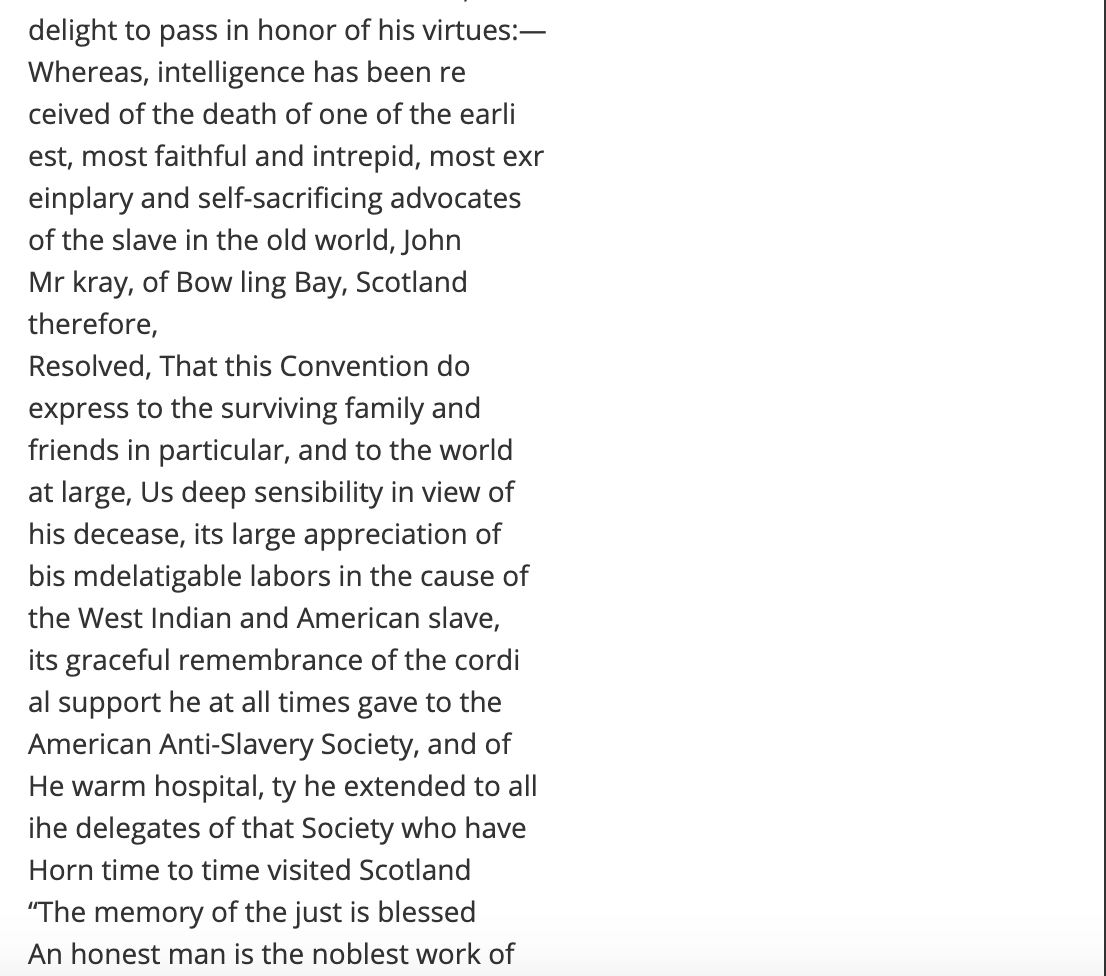
## Step 1: Obtain plain-text file

* Copy and paste the OCR directly from the Library of Congress, or download as a batch if that becomes an option. For reasons unknown, *The North Star* cannot be bulk downloaded at this moment. If that becomes available, the batch numbers were dlc\_bluebell\_ver02 and dlc\_aster\_ver01.
* Copy and pasted output should appear similar to this:



* After copying and pasting all 4 pages into any word processor, ensure that the file is saved as plain text. All files were saved as sn84026365/(date)/ed-1.txt.

## Step 2: Remove punctuation and make lowercase

* This process can be accomplished several ways, but this project specifically used Lexos online (<http://lexos.wheatoncollege.edu/upload>). Their interface allowed for files to be bulk uploaded, edit all files to remove punctuation and make all lowercase, and then bulk downloaded.
* Save the new files in a folder.

## Step 3: Place each word on one line

* Run python script “Write\_Files\_for\_Linebreak\_Removal.ipynb”, which generates new files where each word is on its own line. For words on the linebreak, the linebreak is replaced with a hyphen so that these specific words can be identified in the next step of the process.
* Store these new files in folder Lines\_Broken

## Step 4: Fix linebreak

* Include in the Lines\_Broken folder a copy of the dictionary used for next step “en\_US.txt”
* Run python script “LineRepairer.ipynb”, which looks at hyphenated words and compares them to the dictionary “en\_US.txt”. Key piece of code:
* This takes significant time, and is possibly related to the size of the english language dictionary that is used.
* Code will output files that have removed hyphens and linebreaks, allowing for files to be used in text analysis software.
* Once the script has been completed, save the output files in a new folder. The script also generates a copy of the “en\_US.txt” dictionary. Be sure to delete this, though because of the structure of the code this file should be the same as the original dictionary.

## Step 5: Remove stop words

* For this project, stop words were removed using the --remove-stop-words function in MALLET. This can also be accomplished in Lexos Online by uploading a stop words list.
* As part of this process, I created an additional document that included three extra stop words: thy, thou, and thee.