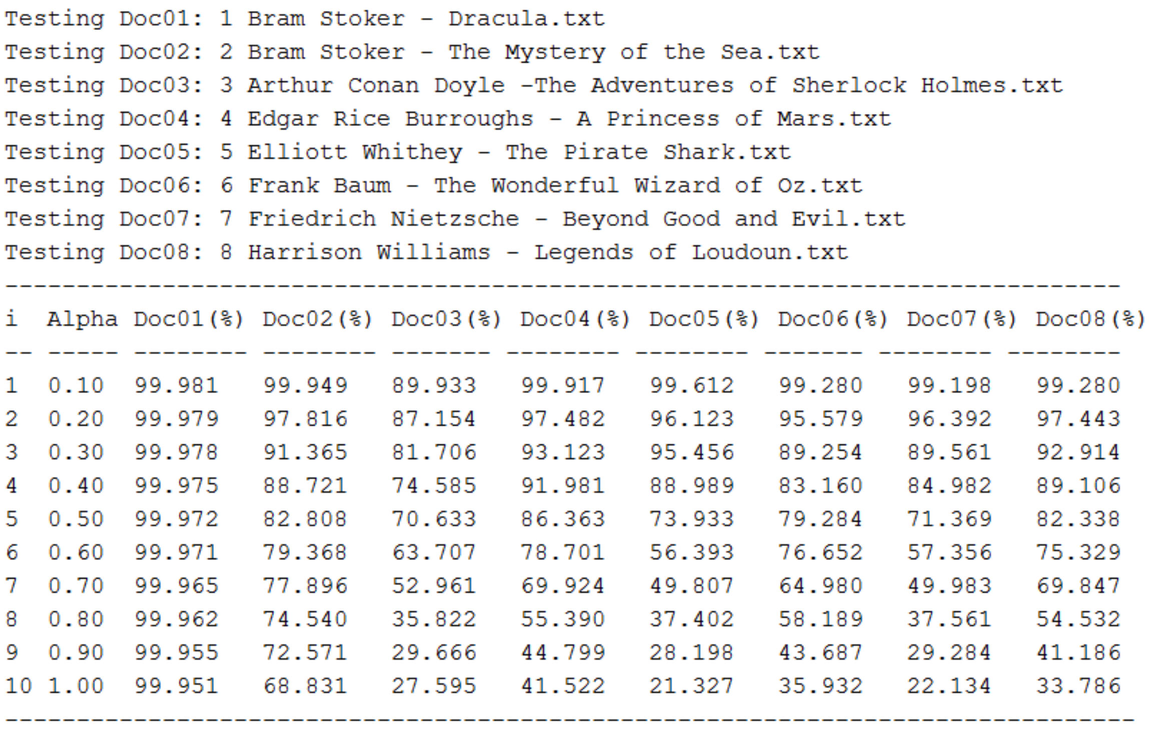
Maximum α: 1.01

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 Doc03: 3 Arthur Conan Doyle -The Adventures of Sherlock Holmes.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
------------------------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%) Doc07(%) Doc08(%)
-- ----- -------- -------- ------- -------- -------- ------- -------- --------
1  0.10  99.981   99.949   89.933   99.917   99.612   99.280   99.198   99.280
2  0.20  99.979   97.816   87.154   97.482   96.123   95.579   96.392   97.443
3  0.30  99.978   91.365   81.706   93.123   95.456   89.254   89.561   92.914
4  0.40  99.975   88.721   74.585   91.981   88.989   83.160   84.982   89.106
5  0.50  99.972   82.808   70.633   86.363   73.933   79.284   71.369   82.338
6  0.60  99.971   79.368   63.707   78.701   56.393   76.652   57.356   75.329
7  0.70  99.965   77.896   52.961   69.924   49.807   64.980   49.983   69.847
8  0.80  99.962   74.540   35.822   55.390   37.402   58.189   37.561   54.532
9  0.90  99.955   72.571   29.666   44.799   28.198   43.687   29.284   41.186
10 1.00  99.951   68.831   27.595   41.522   21.327   35.932   22.134   33.786
-------------------------------------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

Maximum α: 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:

```
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
Testing Doc04: 4 Bram Stoker - Under the Sunset.txt
Testing Doc05: 5 Bram Stoker - Miss Betty.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
------------------------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%) Doc07(%) Doc08(%)
-- ----- -------- -------- ------- -------- -------- ------- -------- --------
1  0.10  99.986   99.749   99.913   99.917   99.112   99.280   99.198   99.280
2  0.20  99.977   97.636   97.174   97.482   96.723   95.579   96.392   97.443
3  0.30  99.975   91.455   91.716   93.123   95.356   89.254   89.561   92.914
4  0.40  99.972   88.421   89.595   91.981   88.389   83.160   84.982   89.106
5  0.50  99.971   83.865   85.663   86.363   83.932   79.284   71.369   82.338
6  0.60  99.970   78.356   83.737   78.701   79.391   76.652   57.356   75.329
7  0.70  99.967   75.833   79.911   77.924   76.808   64.980   49.983   69.847
8  0.80  99.963   74.522   75.822   72.390   72.406   58.189   37.561   54.532
9  0.90  99.959   71.534   74.654   71.799   68.194   43.687   29.284   41.186
10 1.00  99.954   69.451   71.593   70.522   66.322   35.932   22.134   33.786
-------------------------------------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

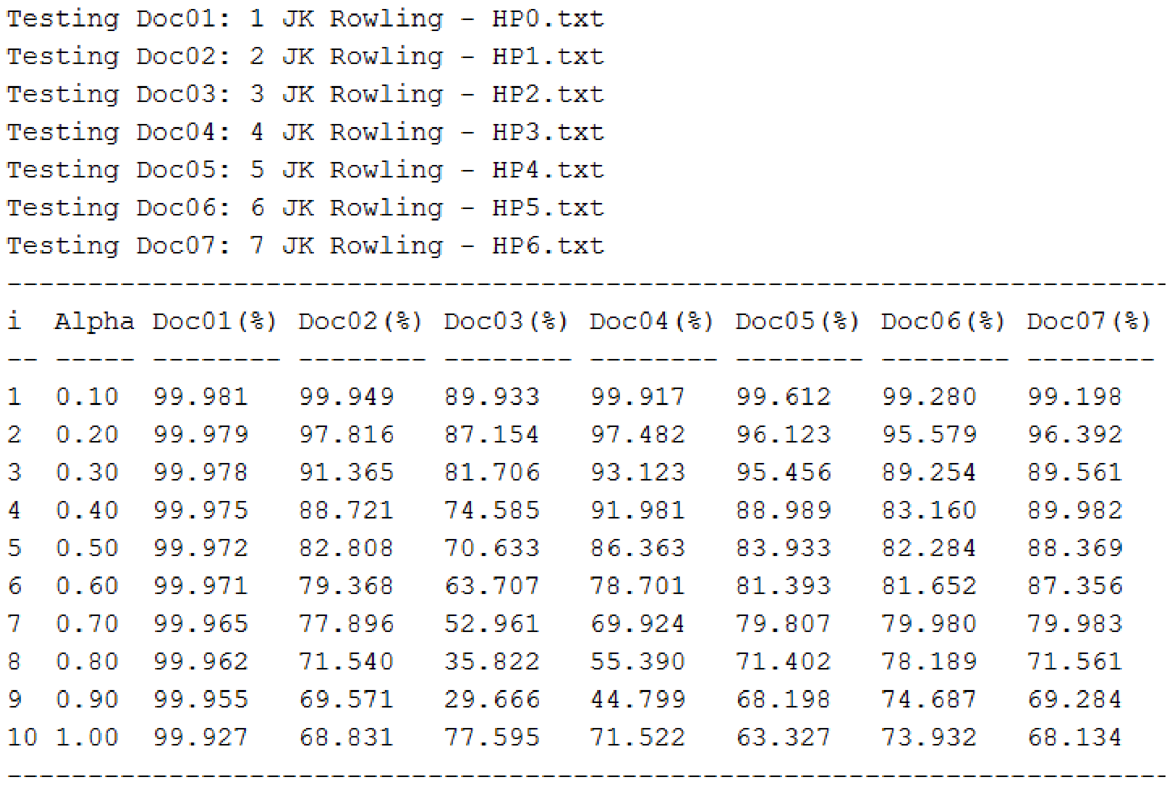
Maximum α: 1.01

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:



```
Testing Doc01: 1 JK Rowling - HP0.txt
Testing Doc02: 2 JK Rowling - HP1.txt
Testing Doc03: 3 JK Rowling - HP2.txt
Testing Doc04: 4 JK Rowling - HP3.txt
Testing Doc05: 5 JK Rowling - HP4.txt
Testing Doc06: 6 JK Rowling - HP5.txt
Testing Doc07: 7 JK Rowling - HP6.txt
------------------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%) Doc07(%)
-- ----- -------- -------- -------- -------- -------- -------- --------
1  0.10  99.981   99.949   89.933   99.917   99.612   99.280   99.198
2  0.20  99.979   97.816   87.154   97.482   96.123   95.579   96.392
3  0.30  99.978   91.365   81.706   93.123   95.456   89.254   89.561
4  0.40  99.975   88.721   74.585   91.981   88.989   83.160   89.982
5  0.50  99.972   82.808   70.633   86.363   83.933   82.284   88.369
6  0.60  99.971   79.368   63.707   78.701   81.393   81.652   87.356
7  0.70  99.965   77.896   52.961   69.924   79.807   79.980   79.983
8  0.80  99.962   71.540   35.822   55.390   71.402   78.189   71.561
9  0.90  99.955   69.571   29.666   44.799   68.198   74.687   69.284
10 1.00  99.927   68.831   77.595   71.522   63.327   73.932   68.134
------------------------------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

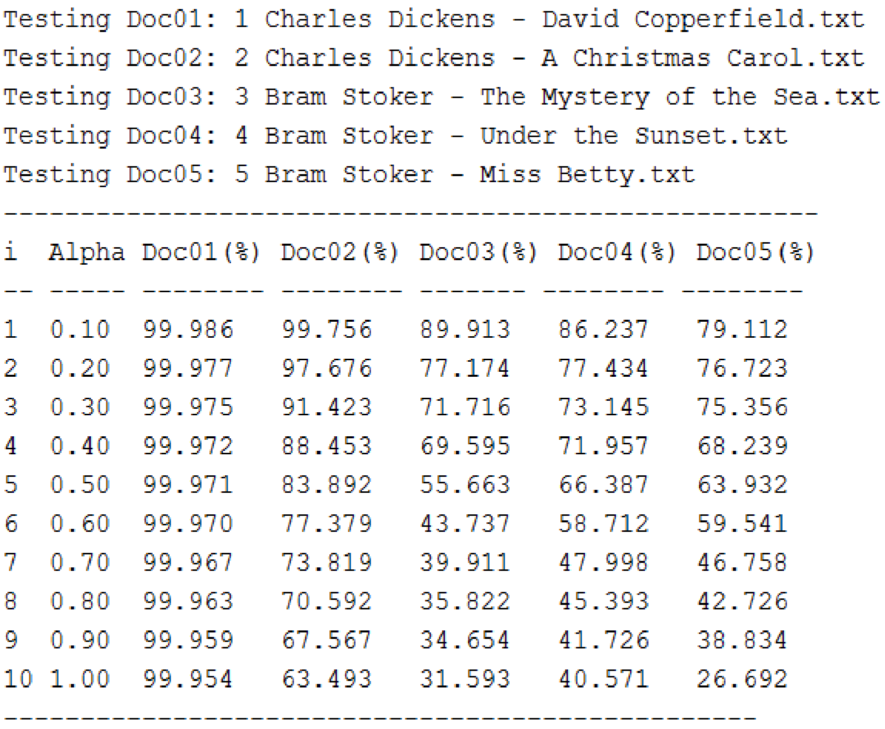
Maximum α: 1.01

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

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 α = 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 α: 0.1

Increment per iteration: 0.1

Maximum α: 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:

```
Testing Doc01: 1 JK Rowling - HP0.txt
Testing Doc02: 2 JK Rowling - HP5.txt
Testing Doc03: 3 Arthur Conan Doyle -The Adventures of Sherlock Holmes.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
----------------------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%) Doc07(%) Doc08(%)
-- ----- --------- --------- -------- --------- --------- --------- --------- ---------
1  0.10  99.981   99.949   89.933   99.917   99.612   99.280   99.198   99.280
2  0.20  99.979   97.816   87.154   97.482   96.123   95.579   96.392   97.443
3  0.30  99.978   91.365   81.706   93.123   95.456   89.254   89.561   92.914
4  0.40  99.975   88.721   74.585   91.981   88.989   83.160   84.982   89.106
5  0.50  99.972   82.808   70.633   86.363   73.933   79.284   71.369   82.338
6  0.60  99.971   79.368   63.707   78.701   56.393   76.652   57.356   75.329
7  0.70  99.965   77.896   52.961   69.924   49.807   64.980   49.983   69.847
8  0.80  99.962   71.540   35.822   55.390   37.402   58.189   37.561   44.532
9  0.90  99.955   69.571   29.666   44.799   28.198   43.687   29.284   38.186
10 1.00  99.927   68.831   27.595   31.522   23.327   33.932   28.134   32.786
-----------------------------------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

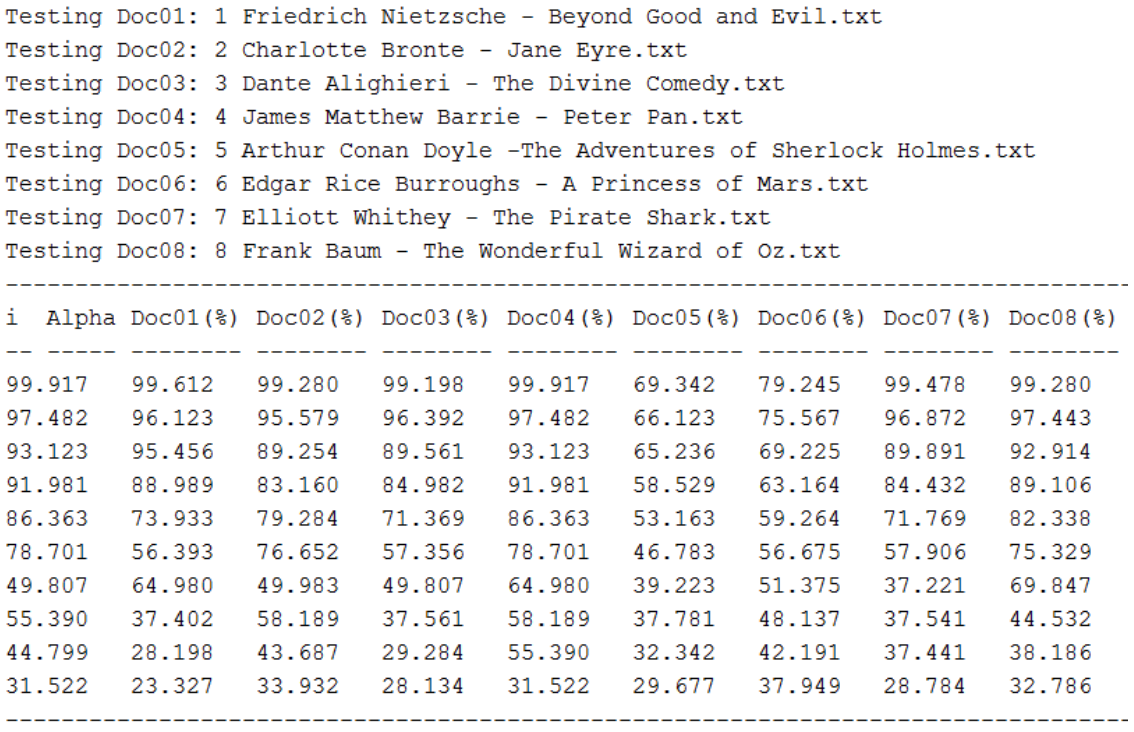
Maximum α: 1.01

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:



```
Testing Doc01: 1 Friedrich Nietzsche - Beyond Good and Evil.txt
Testing Doc02: 2 Charlotte Bronte - Jane Eyre.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
Testing Doc06: 6 Edgar Rice Burroughs - A Princess of Mars.txt
Testing Doc07: 7 Elliott Whithey - The Pirate Shark.txt
Testing Doc08: 8 Frank Baum - The Wonderful Wizard of Oz.txt
--------------------------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%) Doc07(%) Doc08(%)
-- ----- -------- -------- -------- -------- -------- -------- -------- --------
99.917   99.612   99.280   99.198   99.917   69.342   79.245   99.478   99.280
97.482   96.123   95.579   96.392   97.482   66.123   75.567   96.872   97.443
93.123   95.456   89.254   89.561   93.123   65.236   69.225   89.891   92.914
91.981   88.989   83.160   84.982   91.981   58.529   63.164   84.432   89.106
86.363   73.933   79.284   71.369   86.363   53.163   59.264   71.769   82.338
78.701   56.393   76.652   57.356   78.701   46.783   56.675   57.906   75.329
49.807   64.980   49.983   49.807   64.980   39.223   51.375   37.221   69.847
55.390   37.402   58.189   37.561   58.189   37.781   48.137   37.541   44.532
44.799   28.198   43.687   29.284   55.390   32.342   42.191   37.441   38.186
31.522   23.327   33.932   28.134   31.522   29.677   37.949   28.784   32.786
--------------------------------------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

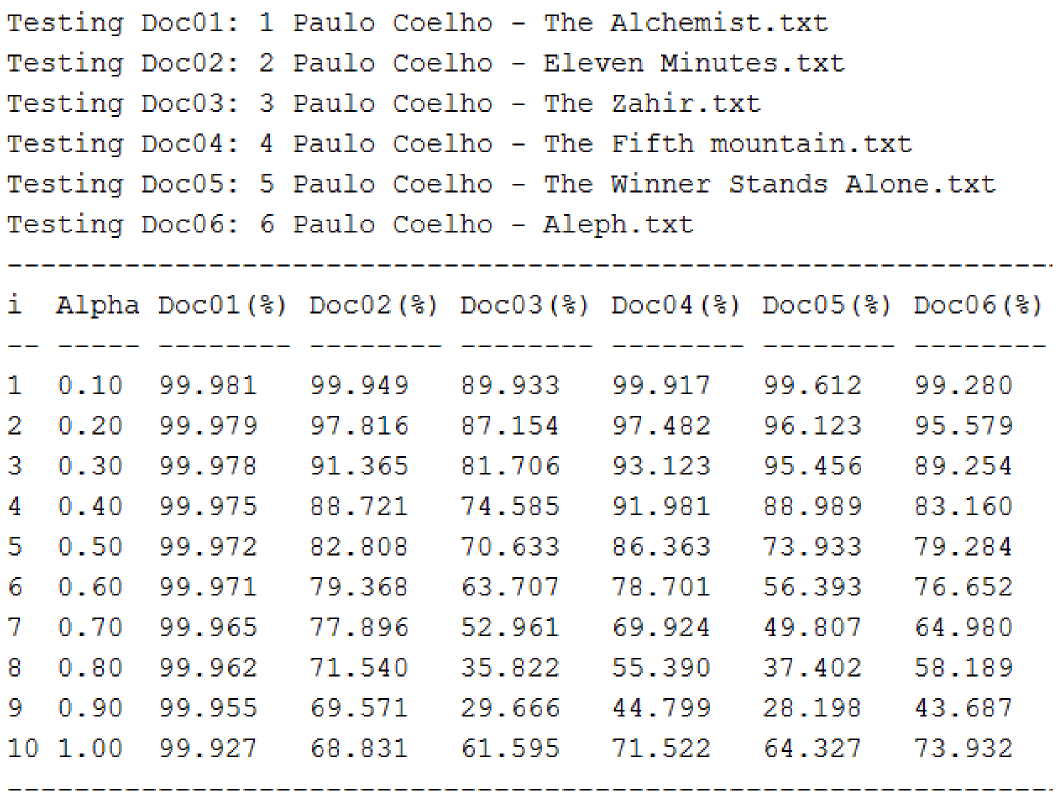
Maximum α: 1.01

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:



```
Testing Doc01: 1 Paulo Coelho - The Alchemist.txt
Testing Doc02: 2 Paulo Coelho - Eleven Minutes.txt
Testing Doc03: 3 Paulo Coelho - The Zahir.txt
Testing Doc04: 4 Paulo Coelho - The Fifth mountain.txt
Testing Doc05: 5 Paulo Coelho - The Winner Stands Alone.txt
Testing Doc06: 6 Paulo Coelho - Aleph.txt
---------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%)
-- ----- -------- -------- -------- -------- -------- --------
1  0.10  99.981   99.949   89.933   99.917   99.612   99.280
2  0.20  99.979   97.816   87.154   97.482   96.123   95.579
3  0.30  99.978   91.365   81.706   93.123   95.456   89.254
4  0.40  99.975   88.721   74.585   91.981   88.989   83.160
5  0.50  99.972   82.808   70.633   86.363   73.933   79.284
6  0.60  99.971   79.368   63.707   78.701   56.393   76.652
7  0.70  99.965   77.896   52.961   69.924   49.807   64.980
8  0.80  99.962   71.540   35.822   55.390   37.402   58.189
9  0.90  99.955   69.571   29.666   44.799   28.198   43.687
10 1.00  99.927   68.831   61.595   71.522   64.327   73.932
---------------------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

Maximum α: 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:

```
Testing Doc01: 1 Paulo Coelho - The Alchemist.txt
Testing Doc02: 2 Paulo Coelho - Eleven Minutes.txt
Testing Doc03: 3 Paulo Coelho - The Zahir.txt
Testing Doc04: 4 Paulo Coelho - The Fifth mountain.txt
Testing Doc05: 5 Paulo Coelho - The Winner Stands Alone.txt
Testing Doc06: 6 Paulo Coelho - Aleph.txt
Testing Doc07: 7 Karl Marx - Das Kapital.txt
-----------------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%) Doc07(%)
-- ----- -------- -------- -------- -------- -------- -------- --------
1  0.10  99.981   99.949   89.933   99.917   99.612   99.280   79.198
2  0.20  99.979   97.816   87.154   97.482   96.123   95.579   76.392
3  0.30  99.978   91.365   81.706   93.123   95.456   89.254   69.561
4  0.40  99.975   88.721   74.585   91.981   88.989   83.160   64.982
5  0.50  99.972   82.808   70.633   86.363   73.933   79.284   61.369
6  0.60  99.971   79.368   63.707   78.701   56.393   76.652   57.356
7  0.70  99.965   77.896   52.961   69.924   49.807   64.980   49.983
8  0.80  99.962   71.540   35.822   55.390   37.402   58.189   27.561
9  0.90  99.955   69.571   29.666   44.799   28.198   43.687   24.284
10 1.00  99.927   68.831   61.595   71.522   64.327   73.932   23.134
-----------------------------------------------------------------------
```

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 α = 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 α: 0.1

Increment per iteration: 0.1

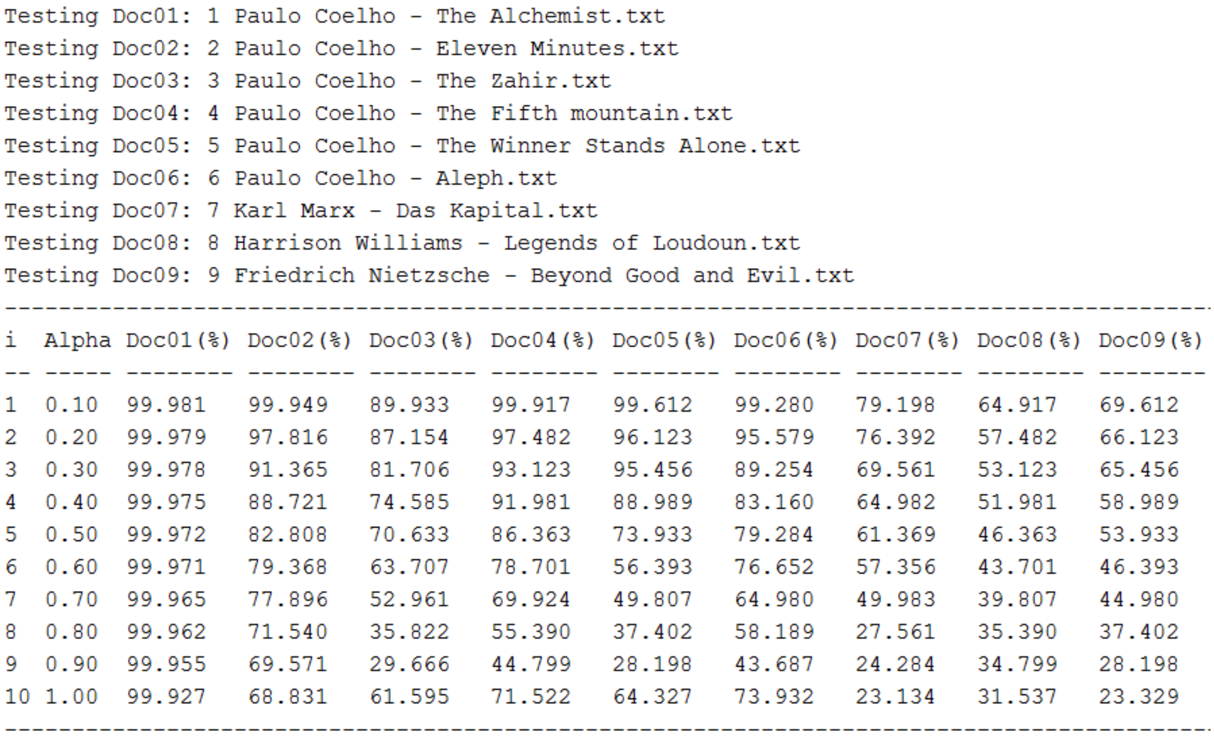
Maximum α: 1.01

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:



```
Testing Doc01: 1 Paulo Coelho - The Alchemist.txt
Testing Doc02: 2 Paulo Coelho - Eleven Minutes.txt
Testing Doc03: 3 Paulo Coelho - The Zahir.txt
Testing Doc04: 4 Paulo Coelho - The Fifth mountain.txt
Testing Doc05: 5 Paulo Coelho - The Winner Stands Alone.txt
Testing Doc06: 6 Paulo Coelho - Aleph.txt
Testing Doc07: 7 Karl Marx - Das Kapital.txt
Testing Doc08: 8 Harrison Williams - Legends of Loudoun.txt
Testing Doc09: 9 Friedrich Nietzsche - Beyond Good and Evil.txt
-----------------------------------------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%) Doc06(%) Doc07(%) Doc08(%) Doc09(%)
-- ----- -------- -------- -------- -------- -------- -------- -------- -------- --------
1  0.10  99.981   99.949   89.933   99.917   99.612   99.280   79.198   64.917   69.612
2  0.20  99.979   97.816   87.154   97.482   96.123   95.579   76.392   57.482   66.123
3  0.30  99.978   91.365   81.706   93.123   95.456   89.254   69.561   53.123   65.456
4  0.40  99.975   88.721   74.585   91.981   88.989   83.160   64.982   51.981   58.989
5  0.50  99.972   82.808   70.633   86.363   73.933   79.284   61.369   46.363   53.933
6  0.60  99.971   79.368   63.707   78.701   56.393   76.652   57.356   43.701   46.393
7  0.70  99.965   77.896   52.961   69.924   49.807   64.980   49.983   39.807   44.980
8  0.80  99.962   71.540   35.822   55.390   37.402   58.189   27.561   35.390   37.402
9  0.90  99.955   69.571   29.666   44.799   28.198   43.687   24.284   34.799   28.198
10 1.00  99.927   68.831   61.595   71.522   64.327   73.932   23.134   31.537   23.329
-----------------------------------------------------------------------------------------
```

Table 19: Result for test case ID: 19

Expected Result: There should be a high percentage match for the books written by the same author when α = 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 α: 0.1

Increment per iteration: 0.1

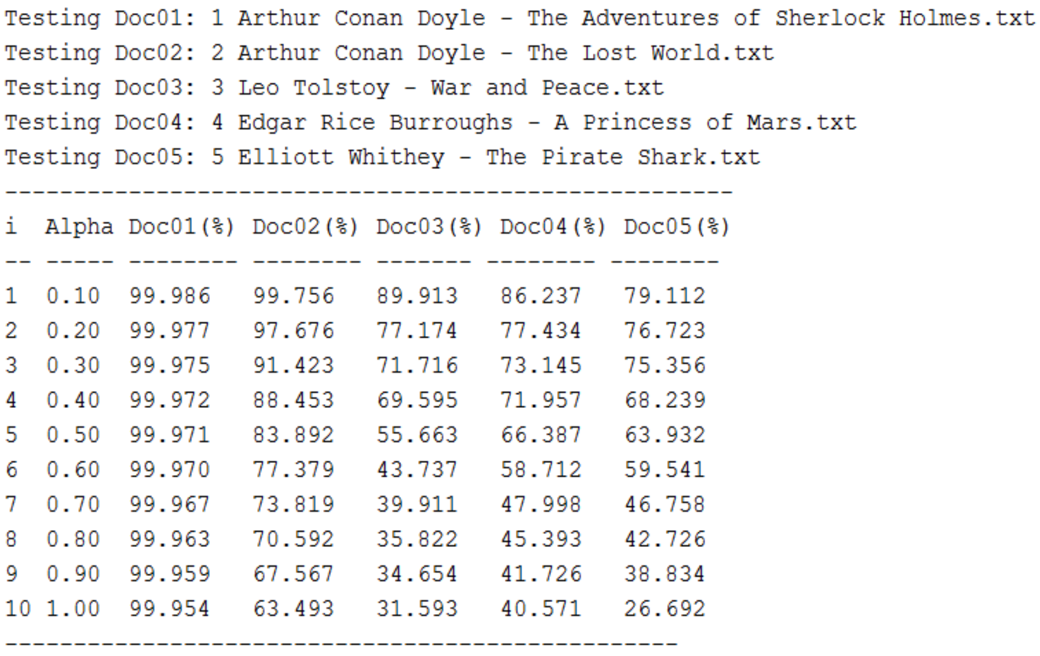
Maximum α: 1.01

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:



```
Testing Doc01: 1 Arthur Conan Doyle - The Adventures of Sherlock Holmes.txt
Testing Doc02: 2 Arthur Conan Doyle - The Lost World.txt
Testing Doc03: 3 Leo Tolstoy - War and Peace.txt
Testing Doc04: 4 Edgar Rice Burroughs - A Princess of Mars.txt
Testing Doc05: 5 Elliott Whithey - The Pirate Shark.txt
------------------------------------------------------
i  Alpha Doc01(%) Doc02(%) Doc03(%) Doc04(%) Doc05(%)
-- ----- -------- -------- ------- -------- --------
1  0.10  99.986   99.756   89.913   86.237   79.112
2  0.20  99.977   97.676   77.174   77.434   76.723
3  0.30  99.975   91.423   71.716   73.145   75.356
4  0.40  99.972   88.453   69.595   71.957   68.239
5  0.50  99.971   83.892   55.663   66.387   63.932
6  0.60  99.970   77.379   43.737   58.712   59.541
7  0.70  99.967   73.819   39.911   47.998   46.758
8  0.80  99.963   70.592   35.822   45.393   42.726
9  0.90  99.959   67.567   34.654   41.726   38.834
10 1.00  99.954   63.493   31.593   40.571   26.692
--------------------------------------------------
```

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 α = 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

1. "Free eBooks by Project Gutenberg,"[Online – May 2013].
   Available: http://www.gutenberg.org/
2. L. G. a. F. Morales, "Function Words,"[Online – June 2013].
   Available: http://www.sequencepublishing.com/academic.html.
3. T. K. a. e. all, Analyzing English Grammar, 6th edition: Longman, 2009.
4. T.Y. Lin, "Kolmogorov Complexity Based Automata," in *IEEE* International Conference on Granular Computing, Beijing, China, 2005.
5. P. Linz, An Introduction to Formal Languages and Automata, 4th edition, Sudbury: Jones and Bartlett Publishers, 2006, pp. 38-39.
6. F. M. Leah Gilner, "Function Words,"[Online – March 2013].
   Available: http://www.sequencepublishing.com
7. T. Y. Lin, "Rough Patterns in Data-Rough Sets and Foundation of Intrusion Detection Systems," Journal of Foundation of Computer Science and Decision Support, Vol.18, No. 3-4, 1993. 225-241.
8. T. Y. Lin, "Neighborhood Systems and Approximation in Database and Knowledge Base Systems", Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems, Poster Session, October 12-15, 1989, pp. 75-86
9. R.C.Carraso and J.Oncina: Learning stochastic regular grammars by means of a state merging method. Proceedings of the 2nd International Colloquium on Grammatical Inference. Lecture Notes in Artificial Intelligence (1994) 139-152.
10. Christopher Bishop: Pattern Recognition and Machine Learning (Information Science and Statistics) (2007)
11. Ferdinand Wagner, Ruedi Schmuki, Thomas Wagner, Peter Wolstenholme.: Modeling Software with Finite State Machines: A Practical Approach (2006)
12. Rajeev Motwani, Jeffrey D. Ullman, John E. Hopcroft: Introduction to Automata Theory, Languages, and Computation (2003)
13. Pierre Baldi, Soren Brunak: The Machine Learning Approach (Adaptive Computation and Machine Learning).
14. S. Zhang, "An Automata Based Authorship Identification System," M.S. Thesis, San Jose State University, San Jose, 2008.
15. Y. Lu, "An Automata Based Text Analysis System," M. S. Thesis, San Jose State University, San Jose, 2009.
16. Q. Yu, "Learning Author's Writing Pattern System," M. S. thesis, San Jose State University, San Jose, 2011.
17. A. Yazdhankhah, "Discovering Pattern using Automata," M. S. thesis, San Jose State University, San Jose, 2011.

APPENDIX A: Development Environment

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

Software Specifications	
Language	Java 1.7 Update 45
Integrated Development Environment	Netbeans 7.2
Operating System	Windows 7 Professional 64 bit

Table 21: Software Specifications

Hardware Specifications	
Model	HP Elitebook
RAM	8 GB
CPU	Intel® Core™ i5 vPro
Speed	3320M @ 2.60 GHz

Table 22: Hardware Specifications

APPENDIX B: List of EBook’s used

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

38	The Mystery of the Sea	Bram Stoker
39	Under the Sunset	Bram Stoker
40	The Wonderful Wizard of Oz	Frank Baum

Table 23: List of Ebook’s