CSC615M - G01 Project 1: Text Cleaning | Carandang, Sityar, Veracruz

Team: The Three Stooges

Name	ID Number
Carandang, Matthew Ryan	12206040
Sityar, Lester Anthony Jr. Sityar	12292639
Veracruz, Sean Riley	12209392

1. Data Selection

Tagalog	English	Bikolano
Ang Biblia Tagalog 1905	Holy Bible - American King James Version	Marahay na Baretea Biblia Central Bikol Bible New Testament
https://bible4u.net/static/bible_files/pdf/ADB1905.pdf	https://openbible.com/pdfs/akjv.pdf	https://bibliamundi.com/wp-content/uploads/2023/09/Central-Bicolano-All-Bible.pdf
Source Text Size: 108,158 Cleaned Size: 107,877	Source Text Size: 109,998 Words Cleaned Size: 104,368 Words	Source Text Size: 102,310 Cleaned Size: 102,389

To fulfill the 100,000-word requirement of the project, the books of ***Genesis, Exodus****, and* ***Numbers*** *from the Old Testament were selected. For Bikolano, Additional verses from Leviticus were added to meet the required word count.*

2. Processing

a. Tagalog

i. Pre-processing (Cleaning)

Step	Search Pattern	Replace With	Explanation
1	`(?i)\b(Genesis|Exodo|Mga\s+Bilang)\b`		Remove book/chapter titles
2	`(?<=\s)\d{1,4}\.?(?=\s)`		Remove stray numbers (not verse numbers)
3	`(\d{1,3})([A-Z])`	`\r\n\1\2`	Insert newline before verse numbers

4	`(\w)\r?\n(\w)`	*`\1 \2`*	Join broken words split by line breaks
5	`([,;:])\r?\n\s*`	*`\1`*	Join lines split after punctuation
6	`\r?\n`		Remove leftover line breaks

ii. Segmentation (Verse and Sentences)

Step	Search Pattern	Replace With	Explanation
7	`\b\d{1,3}(?=[A-Z])`		Remove verse numbers stuck to the start of words
8	`([.!?])\s+`	*`\1\r\n`*	Newline after ., ?, or ! to split into sentences

b. English

i. Pre-processing (Cleaning)

Step	Search Pattern	Replace With	Explanation
1	`\b(?:[1-3]?\s?[A-Z][a-z]+(?:\s[A-Z][a-z]+)*)\s\d+\b`	*(empty)*	Remove book headers (e.g., "Genesis 1")
2	`([A-Za-z ]+)\s+\(([^()]+)\)`	*(empty)*	Remove section headers with scripture references (e.g., The Creation (John 1:1—5))
3	`AKJV  \[Online\]`	*(empty)*	Remove "AKJV [Online]" headers
4	`[ ]{2,}`	*whitespace*	Collapse multiple spaces into one
5	*(note: NNBSP)*	*(empty)*	Insert line break before verse numbers followed by a non-breaking space

ii. Segmentation (Verse & Sentence)

Step	Search Pattern	Replace With	Explanation
6	`(?<!\n)(?<!^)([0-9]+)(?=[A-Z])`	`\n\1`	Insert line break before verse numbers directly followed by text
7	`(?m)^\d+(?=[A-Z])`	*(empty)*	Remove verse numbers at the start of a line
8	`^\s*\n`	*(empty)*	Remove empty lines

c. Bikolano

i. Pre-processing (Cleaning)

Step	Search Pattern	Replace With	Explanation
1	`\s+`	*whitespace*	Removes multiple spaces and replaces them with a single space. Uniform formatting

ii. Segmentation (Verse & Sentence)

Step	Search Pattern	Replace With	Explanation
2	`\bGenesis\s+(\d+)`	*`\nGenesis $1:\n`*	Fixes the formatting of the Book of Genesis chapters and separates them from the verses. (Was done for Readability purposes)
3	`\bExodo\s+(\d+)`	*`\nExodo $1:\n`*	Fixes the formatting of the Book of Exodus chapters and separates them from the verses. (Was done for Readability purposes)
4	`\bMga\sBilang\s+(\d+)`	*`\nMga Bilang $1:\n`*	Fixes the formatting of the Book of Numbers chapters and separates them from the verses. (Was done for Readability purposes)
5	`\bLevitico\s+(\d+)`	*`\nLevitico $1:\n`*	Fixes the formatting of the Book of Leviticus chapters and separates them from the verses. (Was done for Readability purposes)
6	`(\d+)(?=[A-Za-z“"])`	*`\n[$1]`*	Used to separate the verse numbers that were attached to the start of sentences
7	`(?<=[.?!])\s+(?=[A-Z“])`	*`\n`*	Used to separate the sentences in the verses if the verse contained 2 or more sentences

3. Declaration of AI (per person)

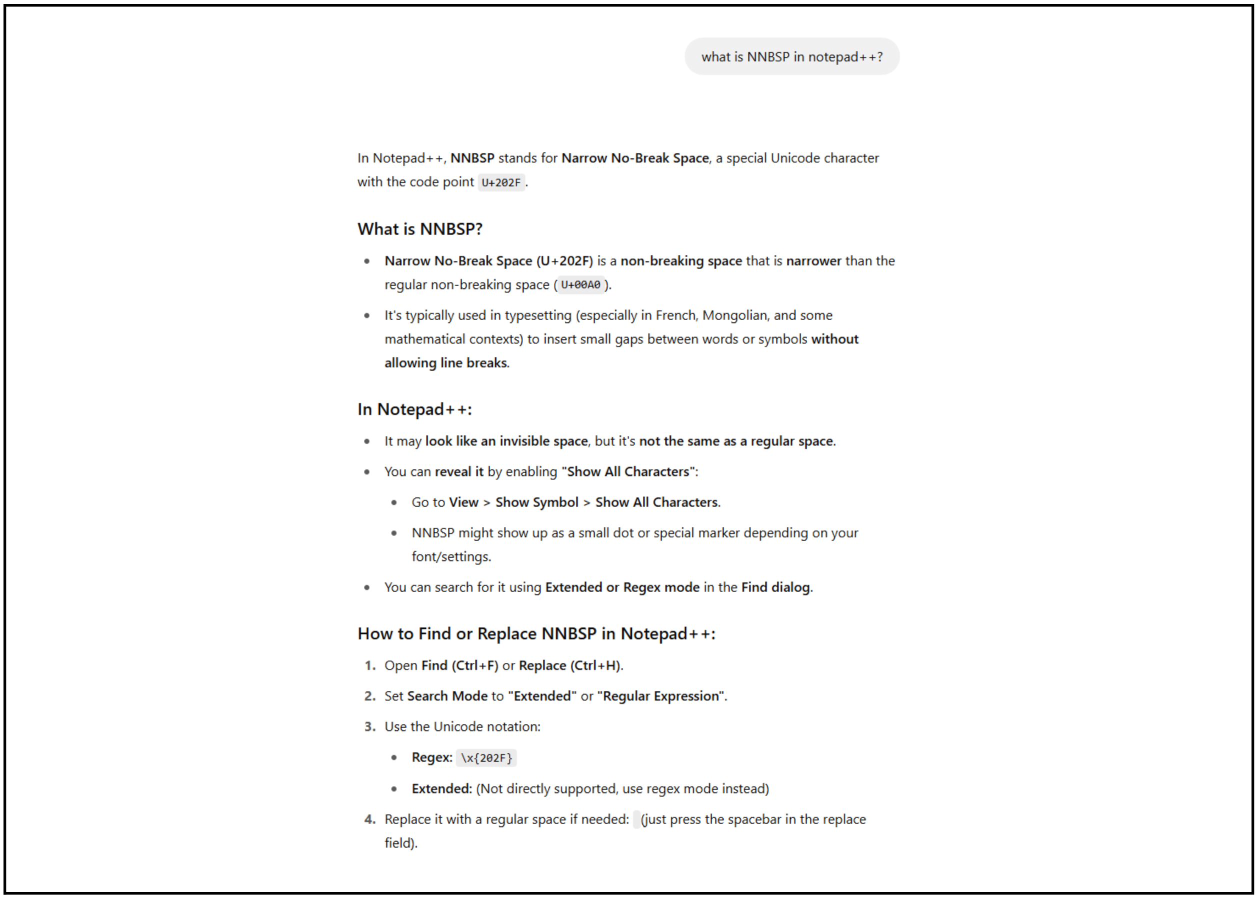
a. Carandang

AI Tools Used	Extent of Use
ChatGPT	Minimal
Description of AI Use	

I used ChatGPT to understand the formatting used by the bible. Since I was also not familiar with Bikolano, I checked whether or not there were problems in terms of grammar and punctuation.
Reflection
With the use of ChatGPT, I was able to understand the things that are important to clean and take note of when cleaning. Things like copying the book name and chapter help me keep them in mind. As well as the verse number being attached to the first sentence of the verse.

b. Sityar

AI Tools Used	Extent of Use
ChatGPT	Minimal
Description of AI Use	
I used ChatGPT to help me better understand the structure and formatting of the Bible text I was working with. It assisted in identifying specific challenges, especially non-breaking spaces.	
Reflection	
My use of AI helped clarify the unique formatting issues in the text, which made it easier for me to create appropriate Regex operations. While I wrote the expressions and handled the processing myself, the insights provided me with better insight.	



what is NNBSP in notepad++?

In Notepad++, **NNBSP** stands for **Narrow No-Break Space**, a special Unicode character with the code point `U+202F`.

What is NNBSP?

- **Narrow No-Break Space (U+202F)** is a **non-breaking space** that is **narrower** than the regular non-breaking space (`U+00A0`).
- It's typically used in typesetting (especially in French, Mongolian, and some mathematical contexts) to insert small gaps between words or symbols **without allowing line breaks**.

In Notepad++:

- It may **look like an invisible space**, but it's **not the same as a regular space**.
- You can **reveal it** by enabling **"Show All Characters"**:
  - Go to **View** > **Show Symbol** > **Show All Characters**.
  - NNBSP might show up as a small dot or special marker depending on your font/settings.
- You can search for it using **Extended or Regex mode** in the **Find dialog**.

How to Find or Replace NNBSP in Notepad++:

1. Open **Find (Ctrl+F)** or **Replace (Ctrl+H)**.
2. Set **Search Mode** to **"Extended"** or **"Regular Expression"**.
3. Use the Unicode notation:
   - **Regex:** `\x{202F}`
   - **Extended:** (Not directly supported, use regex mode instead)
4. Replace it with a regular space if needed: (just press the spacebar in the replace field).

c. Veracruz

AI Tools Used	Extent of Use
ChatGPT	Minimal
Description of AI Use	
I used ChatGPT to get an overview or summary of what possible formatting/structural changes I would need to implement in order to properly clean and segment the text into valid sentences through feeding it a few of the verses.	
Reflection	
The use of AI helped improve productivity and speed of working by providing me direction in what changes I may need to do, which allowed me to focus solely on writing and refining RegEx.	