CSC733 Project Fall 2024

This page is under revision, so much of it may change!!!!!!!!!!!!!

Preliminary:

Choose a book from any source.

Suggestions: [Project Gutenberg](https://www.gutenberg.org/) [Google Books](https://books.google.com/)

Follow steps in Project deliverables #1a

The rest

Choose a book, story, or article. It must be a minimum of 30 pages.

If you find an MS word document you will save a few steps.

You may NOT use [Project Gutenberg](https://www.gutenberg.org/)

[Internet Archive](https://archive.org/) is an excellent source.

You may use a printed article or short story and take pictures of the pages with your phone.

If you choose an image, \*.png or \*.img, convert it into a \*.txt file as demonstrated in class.

Here is a trick which will make your work a lot easier.

The Gutenberg corpus has a lot of NLTK functions already attached to it.

To take advantage of these, you may add your text to your copy of Gutenberg by following these steps (the example is my project):

import nltk

from nltk.corpus import PlaintextCorpusReader

# Define the directory where your custom corpus is stored

corpus\_root = 'C:Users\Emile\AppData\Local\Programs\Python\Python312\MyTexts'

# Create a PlaintextCorpusReader object

custom\_corpus = PlaintextCorpusReader(corpus\_root, '.\*\.txt') # Load all .txt files in the directory

# Access the files in your custom corpus

file\_ids = custom\_corpus.fileids()

print(file\_ids) # This will print the names of the text files in your corpus

# Read the content of a specific file (your book)

book\_text = custom\_corpus.raw('my\_book.txt')

print(book\_text[:500]) # Print the first 500 characters of the book