Factoring Common Text in Documents

CPE 593: Applied Data Structures and Algorithms

14 May 2021

Trevor Dawideit

Gousemoodhin Nadaf

I pledge my honor that I have abided

by the Stevens Honor System

Table of Contents

Abstract 1

Introduction 1

Problem 1

Method 2

Pseudocode 8

Results 10

Conclusion 16

Abstract

The Project Gutenberg website provides free access to online eBooks. Each eBook on the website contains important legal information at the top and bottom of the document. This is referred to as the top and bottom boilerplates. The team has developed an algorithm that will extract the top and bottom boilerplates from a set of eBooks and are replaced with hypertext links to a separate file containing the boilerplate information. This combines all boilerplates into a few files as slight changes to them have been made over the years. The algorithm runs in O(n) time, where n is the total number of lines in all the files. The algorithm developed works as intended but code optimizations exist that can reduce runtime by a considerable amount.

Introduction

Project Gutenberg is an online library that provides free eBooks. The beginning and ending of each eBook contains important copyright and legal information about the use of the eBook both inside and outside the United States. These sections are called boilerplates. Project Gutenberg now has a collection of over 60,000 eBooks available for anyone to read. This means there are over 60,000 boilerplates on Project Gutenberg.

Problem

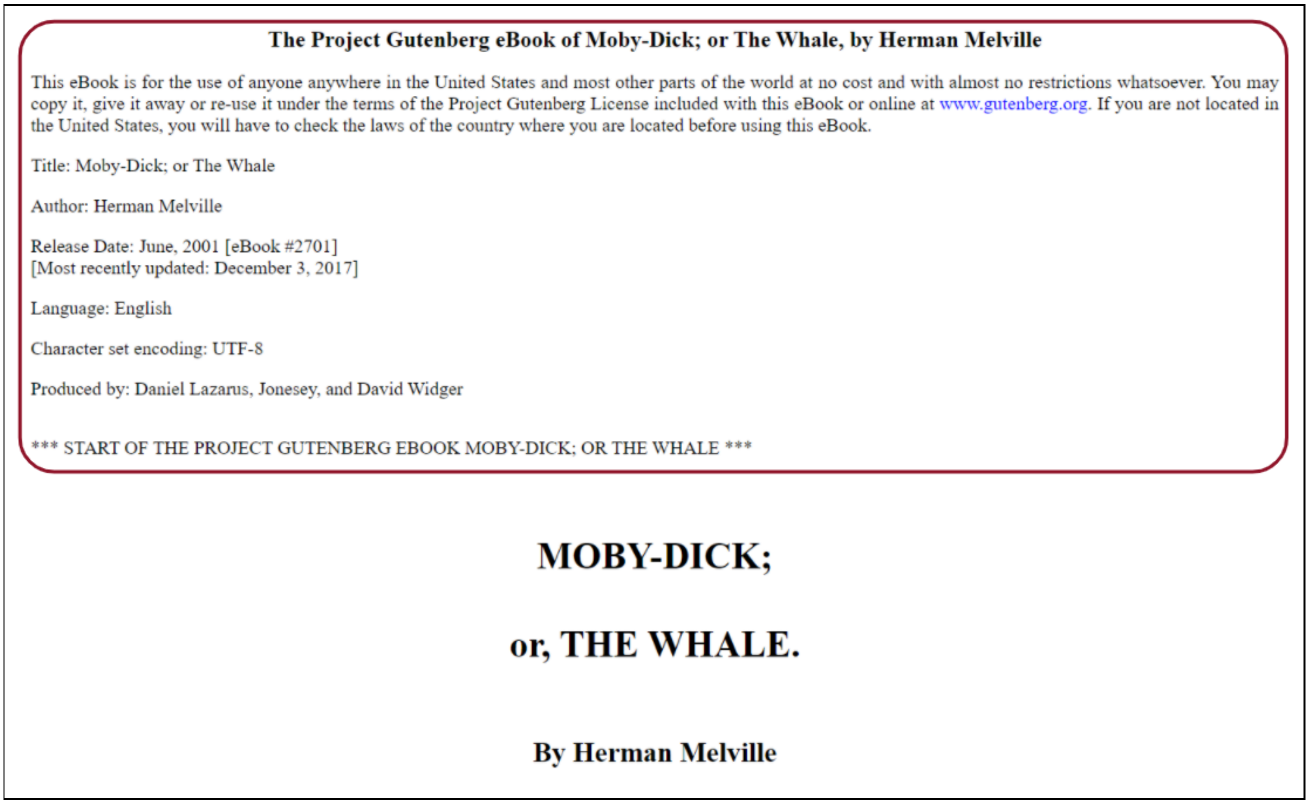
As Project Gutenberg’s collection of eBooks grows, constantly adding the top and bottom boilerplates can waste a lot of space and be difficult to maintain. If the boilerplate information were to be moved to a separate file, with each document containing hyperlinks to the information, it would save space, reduce redundancy, and allow the user to get reading right away. Unfortunately, over the years, Project Gutenberg has changed their legal information and therefore have changed their boilerplate.

Method

The team's solution is to implement an algorithm that replaces the top and bottom boilerplates with links to files that contain said boilerplates. This would save space as multiple files can point to the same set of boilerplates. Moreover, maintaining the legal information will be much easier since only the boilerplate files need to be updated.

The first part of the algorithm collects the top and bottom text from a Project Gutenberg eBook HTML file. This is done by parsing through the entire file line by line. As the algorithm reads in each line, it is stored in one of four different vectors. These vectors are named topFileContent, topBoilerplate, mainFileContent, and bottomBoilerplate. Lines are initially sent to topFileContent. The topFileContent vector then contains all the header information of the HTML file. Once the top boilerplate is reached, the collection vector is switched to topBoilerplate.

Figure 1 shows an example of a top boilerplate for the ebook Moby-Dick. The top boilerplate is outlined in the red box and ends at the line: “*** START OF THE PROJECT GUTENBERG EBOOK MOBY-DICK; OR THE WHALE ***.” However, there are multiple lines of text that will be specific to each eBook. This is information such as the title, author, and release of the eBook. Therefore, this information will need to be excluded from the common text collection.



The Project Gutenberg eBook of Moby-Dick; or The Whale, by Herman Melville

This eBook is for the use of anyone anywhere in the United States and most other parts of the world at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org. If you are not located in the United States, you will have to check the laws of the country where you are located before using this eBook.

Title: Moby-Dick; or The Whale

Author: Herman Melville

Release Date: June, 2001 [eBook #2701]
[Most recently updated: December 3, 2017]

Language: English

Character set encoding: UTF-8

Produced by: Daniel Lazarus, Jonesey, and David Widger

*** START OF THE PROJECT GUTENBERG EBOOK MOBY-DICK; OR THE WHALE ***

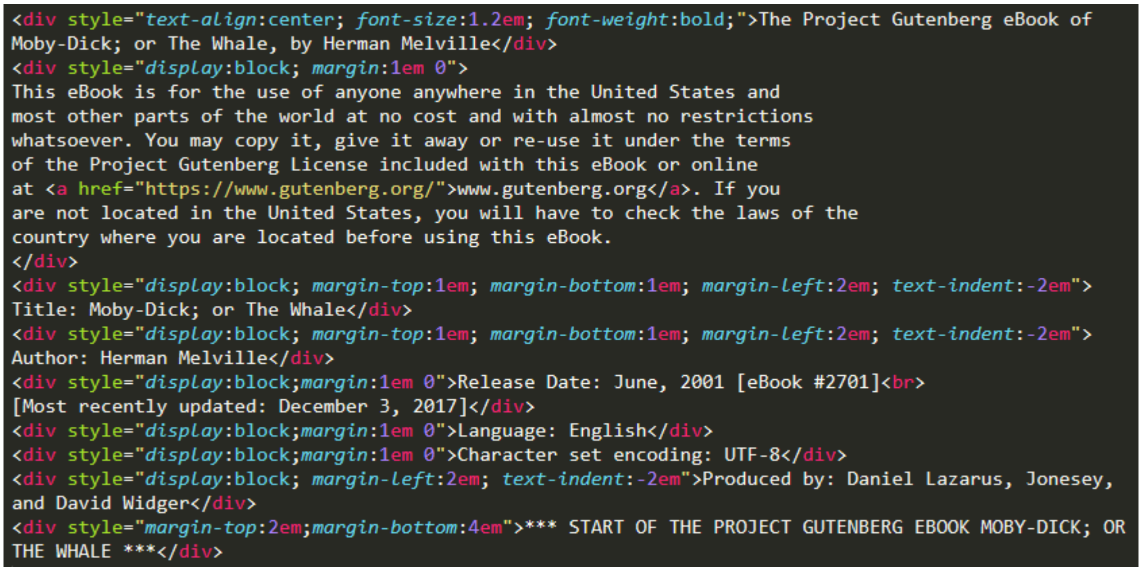
MOBY-DICK;

or, THE WHALE.

By Herman Melville

Figure 1: Example top boilerplate

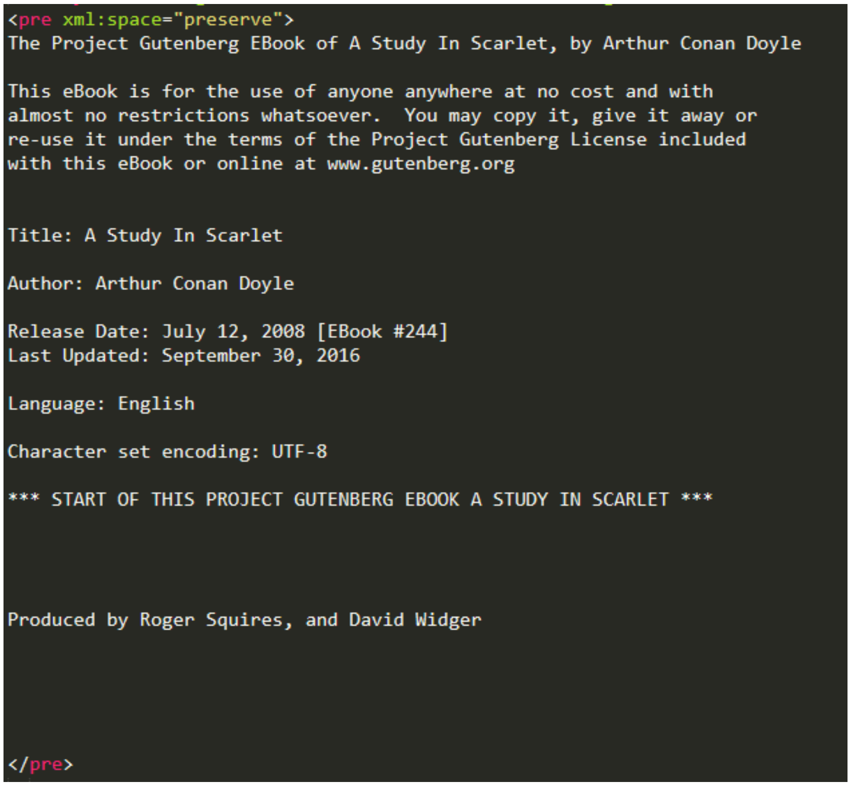
Figure 2 shows the HTML code for the top boilerplate. Looking at the HTML code, the common text that needs to be extracted is located in an HTML <div> tag. Therefore, the point where the algorithm switches collection vectors from topFileContent to topBoilerplate is the line after the <div> tag that contains the title of the eBook. The top boilerplate collection then runs until it reaches the closing </div> tag. Once there, the collection vector changes to mainFileContent.



```
<div style="text-align:center; font-size:1.2em; font-weight:bold;">The Project Gutenberg eBook of
Moby-Dick; or The Whale, by Herman Melville</div>
<div style="display:block; margin:1em 0">
This eBook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this eBook or online
at <a href="https://www.gutenberg.org/">www.gutenberg.org</a>. If you
are not located in the United States, you will have to check the laws of the
country where you are located before using this eBook.
</div>
<div style="display:block; margin-top:1em; margin-bottom:1em; margin-left:2em; text-indent:-2em">
Title: Moby-Dick; or The Whale</div>
<div style="display:block; margin-top:1em; margin-bottom:1em; margin-left:2em; text-indent:-2em">
Author: Herman Melville</div>
<div style="display:block;margin:1em 0">Release Date: June, 2001 [eBook #2701]<br>
[Most recently updated: December 3, 2017]</div>
<div style="display:block;margin:1em 0">Language: English</div>
<div style="display:block;margin:1em 0">Character set encoding: UTF-8</div>
<div style="display:block; margin-left:2em; text-indent:-2em">Produced by: Daniel Lazarus, Jonesey,
and David Widger</div>
<div style="margin-top:2em;margin-bottom:4em">*** START OF THE PROJECT GUTENBERG EBOOK MOBY-DICK; OR
THE WHALE ***</div>
```

Figure 2: HTML for top boilerplate

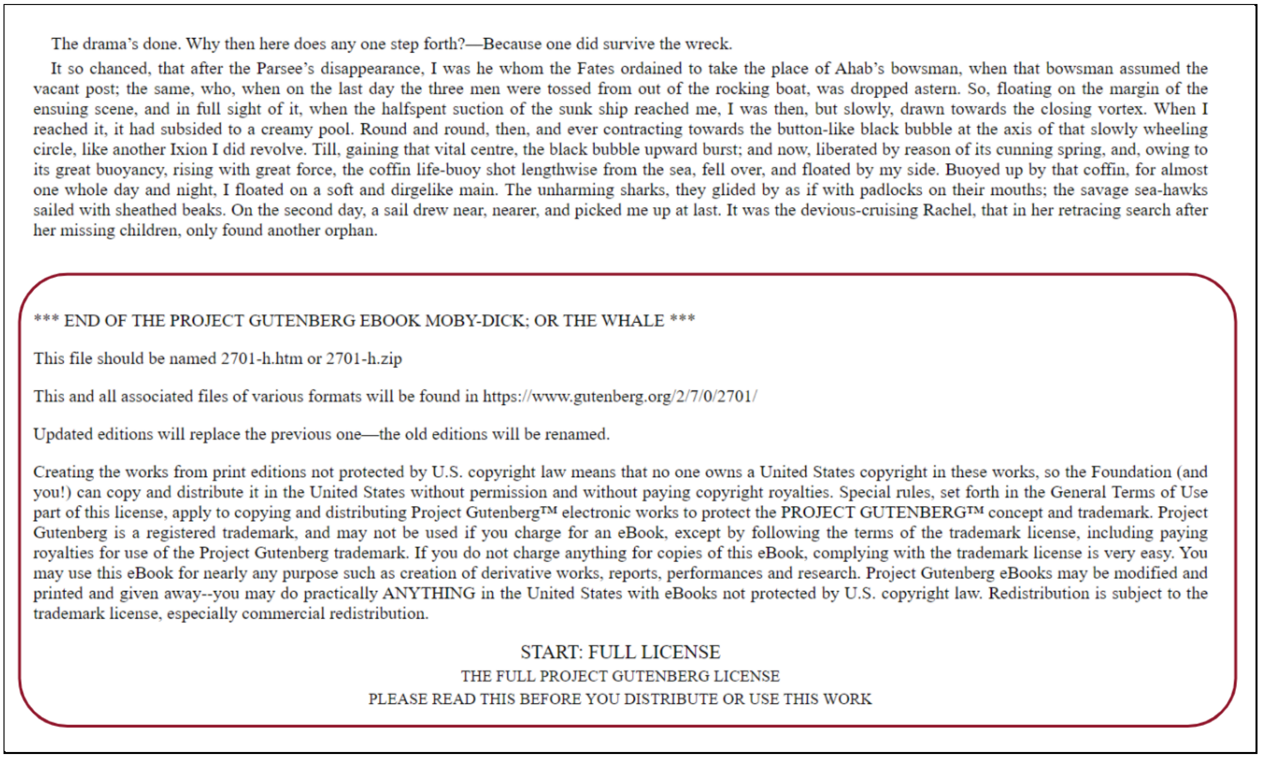
However, this case for the top boilerplate is not always true. As seen in figure 3, the HTML that contains the top boilerplate can vary from file to file. In this file, the top boilerplate information is held within a <pre> tag. For this, the ending condition must be changed. For the <pre> tag case, topBoilerplate collection ends once the text “Title: ” is found. The line with “Title: ” is then stored in mainFileContent and the collection vector is changed to mainFileContent. For all of the 20 eBooks chosen, they were either one of these two cases.



```
<pre xml:space="preserve">
The Project Gutenberg EBook of A Study In Scarlet, by Arthur Conan Doyle

This eBook is for the use of anyone anywhere at no cost and with
almost no restrictions whatsoever.  You may copy it, give it away or
re-use it under the terms of the Project Gutenberg License included
with this eBook or online at www.gutenberg.org


Title: A Study In Scarlet

Author: Arthur Conan Doyle

Release Date: July 12, 2008 [EBook #244]
Last Updated: September 30, 2016

Language: English

Character set encoding: UTF-8

*** START OF THIS PROJECT GUTENBERG EBOOK A STUDY IN SCARLET ***




Produced by Roger Squires, and David Widger





</pre>
```

Figure 3: Alternate HTML for top boilerplate

Figure 4 shows an example beginning of the bottom boilerplate, outlined in red. The bottom boilerplate begins after the line: “*** END OF THE PROJECT GUTENBERG EBOOK MOBY-DICK; OR THE WHALE ***.” However, much like the top boilerplate, the bottom has a few lines after that are specific to each eBook. These lines contain information about the naming conversion for the eBook. Therefore, this information will have to be skipped as well.



The drama's done. Why then here does any one step forth?—Because one did survive the wreck.

It so chanced, that after the Parsee's disappearance, I was he whom the Fates ordained to take the place of Ahab's bowsman, when that bowsman assumed the vacant post; the same, who, when on the last day the three men were tossed from out of the rocking boat, was dropped astern. So, floating on the margin of the ensuing scene, and in full sight of it, when the halfspent suction of the sunk ship reached me, I was then, but slowly, drawn towards the closing vortex. When I reached it, it had subsided to a creamy pool. Round and round, then, and ever contracting towards the button-like black bubble at the axis of that slowly wheeling circle, like another Ixion I did revolve. Till, gaining that vital centre, the black bubble upward burst; and now, liberated by reason of its cunning spring, and, owing to its great buoyancy, rising with great force, the coffin life-buoy shot lengthwise from the sea, fell over, and floated by my side. Buoyed up by that coffin, for almost one whole day and night, I floated on a soft and dirgelike main. The unharming sharks, they glided by as if with padlocks on their mouths; the savage sea-hawks sailed with sheathed beaks. On the second day, a sail drew near, nearer, and picked me up at last. It was the devious-cruising Rachel, that in her retracing search after her missing children, only found another orphan.

*** END OF THE PROJECT GUTENBERG EBOOK MOBY-DICK; OR THE WHALE ***

This file should be named 2701-h.htm or 2701-h.zip

This and all associated files of various formats will be found in https://www.gutenberg.org/2/7/0/2701/

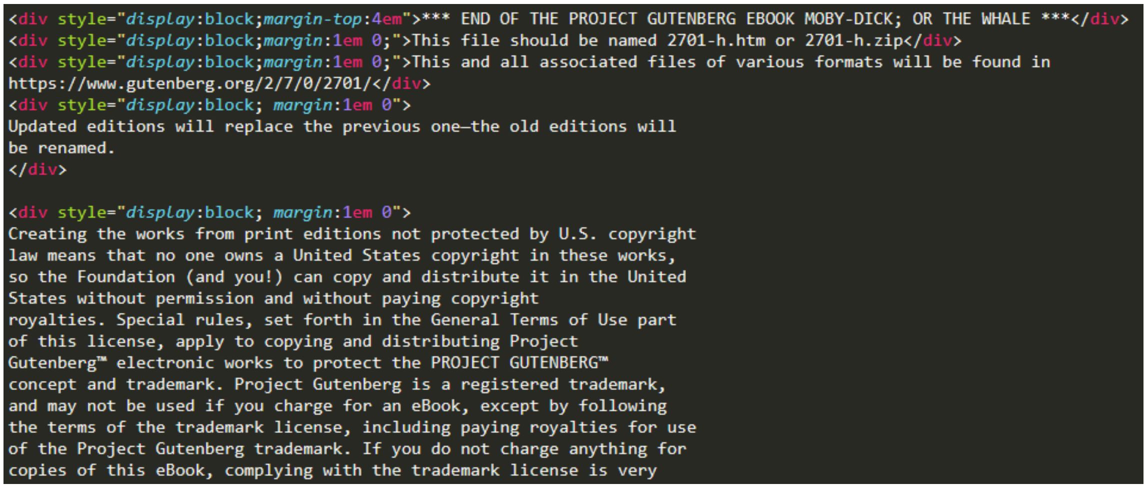
Updated editions will replace the previous one—the old editions will be renamed.

Creating the works from print editions not protected by U.S. copyright law means that no one owns a United States copyright in these works, so the Foundation (and you!) can copy and distribute it in the United States without permission and without paying copyright royalties. Special rules, set forth in the General Terms of Use part of this license, apply to copying and distributing Project Gutenberg™ electronic works to protect the PROJECT GUTENBERG™ concept and trademark. Project Gutenberg is a registered trademark, and may not be used if you charge for an eBook, except by following the terms of the trademark license, including paying royalties for use of the Project Gutenberg trademark. If you do not charge anything for copies of this eBook, complying with the trademark license is very easy. You may use this eBook for nearly any purpose such as creation of derivative works, reports, performances and research. Project Gutenberg eBooks may be modified and printed and given away--you may do practically ANYTHING in the United States with eBooks not protected by U.S. copyright law. Redistribution is subject to the trademark license, especially commercial redistribution.

START: FULL LICENSE
THE FULL PROJECT GUTENBERG LICENSE
PLEASE READ THIS BEFORE YOU DISTRIBUTE OR USE THIS WORK

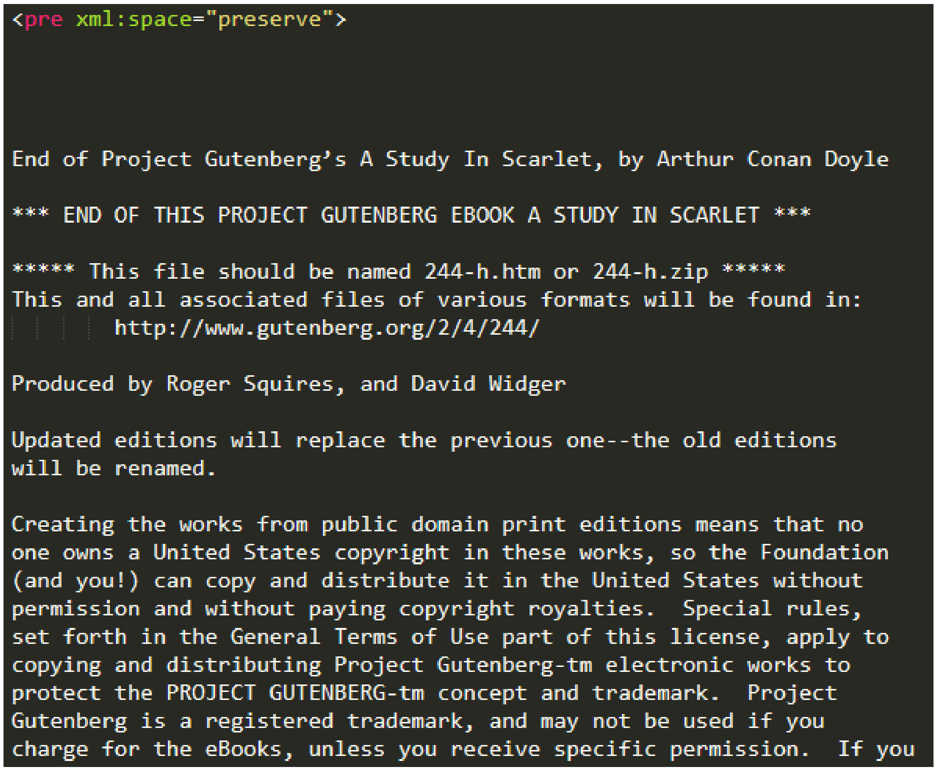
Figure 4: Example bottom boilerplate

Moreover, figures 5 and 6 show two different HTML outputs for the bottom boilerplate. It is again like the top boilerplate case with files either using <div> (Figure 5) or <pre> (Figure 6) tags. To determine with to switch the collection vector from mainFileContent to bottomBoilerplate, the algorithm looks for the line "*** END OF THE PROJECT GUTENBERG EBOOK MOBY-DICK; OR THE WHALE ***." Then, depending on what tags are being used, a countdown variable is initiated. Once the variable reaches zero, the collection vector changes to bottomBoilerplate and runs until the end of the file is reached.



```
<div style="display:block;margin-top:4em">*** END OF THE PROJECT GUTENBERG EBOOK MOBY-DICK; OR THE WHALE ***</div>
<div style="display:block;margin:1em 0;">This file should be named 2701-h.htm or 2701-h.zip</div>
<div style="display:block;margin:1em 0;">This and all associated files of various formats will be found in
https://www.gutenberg.org/2/7/0/2701/</div>
<div style="display:block; margin:1em 0">
Updated editions will replace the previous one—the old editions will
be renamed.
</div>

<div style="display:block; margin:1em 0">
Creating the works from print editions not protected by U.S. copyright
law means that no one owns a United States copyright in these works,
so the Foundation (and you!) can copy and distribute it in the United
States without permission and without paying copyright
royalties. Special rules, set forth in the General Terms of Use part
of this license, apply to copying and distributing Project
Gutenberg™ electronic works to protect the PROJECT GUTENBERG™
concept and trademark. Project Gutenberg is a registered trademark,
and may not be used if you charge for an eBook, except by following
the terms of the trademark license, including paying royalties for use
of the Project Gutenberg trademark. If you do not charge anything for
copies of this eBook, complying with the trademark license is very
```

Figure 5: HTML for bottom boilerplate, <div> case



```
<pre xml:space="preserve">




End of Project Gutenberg's A Study In Scarlet, by Arthur Conan Doyle

*** END OF THIS PROJECT GUTENBERG EBOOK A STUDY IN SCARLET ***

***** This file should be named 244-h.htm or 244-h.zip *****
This and all associated files of various formats will be found in:
        http://www.gutenberg.org/2/4/244/

Produced by Roger Squires, and David Widger

Updated editions will replace the previous one--the old editions
will be renamed.

Creating the works from public domain print editions means that no
one owns a United States copyright in these works, so the Foundation
(and you!) can copy and distribute it in the United States without
permission and without paying copyright royalties.  Special rules,
set forth in the General Terms of Use part of this license, apply to
copying and distributing Project Gutenberg-tm electronic works to
protect the PROJECT GUTENBERG-tm concept and trademark.  Project
Gutenberg is a registered trademark, and may not be used if you
charge for the eBooks, unless you receive specific permission.  If you
```

Figure 6: HTML for bottom boilerplate, <pre> case

After all the common text has been collected, the top and bottom boilerplates are compared with all currently existing top and bottom boilerplates. This is done by comparing if the vectors are the same size. If they are, then each element is checked to see if they are the same. If they do not match any of the currently existing boilerplates, a new file is created with the collected text. The algorithm then links the eBook to the appropriate boilerplate file. Finally, the top and bottom common text in the eBook is replaced with hypertext links. This is done by creating a new file, writing all the text in topFileContent, then inserting the top boilerplate link, writing all the text in mainFileContent, and then inserting the bottom boilerplate link.

The algorithm will be tested using twenty eBooks from Project Gutenberg. These eBooks are chosen at random from the top 100 pages on the Project Gutenberg website. In order to test the flexibility of the algorithm, of the twenty randomly selected, at least two different top and bottom boilerplates are required.

Pseudocode

```
For each file in path: O(F), F = number of files
  If (file ends in .html):
    Create a new eBook
    Start location at topFileContent
    While (file has line): O(M) M = number of lines in file
      If not in preMode:
        If line is not empty:
          Add line to location
          If location is topFileContent and line does not have “<title>”:
            Change location to topBoilerPlate
          Else If location is topBoilerplate and line contains “</div>”:
            Change location to mainFileContent
          Else If location is topBoilerplate and line contains “Title: ”:
            Change location to mainFileContent
            Enable preMode
            Move line from topBoilerplate to mainFileContent
          Else If location is mainFileContent and line contains “*** END OF”:
            Enable startbottomBPCountdown
          Else If startbottomBPCountdown is enabled:
            Decrease bottomBPCount by 1
            If bottomBPCount is zero:
              Change location to bottomBoilerplate
              Disable startbottomBPCountdown
            End
          End
        End
      Else:
        Add line to location
        If location is mainFileContent and line contains “*** END OF”:
          Enable startbottomBPCountdown
        Else If startbottomBPCountdown is enabled and line contains “Creating ”:
          Change location to bottomBoilerplate
          Disable startbottomBPCountdown
          Move line from mainFileContentto bottomBoilerplate
        End
      End
    End
    Compare topBoilerplate to each eBook already created O(K), K = length of boilerplate * number
of existing boilerplates)
    if(topBoilerplate  does not already exist):
      Create new html file with topBP contents
    Else:
      Link eBook to already existing html file that contains matching text
    End
    Compare bottomBoilerplate to each eBook already created O(K), K = length of boilerplate *
number of existing boilerplates)
    if(bottomBoilerplate  does not already exist):
      Create new html file with bottomBP contents
    Else:
      Link eBook to already existing html file that contains matching text
    End
    Create new html file O(L), L = length of topFileContent + length of mainFileContent
    Insert all topFileContent to file
    Insert link to topBoilerplate to file
    Insert all mainFileContent to file
    Insert link to bottomBoilerplate to file
  Else:
    Copy file contents to output location
  End
End
```

The outer loop runs for all files making the outer loop O(F), where F is the total number of files. The inner loop runs for every line in a file making it O(M), where M is the number of lines in the file. Once outside of the inner loop, the boilerplate comparison is O(K), where K is the length of boilerplate * number of existing boilerplates. This comparison is run for both top and bottom boilerplates. Finally, the output file creation is O(L) where L is the length of topFileContent + length of mainFileContent. Combining this results in:

$$O(F * (M + 2K + L))$$

This can be simplified down further. L can be treated as M times some number between 0 and 1 as topFileContent and mainFileContext are all the lines in the file excluding the top and bottom boilerplates. Let's call this multiplier X. The number of lines in the boilerplates would be all the information left in the file and can be rewritten as (1-X)*M. Therefore, K can be changed to (1-X)*M*I where I is the number of existing boilerplates. This changes the complexity to:

$$O(F * (M + 2(1-X)*M*I + X*M)) = O(FM * (1 + 2(1-X)*I + X))$$

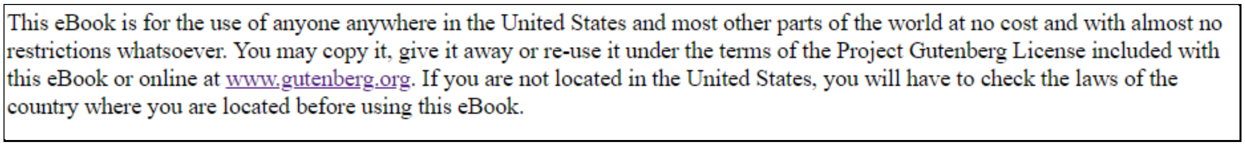
The (1 + 2(1-X)*I + X) will be much smaller than the FM term and can therefore be dropped, leaving the total complexity as O(FM), where F is the number of files and M is the number of lines per file. This can be represented as O(n) where n is the total number of lines in all the files.

For sigma notation, it will be the same. While the boilerplate comparison can be short circuited if none of the boilerplates have the same length, the inner and outer loops will run in full. This gives a best case run scenario of Ω(FM), where F is the number of files and M is the number of lines per file. Much like the big O case, this can be represented as Ω(n) where n is the total number of lines in all the files.

Results

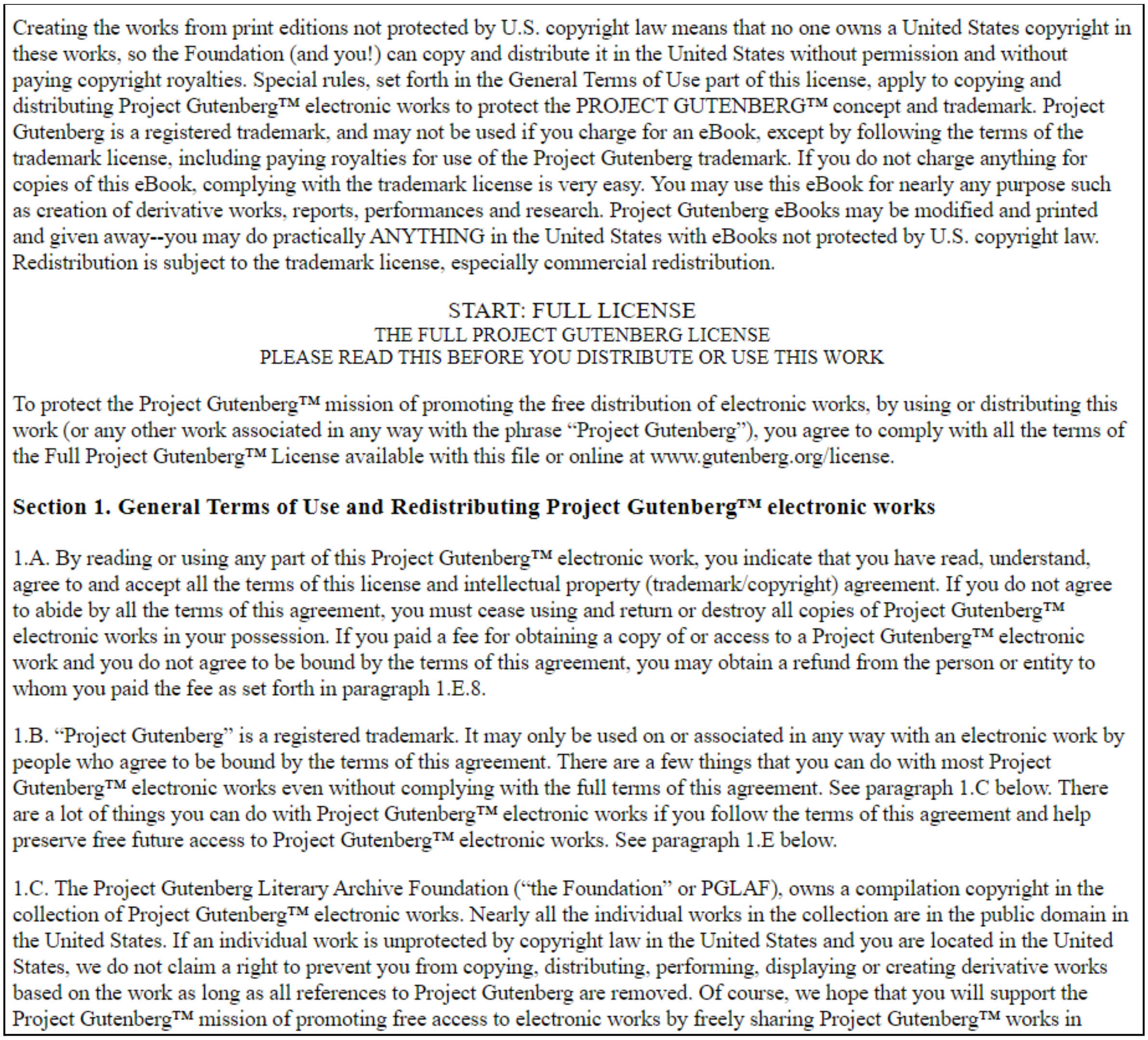
The developed algorithm ran successfully with no errors with the twenty test eBook HTML files. Three different top boilerplate HTML files and four different bottom boilerplate HTML files were created. This implies that the algorithm did find common text between the twenty eBooks as there are less boilerplate files than there are eBooks tested.

Figure 3 shows one of the top boilerplate HTML files and figure 4 shows one of the bottom boilerplate HTML files. Looking at figures 3 and 4, it is clear that the common text extraction is working as intended. The top and bottom common text is successfully separated from the eBook and placed into separate files.



This eBook is for the use of anyone anywhere in the United States and most other parts of the world at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org. If you are not located in the United States, you will have to check the laws of the country where you are located before using this eBook.

Figure 3: Top boilerplate HTML file text



Creating the works from print editions not protected by U.S. copyright law means that no one owns a United States copyright in these works, so the Foundation (and you!) can copy and distribute it in the United States without permission and without paying copyright royalties. Special rules, set forth in the General Terms of Use part of this license, apply to copying and distributing Project Gutenberg™ electronic works to protect the PROJECT GUTENBERG™ concept and trademark. Project Gutenberg is a registered trademark, and may not be used if you charge for an eBook, except by following the terms of the trademark license, including paying royalties for use of the Project Gutenberg trademark. If you do not charge anything for copies of this eBook, complying with the trademark license is very easy. You may use this eBook for nearly any purpose such as creation of derivative works, reports, performances and research. Project Gutenberg eBooks may be modified and printed and given away--you may do practically ANYTHING in the United States with eBooks not protected by U.S. copyright law. Redistribution is subject to the trademark license, especially commercial redistribution.

START: FULL LICENSE
THE FULL PROJECT GUTENBERG LICENSE
PLEASE READ THIS BEFORE YOU DISTRIBUTE OR USE THIS WORK

To protect the Project Gutenberg™ mission of promoting the free distribution of electronic works, by using or distributing this work (or any other work associated in any way with the phrase "Project Gutenberg"), you agree to comply with all the terms of the Full Project Gutenberg™ License available with this file or online at www.gutenberg.org/license.

Section 1. General Terms of Use and Redistributing Project Gutenberg™ electronic works

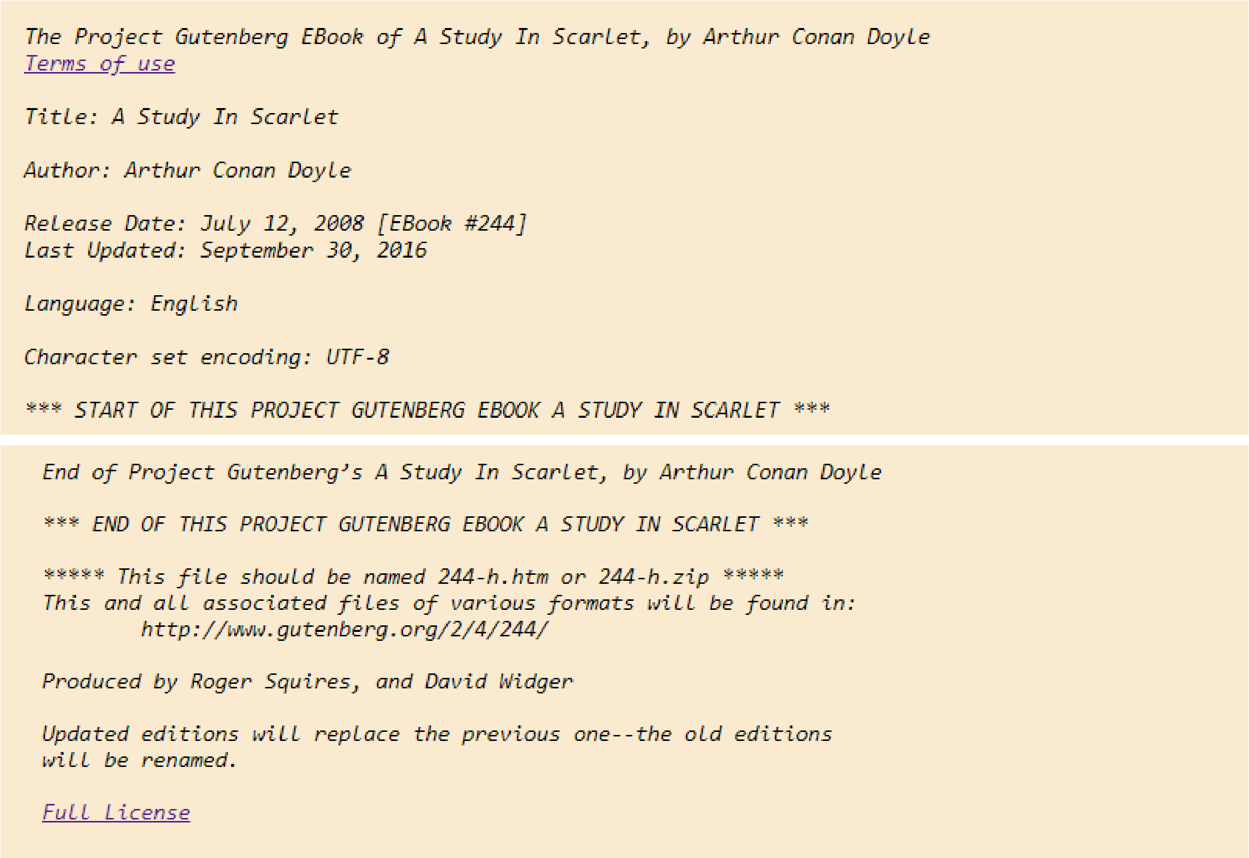
1.A. By reading or using any part of this Project Gutenberg™ electronic work, you indicate that you have read, understand, agree to and accept all the terms of this license and intellectual property (trademark/copyright) agreement. If you do not agree to abide by all the terms of this agreement, you must cease using and return or destroy all copies of Project Gutenberg™ electronic works in your possession. If you paid a fee for obtaining a copy of or access to a Project Gutenberg™ electronic work and you do not agree to be bound by the terms of this agreement, you may obtain a refund from the person or entity to whom you paid the fee as set forth in paragraph 1.E.8.

1.B. "Project Gutenberg" is a registered trademark. It may only be used on or associated in any way with an electronic work by people who agree to be bound by the terms of this agreement. There are a few things that you can do with most Project Gutenberg™ electronic works even without complying with the full terms of this agreement. See paragraph 1.C below. There are a lot of things you can do with Project Gutenberg™ electronic works if you follow the terms of this agreement and help preserve free future access to Project Gutenberg™ electronic works. See paragraph 1.E below.

1.C. The Project Gutenberg Literary Archive Foundation ("the Foundation" or PGLAF), owns a compilation copyright in the collection of Project Gutenberg™ electronic works. Nearly all the individual works in the collection are in the public domain in the United States. If an individual work is unprotected by copyright law in the United States and you are located in the United States, we do not claim a right to prevent you from copying, distributing, performing, displaying or creating derivative works based on the work as long as all references to Project Gutenberg are removed. Of course, we hope that you will support the Project Gutenberg™ mission of promoting free access to electronic works by freely sharing Project Gutenberg™ works in

Figure 4: Bottom boilerplate HTML file text

Figure 5 is a screenshot of the and bottom of an eBook after the program is run. Looking at figure 5, the common text is successfully being replaced with hypertext links to the extracted text. The top boilerplate link is named "Terms of use" and the bottom boilerplate link is named "Full License."

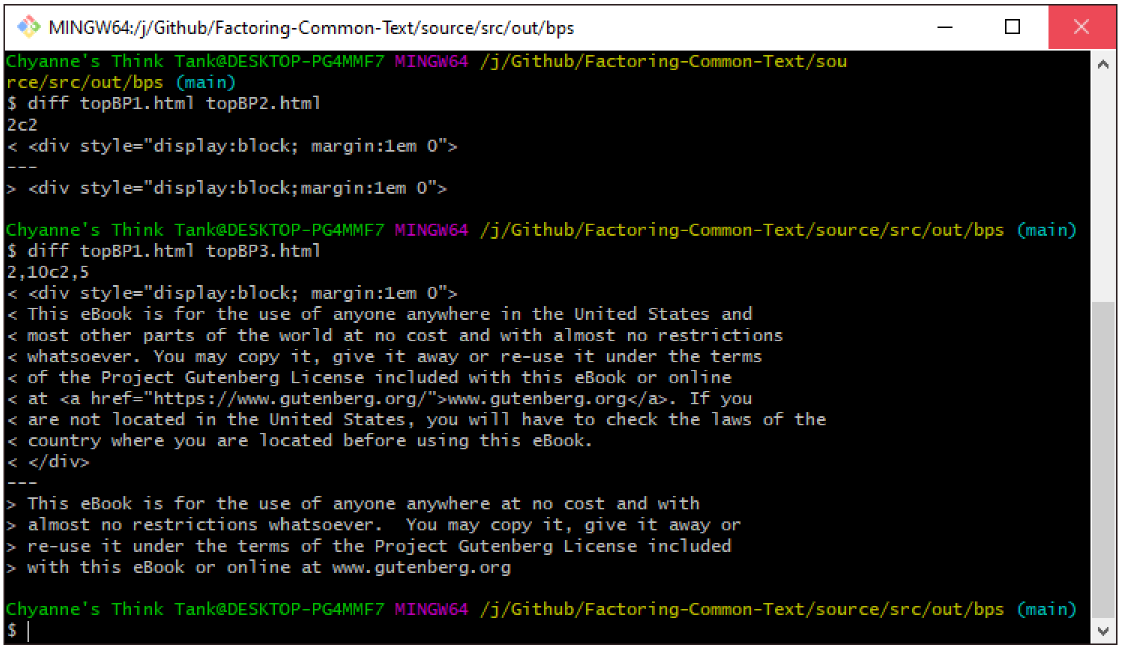


```
The Project Gutenberg EBook of A Study In Scarlet, by Arthur Conan Doyle
Terms of use

Title: A Study In Scarlet

Author: Arthur Conan Doyle

Release Date: July 12, 2008 [EBook #244]
Last Updated: September 30, 2016

Language: English

Character set encoding: UTF-8

*** START OF THIS PROJECT GUTENBERG EBOOK A STUDY IN SCARLET ***
```

```
End of Project Gutenberg's A Study In Scarlet, by Arthur Conan Doyle

*** END OF THIS PROJECT GUTENBERG EBOOK A STUDY IN SCARLET ***

***** This file should be named 244-h.htm or 244-h.zip *****
This and all associated files of various formats will be found in:
        http://www.gutenberg.org/2/4/244/

Produced by Roger Squires, and David Widger

Updated editions will replace the previous one--the old editions
will be renamed.

Full License
```

Figure 5: Top and bottom of eBook after program is run

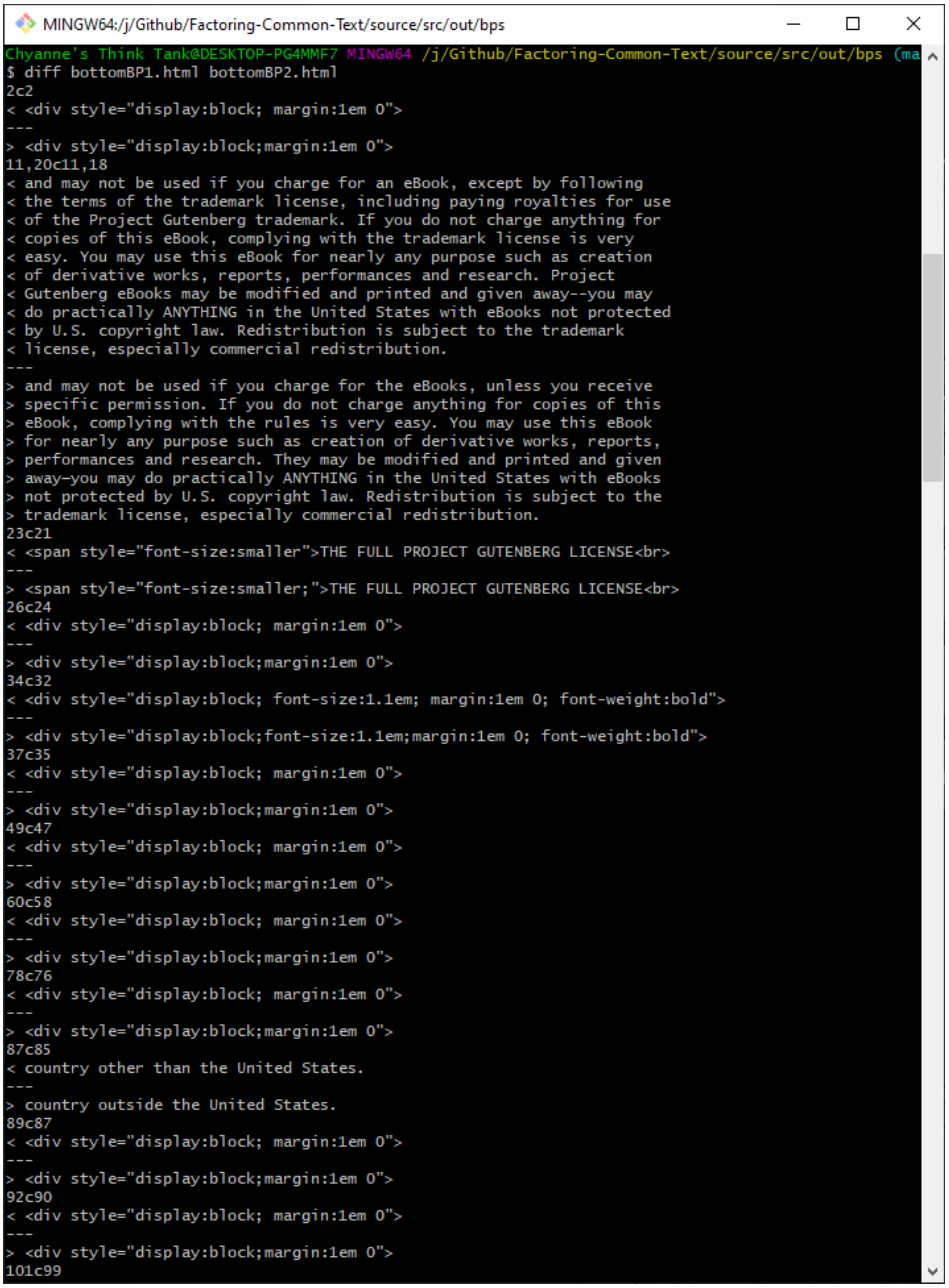
Figure 6 below shows the output of the 'diff' command run for the 3 different top boilerplates. This command outputs the differences between two files and is useful in showing if the boilerplates are truly different. Looking at the diff outputs, the differences between top boilerplates 1 and 2 are minimal. The only difference is an extra whitespace character within an HTML tag. Therefore, the user will not see any difference between the first two boilerplates. The differences between top boilerplates 1 and 3 are much more noticeable. The main text shown to the user has changed. This proves that the program successfully identifies different top boilerplates.



```
Chyanne's Think Tank@DESKTOP-PG4MMF7 MINGW64 /j/Github/Factoring-Common-Text/sou
rce/src/out/bps (main)
$ diff topBP1.html topBP2.html
2c2
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">

Chyanne's Think Tank@DESKTOP-PG4MMF7 MINGW64 /j/Github/Factoring-Common-Text/source/src/out/bps (main)
$ diff topBP1.html topBP3.html
2,10c2,5
< <div style="display:block; margin:1em 0">
< This eBook is for the use of anyone anywhere in the United States and
< most other parts of the world at no cost and with almost no restrictions
< whatsoever. You may copy it, give it away or re-use it under the terms
< of the Project Gutenberg License included with this eBook or online
< at <a href="https://www.gutenberg.org/">www.gutenberg.org</a>. If you
< are not located in the United States, you will have to check the laws of the
< country where you are located before using this eBook.
< </div>
---
> This eBook is for the use of anyone anywhere at no cost and with
> almost no restrictions whatsoever.  You may copy it, give it away or
> re-use it under the terms of the Project Gutenberg License included
> with this eBook or online at www.gutenberg.org

Chyanne's Think Tank@DESKTOP-PG4MMF7 MINGW64 /j/Github/Factoring-Common-Text/source/src/out/bps (main)
$ |
```

Figure 6: diff for top boilerplates

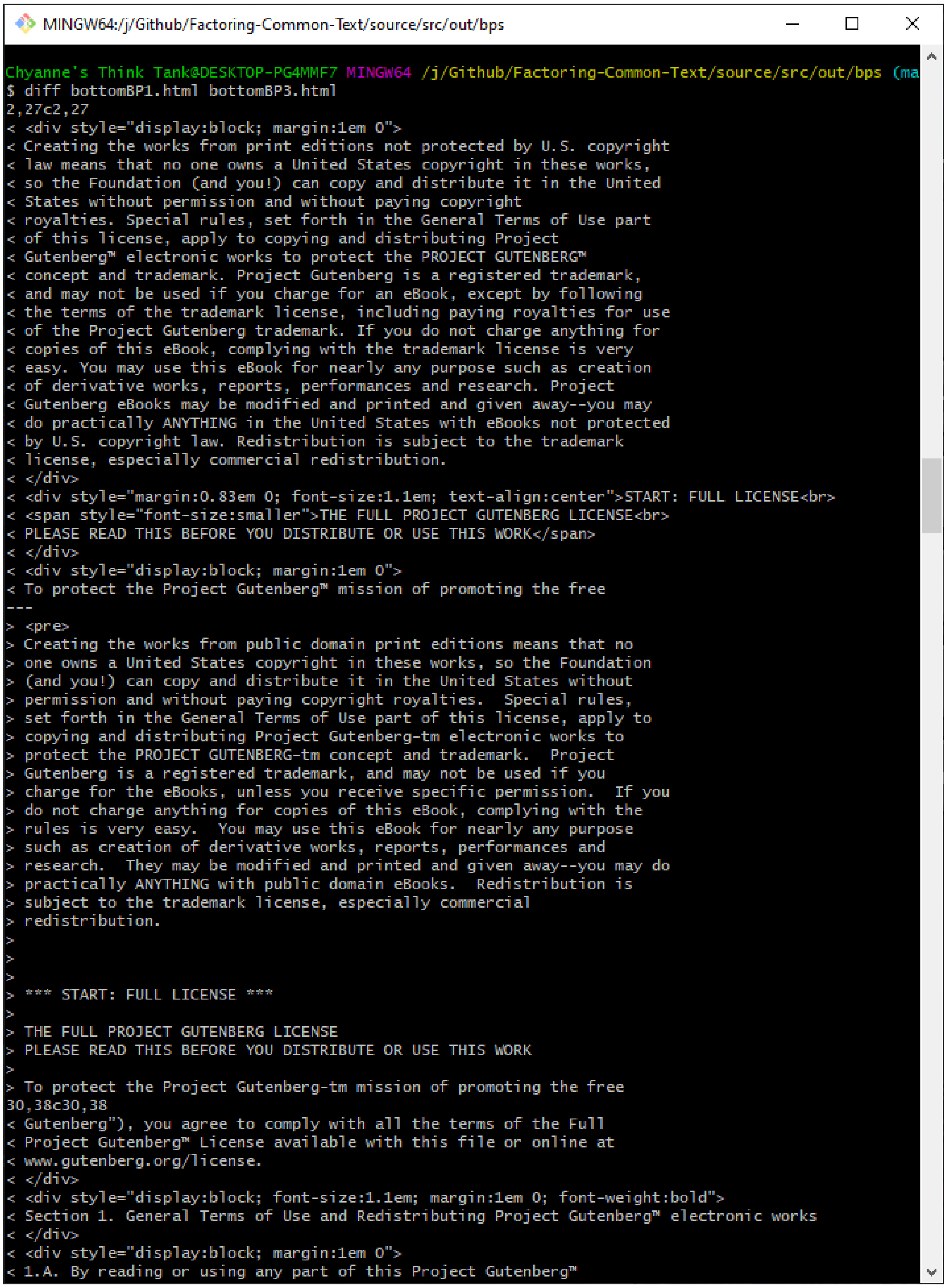
Figure 7 below shows the output of the ‘diff’ command run for the bottom boilerplates 1 and 2. Most of the differences between boilerplates 1 and 2 are whitespace differences between HTML tags. However, there is one text difference between the two boilerplates. This means the program correctly identified two different boilerplates.



```
Chyanne's Think Tank@DESKTOP-PG4MMF7 MINGW64 /j/Github/Factoring-Common-Text/source/src/out/bps (ma
$ diff bottomBP1.html bottomBP2.html
2c2
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
11,20c11,18
< and may not be used if you charge for an eBook, except by following
< the terms of the trademark license, including paying royalties for use
< of the Project Gutenberg trademark. If you do not charge anything for
< copies of this eBook, complying with the trademark license is very
< easy. You may use this eBook for nearly any purpose such as creation
< of derivative works, reports, performances and research. Project
< Gutenberg eBooks may be modified and printed and given away--you may
< do practically ANYTHING in the United States with eBooks not protected
< by U.S. copyright law. Redistribution is subject to the trademark
< license, especially commercial redistribution.
---
> and may not be used if you charge for the eBooks, unless you receive
> specific permission. If you do not charge anything for copies of this
> eBook, complying with the rules is very easy. You may use this eBook
> for nearly any purpose such as creation of derivative works, reports,
> performances and research. They may be modified and printed and given
> away—you may do practically ANYTHING in the United States with eBooks
> not protected by U.S. copyright law. Redistribution is subject to the
> trademark license, especially commercial redistribution.
23c21
< <span style="font-size:smaller">THE FULL PROJECT GUTENBERG LICENSE<br>
---
> <span style="font-size:smaller;">THE FULL PROJECT GUTENBERG LICENSE<br>
26c24
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
34c32
< <div style="display:block; font-size:1.1em; margin:1em 0; font-weight:bold">
---
> <div style="display:block;font-size:1.1em;margin:1em 0; font-weight:bold">
37c35
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
49c47
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
60c58
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
78c76
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
87c85
< country other than the United States.
---
> country outside the United States.
89c87
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
92c90
< <div style="display:block; margin:1em 0">
---
> <div style="display:block;margin:1em 0">
101c99
```

Figure 7: diff for bottom boilerplates 1 and 2

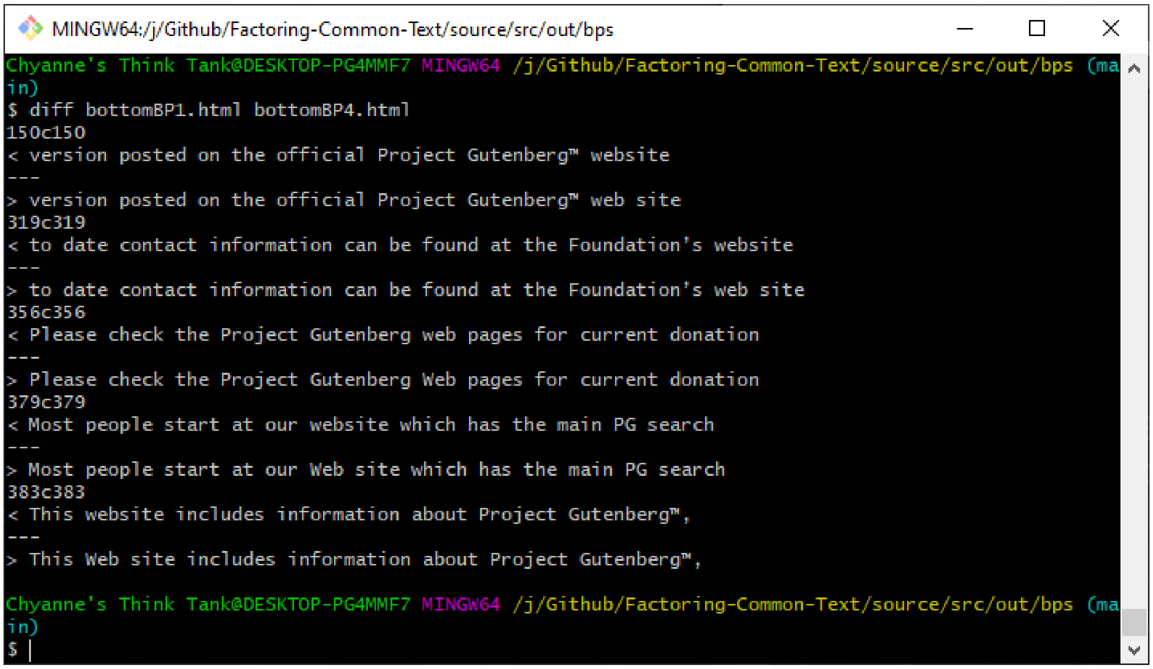
Figure 8 below shows the output of the ‘diff’ command run for bottom boilerplates 1 and 3. Boilerplate 3 is the only boilerplate that uses <pre> HTML tag while all other boilerplates use <div> tags. This makes the two boilerplates very different. Therefore, this further proves that the program correctly identifies different boilerplates.



```
Chyanne's Think Tank@DESKTOP-PG4MMF7 MINGW64 /j/Github/Factoring-Common-Text/source/src/out/bps (ma
$ diff bottomBP1.html bottomBP3.html
2,27c2,27
< <div style="display:block; margin:1em 0">
< Creating the works from print editions not protected by U.S. copyright
< law means that no one owns a United States copyright in these works,
< so the Foundation (and you!) can copy and distribute it in the United
< States without permission and without paying copyright
< royalties. Special rules, set forth in the General Terms of Use part
< of this license, apply to copying and distributing Project
< Gutenberg™ electronic works to protect the PROJECT GUTENBERG™
< concept and trademark. Project Gutenberg is a registered trademark,
< and may not be used if you charge for an eBook, except by following
< the terms of the trademark license, including paying royalties for use
< of the Project Gutenberg trademark. If you do not charge anything for
< copies of this eBook, complying with the trademark license is very
< easy. You may use this eBook for nearly any purpose such as creation
< of derivative works, reports, performances and research. Project
< Gutenberg eBooks may be modified and printed and given away--you may
< do practically ANYTHING in the United States with eBooks not protected
< by U.S. copyright law. Redistribution is subject to the trademark
< license, especially commercial redistribution.
< </div>
< <div style="margin:0.83em 0; font-size:1.1em; text-align:center">START: FULL LICENSE<br>
< <span style="font-size:smaller">THE FULL PROJECT GUTENBERG LICENSE<br>
< PLEASE READ THIS BEFORE YOU DISTRIBUTE OR USE THIS WORK</span>
< </div>
< <div style="display:block; margin:1em 0">
< To protect the Project Gutenberg™ mission of promoting the free
---
> <pre>
> Creating the works from public domain print editions means that no
> one owns a United States copyright in these works, so the Foundation
> (and you!) can copy and distribute it in the United States without
> permission and without paying copyright royalties.  Special rules,
> set forth in the General Terms of Use part of this license, apply to
> copying and distributing Project Gutenberg-tm electronic works to
> protect the PROJECT GUTENBERG-tm concept and trademark.  Project
> Gutenberg is a registered trademark, and may not be used if you
> charge for the eBooks, unless you receive specific permission.  If you
> do not charge anything for copies of this eBook, complying with the
> rules is very easy.  You may use this eBook for nearly any purpose
> such as creation of derivative works, reports, performances and
> research.  They may be modified and printed and given away--you may do
> practically ANYTHING with public domain eBooks.  Redistribution is
> subject to the trademark license, especially commercial
> redistribution.
>
>
>
> *** START: FULL LICENSE ***
>
> THE FULL PROJECT GUTENBERG LICENSE
> PLEASE READ THIS BEFORE YOU DISTRIBUTE OR USE THIS WORK
>
> To protect the Project Gutenberg-tm mission of promoting the free
30,38c30,38
< Gutenberg"), you agree to comply with all the terms of the Full
< Project Gutenberg™ License available with this file or online at
< www.gutenberg.org/license.
< </div>
< <div style="display:block; font-size:1.1em; margin:1em 0; font-weight:bold">
< Section 1. General Terms of Use and Redistributing Project Gutenberg™ electronic works
< </div>
< <div style="display:block; margin:1em 0">
< 1.A. By reading or using any part of this Project Gutenberg™
```

Figure 8: diff for bottom boilerplates 1 and 3

Figure 9 below shows the output of the ‘diff’ command run for bottom boilerplates 1 and 4. Similar to the diff for boilerplates 1 and 2, the differences are minimal. Looking at the output, the difference between the two is the spelling of the word website. Boilerplate 1 spells website as one word, “website.” While boilerplate 4 spells it as two words, “web site.” This is a small difference and the fact that the program correctly identified the difference proves the algorithm is very precise when comparing common text.



```
Chyanne's Think Tank@DESKTOP-PG4MMF7 MINGW64 /j/Github/Factoring-Common-Text/source/src/out/bps (ma
in)
$ diff bottomBP1.html bottomBP4.html
150c150
< version posted on the official Project Gutenberg™ website
---
> version posted on the official Project Gutenberg™ web site
319c319
< to date contact information can be found at the Foundation’s website
---
> to date contact information can be found at the Foundation’s web site
356c356
< Please check the Project Gutenberg web pages for current donation
---
> Please check the Project Gutenberg Web pages for current donation
379c379
< Most people start at our website which has the main PG search
---
> Most people start at our Web site which has the main PG search
383c383
< This website includes information about Project Gutenberg™,
---
> This Web site includes information about Project Gutenberg™,

Chyanne's Think Tank@DESKTOP-PG4MMF7 MINGW64 /j/Github/Factoring-Common-Text/source/src/out/bps (ma
in)
$ |
```

Figure 9: diff for bottom boilerplates 1 and 4

Conclusion

Factoring common text in only the beginning and end of a file presents a unique issue. It is very hard to be able to generally define where the top and bottom boilerplates are located. This led to the team using HTML tags or lines of text that are always included in the eBooks. Therefore, if Project Gutenberg were to change the format of the eBooks, this algorithm would stop working as intended. There are already two different cases within the algorithm written as eBooks either use <div> or <pre> tags to denote the top boilerplate. This flaw makes the code unusable for any other website.

Furthermore, the text extraction skips empty lines and copies everything else. This means that small whitespace differences will cause extra boilerplate files to be created. This is seen in the difference between top boilerplates 1 and 2. The only difference is one space in an HTML <div> tag. If all whitespace was skipped, the main text would not have any spaces either, making it unreadable for the user. An improvement could be to ignore differences that are within brackets or skip adding them all together. This would make the HTML tags irrelevant to the common text extraction, leaving only the text that is displayed. However, this could lead to improper formatting of the boilerplates as all formatting information is held within the HTML tags. Moreover, looking at the difference between bottom boilerplates 1 and 4, the only difference is the spelling of the word ‘website.’ To help reduce redundancy even further, the algorithm can be modified so that if the difference between two boilerplates is minimal, they will be treated as the same boilerplate. This has the potential to solve both the HTML tag issue and the small spelling differences that may exist across all the boilerplates.

Finally, the team's algorithm is not fully optimized. For example, the existing boilerplate check can be sped up by using a hashmap. This would make the time it takes to see if a boilerplate already exists O(1), cutting the runtime by a considerable amount. Despite this, the current algorithm performs well. The algorithm is O(n) where n is the total number of lines in all the files, and all tested files output the correct eBook with the hypertext links inserted.