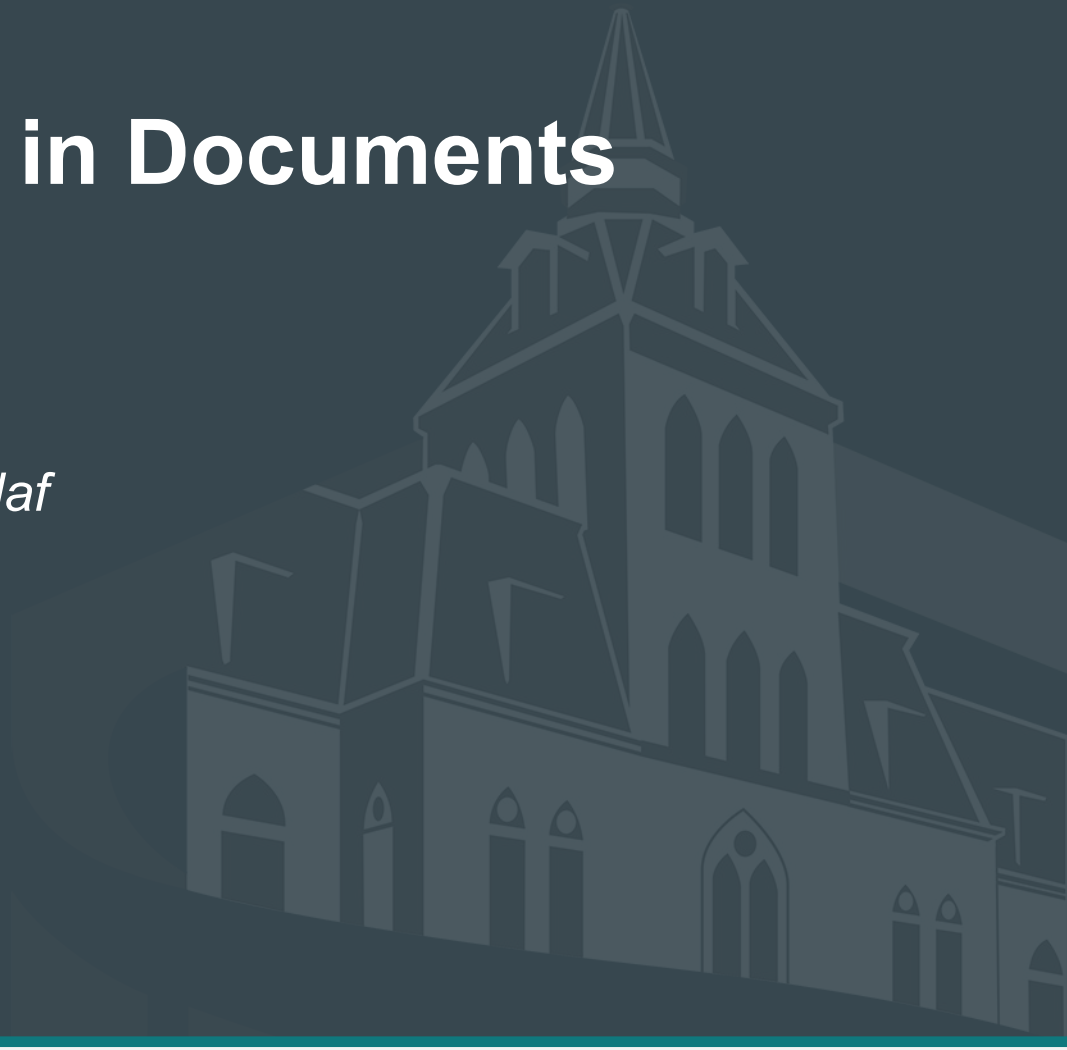
STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Factoring Common Text in Documents

Trevor Dawideit

Gousemoodhin Nadaf



I pledge my honor that I have abided by the Stevens Honor System.

Agenda

- Introduction
- Problem
- Solution
- Results
- Demo
- Improvements
- Q&A

Introduction

Problem

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.

*** 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

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 ***

A STUDY IN SCARLET.

By A. Conan Doyle

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/

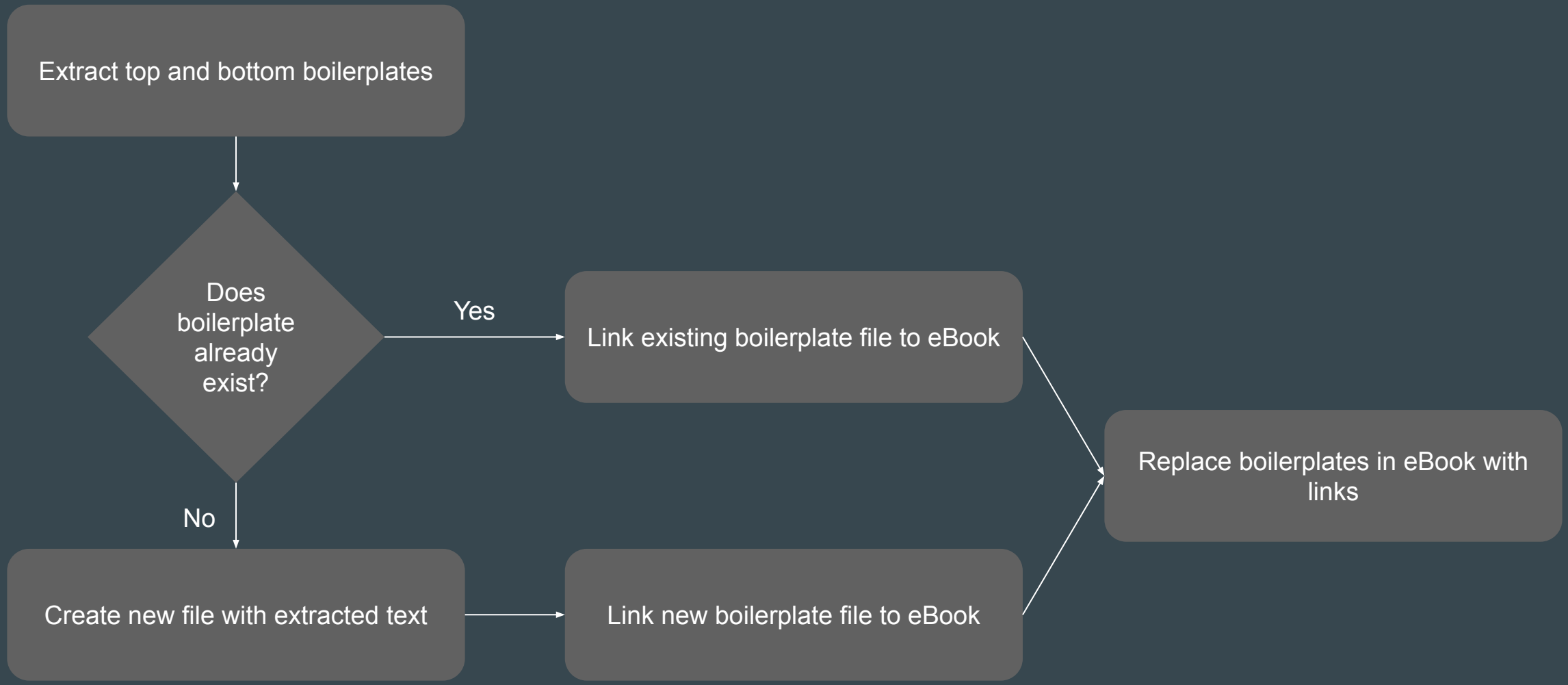
Produced by Roger Squires, and David Widger

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 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 ***

Solution


Extract top and bottom boilerplates
Does boilerplate already exist?
Yes
No
Link existing boilerplate file to eBook
Create new file with extracted text
Link new boilerplate file to eBook
Replace boilerplates in eBook with links

Solution (Extract top and bottom boilerplates)

Create 4 vectors:

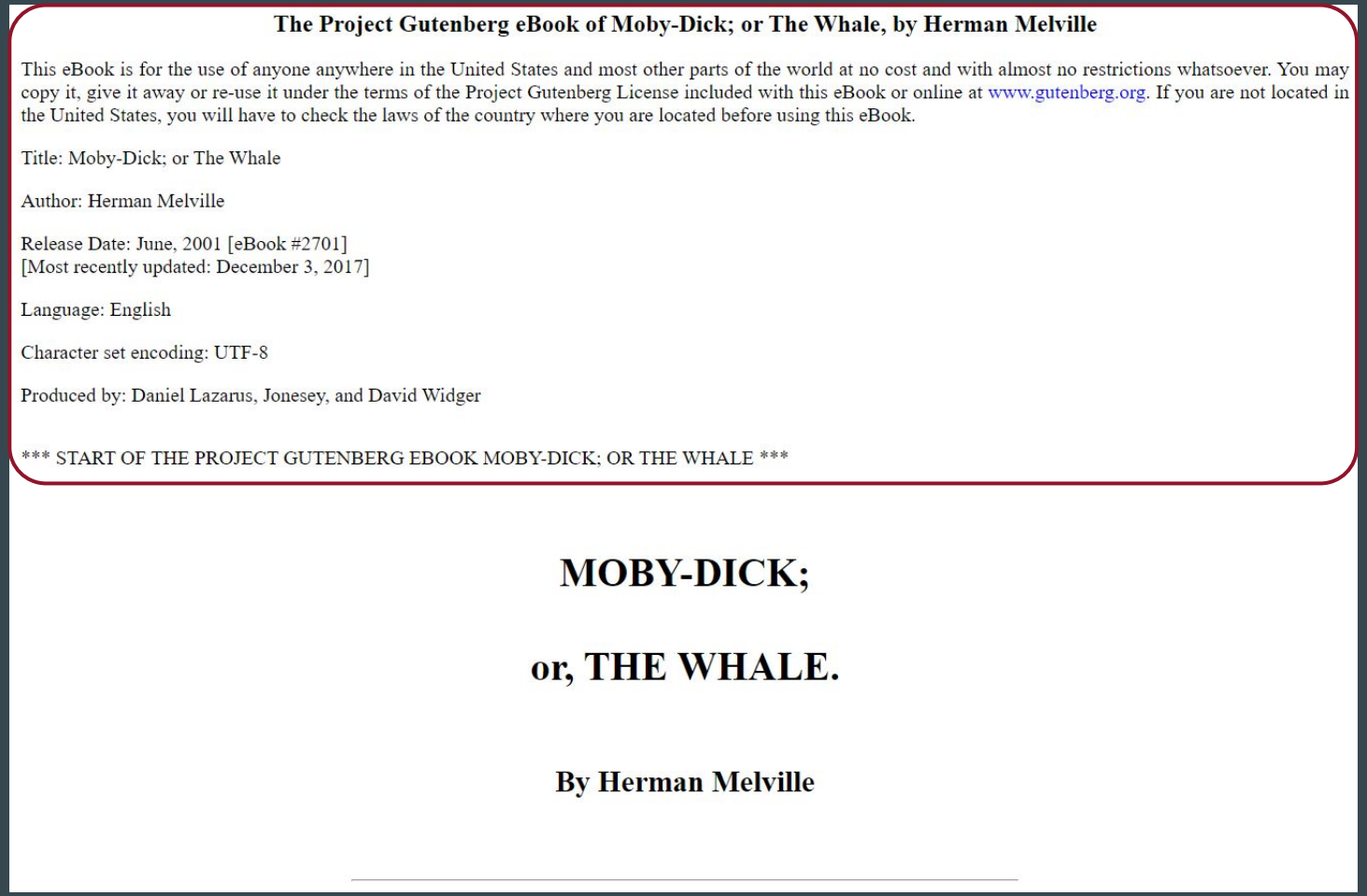
topFileContent

topBoilerplate

mainFileContent

bottomBoilerplate

Solution (Extract Top Boilerplate)



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

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

Solution (Extract Top Boilerplate cont.)

```
<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>
```

Solution (Extract Top Boilerplate cont.)

```
<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>
```

Solution (Extract Bottom Boilerplate)

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

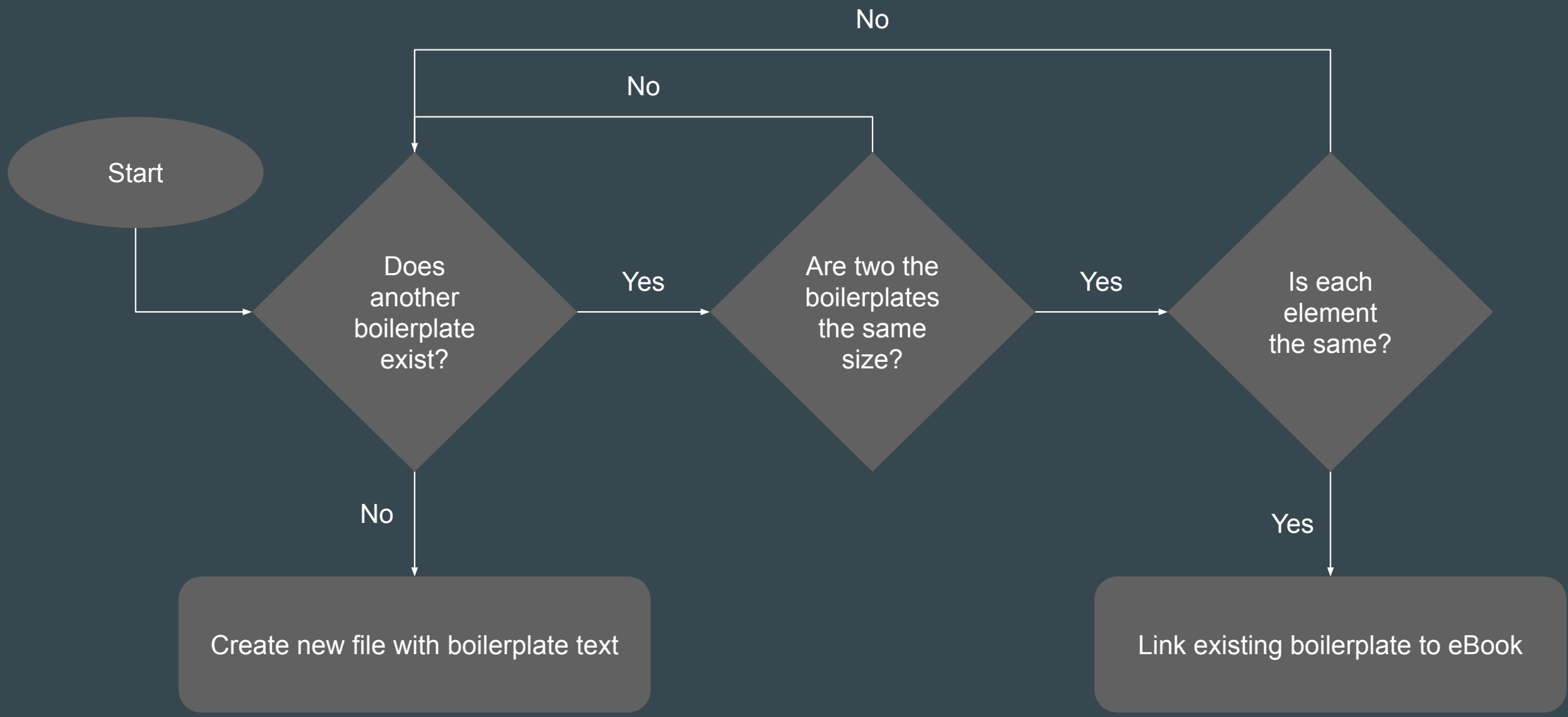
It so chanced, that after the Parsee's disappearance, I was he whom the Fates ordained to take the place of Ahab's bowsman, when that bowsman assumed the vacant post; the same, who, when on the last day the three men were tossed from out of the rocking boat, was dropped astern. So, floating on the margin 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.

Solution (Extract Bottom Boilerplate cont.)

```
<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
```

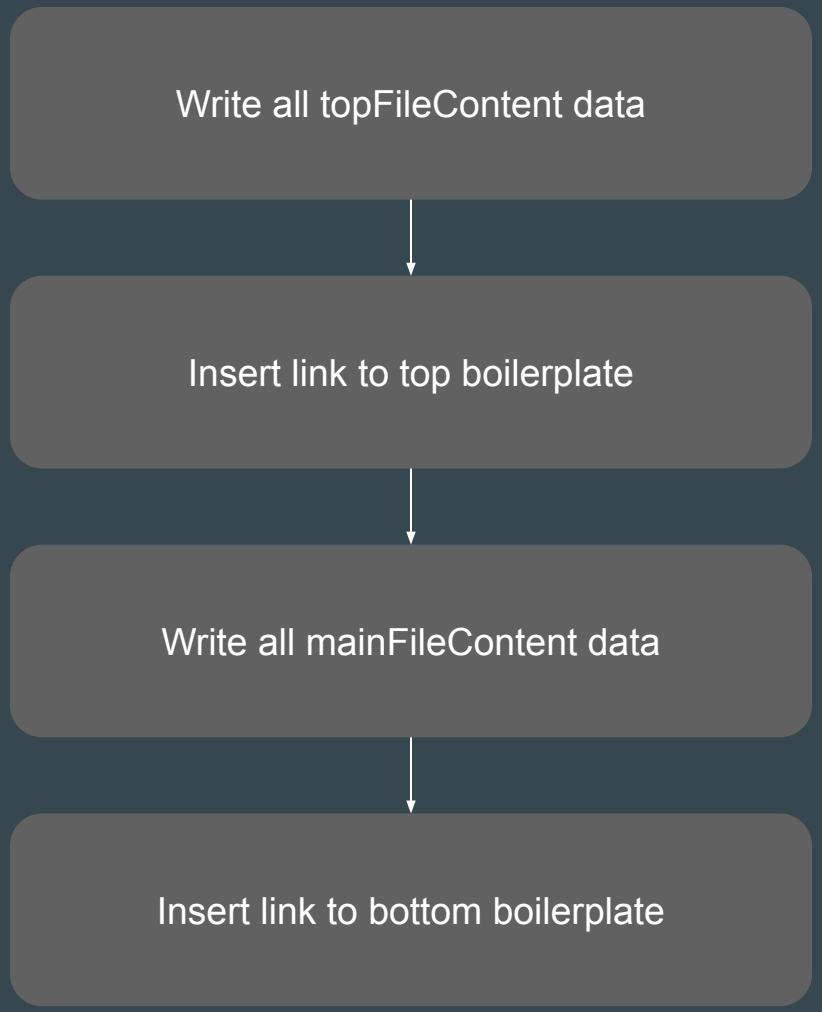
```
<pre xml:space="preserve">



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
```

Solution (Check if boilerplate already exists)



Solution (Replace text with link)

Create new file:



Results (Complexity)

- Outer loop = O(F), F = Number of files
- Inner loop = O(M), M = Number of lines in file
- Boilerplate comparison = O(K), K = Length of boilerplate * Number of existing boilerplates
  - This comparison is run for both top and bottom boilerplates.
- Output file creation = O(L), L = Length of topFileContent + length of mainFileContent

Total Complexity:

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

O(n), n = total number of lines in all files

Live Demo

Improvements

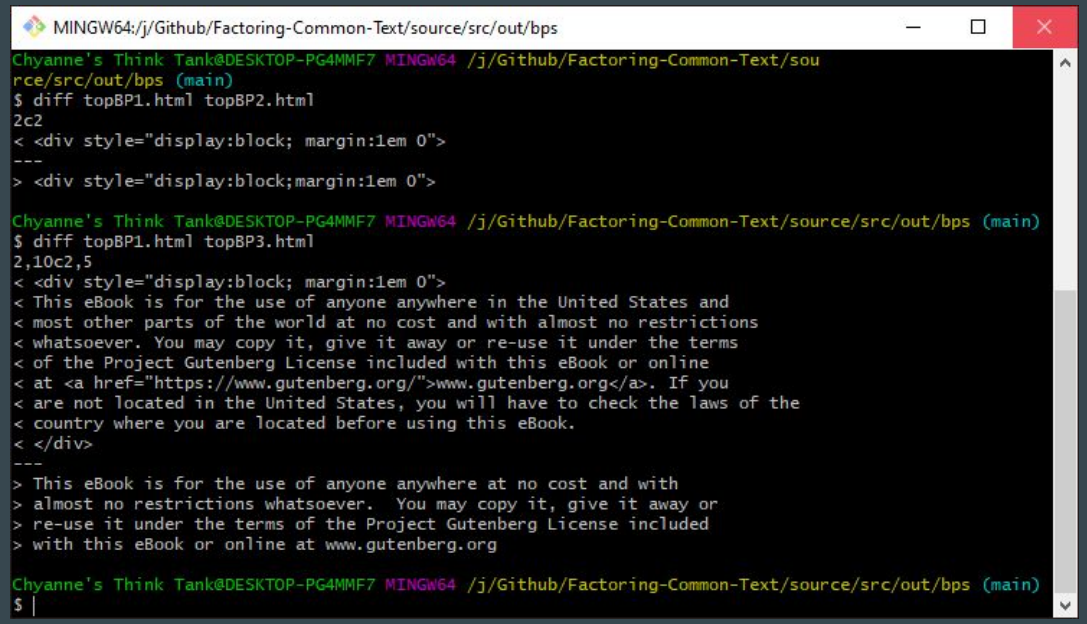
Boilerplate Detection

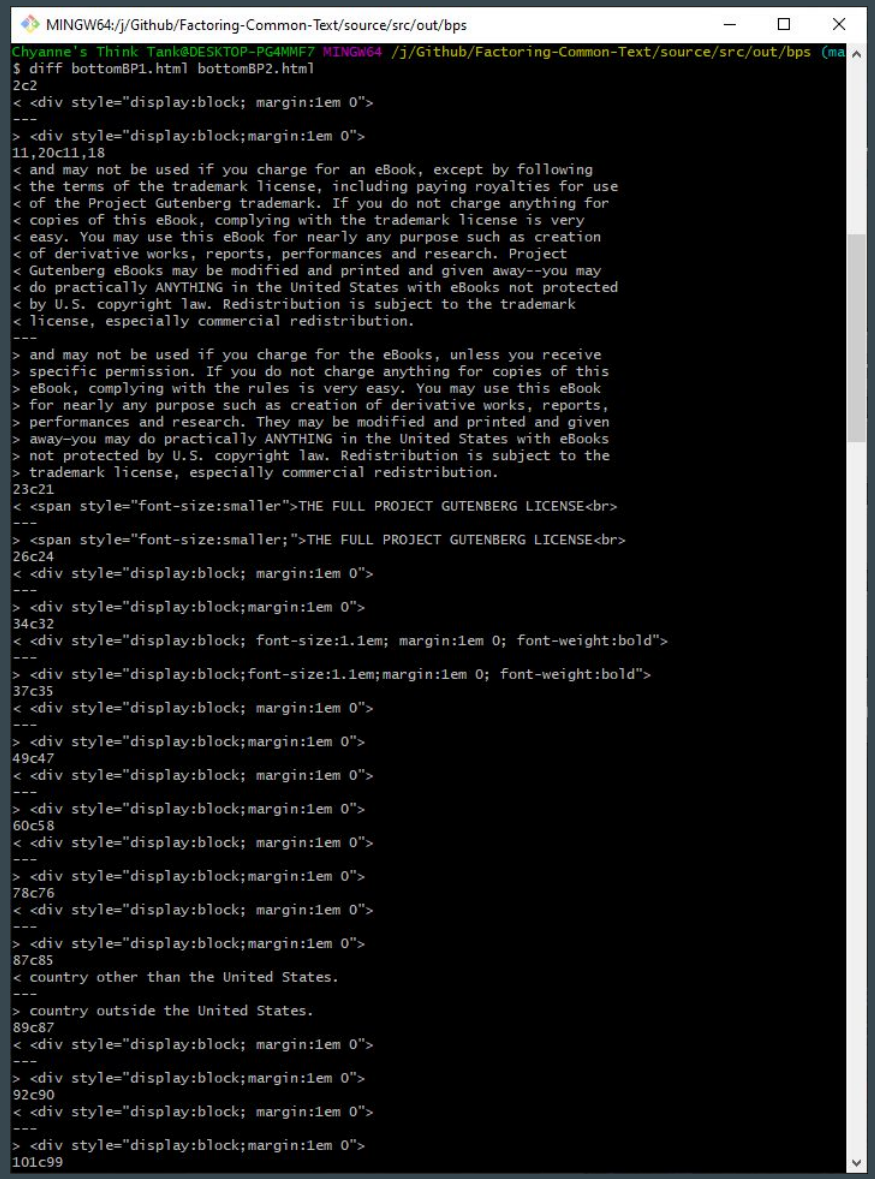
```
<pre xml:space="preserve">
The Project Gutenberg EBook of A Study In Scarlet, by Arthur Conan Doyle

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>
```

```
<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: Anthem</div>
<div style="display:block; margin-top:1em; margin-bottom:1em; margin-left:2em; text-indent:-2em">Author: Ayn Rand</div>
<div style="display:block;margin:1em 0">Release Date: March, 1998 [eBook #1250]<br>
[Most recently updated: January 21, 2021]</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: An anonymous group of volunteers, and David Widger</div>
<div style="margin-top:2em;margin-bottom:4em">*** START OF THE PROJECT GUTENBERG EBOOK ANTHEM ***</div>
```

Improvements

Text Extraction





Questions?

STEVENS

INSTITUTE of TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu