Part A: Compression (40 points)

This exercise is an analogy, not the algorithm used by LZW or Huffman encoding.

Compression algorithms substitute repeating strings in the original text with a unique token and keep a record of each substitution in a dictionary. The dictionary is stored with the compressed text for later decompression.

An old quote from Vangie Beal, managing editor of Webopedia. (simplified in all lower case)

```
data compression is particularly useful in communications because it enables
devices to transmit or store the same amount of data in fewer bits. there
are a variety of data compression techniques, but only a few have been
standardized. the ccitt has defined a standard data compression technique
for transmitting and a compression standard for data communications through
modems. in addition, there are file compression formats, such as arc and
zip.
```

Compressed text:

```
!@is particularly useful in #because it enables
devices to $ or store the same amount of !in fewer bits.
%a variety of !@*s, but only a few have been
&ized. the ccitt has defined a & !@*
for $ting and a @& for !#through
modems. in addition,%file @formats, such as arc and zip.
```

Total original Characters (with spaces) count from Word doc		449		
token	compressed text	Dictionary size	*N* occurrences	Savings = Length × (*N*–1) - 1 – *N*
!	data	6	5	14
@	compression	13	5	42
#	communications	16	2	12
$	transmit	9	2	5
%	there are	12	2	8
&	standard	9	3	12
*	technique	10	2	6
Dictionary size		75		
Compressed Characters (with spaces) count from Word doc		275		

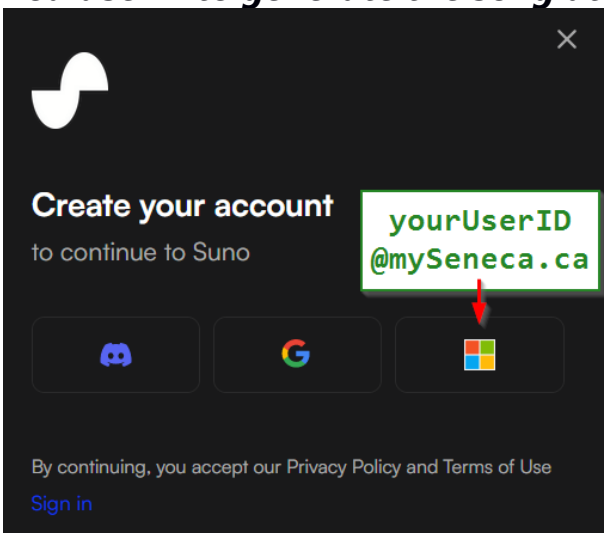
Saving = Total original Characters *subtract* (Dictionary size + Compressed Characters)	99		99
Compression = (Dictionary size + Compressed Characters) ÷ Total original Characters	78%		22%
Percentages should sum to 100%		100.00%	

Above is the token / compressed text dictionary, kept in the CP4P_Compression_Activity_calculator.xlsx Excel file, used to decompress tokens in the compressed text back to the original text.

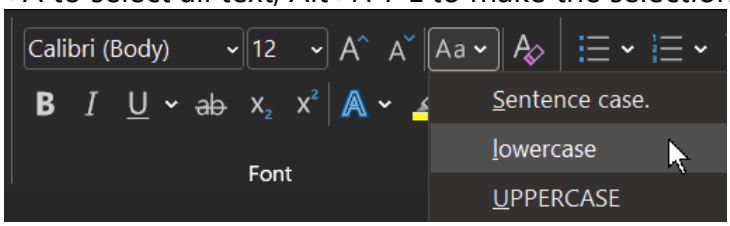
- Excel calculates the overhead of the dictionary in assessing the size of the compressed text versus the original plain text.

➔ How much can you compress the **lyrics to a song using the ideas above?**

You use AI to generate the song at Suno



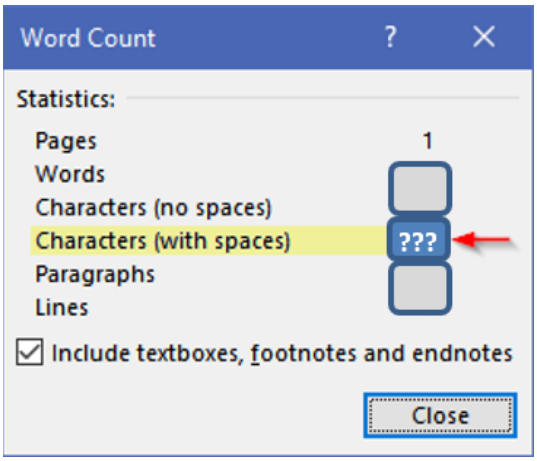
➔ Copy the lyrics of your generated song to a new MS-Word document (Ctrl+N).
➔ To reduce complexity, make all letters lower case:
Ctrl+A to select all text, Alt+H 7 L to make the selection lower case.



In the bottom left of the Word display, click "### words".

e.g. 

The Word Count dialog will pop up showing the number of characters with spaces.



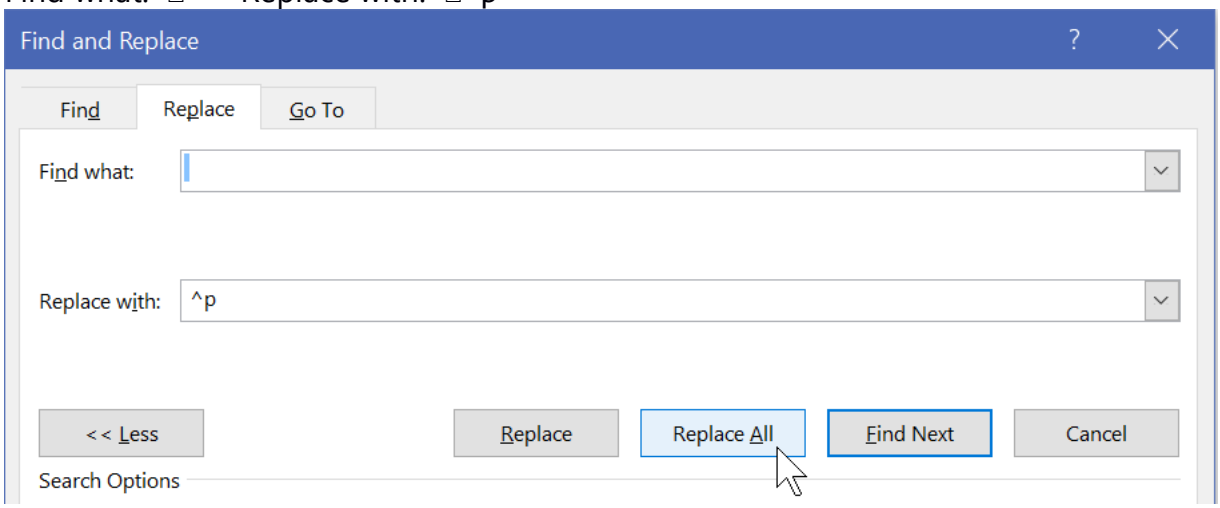
N.B. paragraph formatting codes are not counted by Word which is fine. Our exercise here is concerned only with the text.

Word *does* count new line (Shift + Enter) characters.

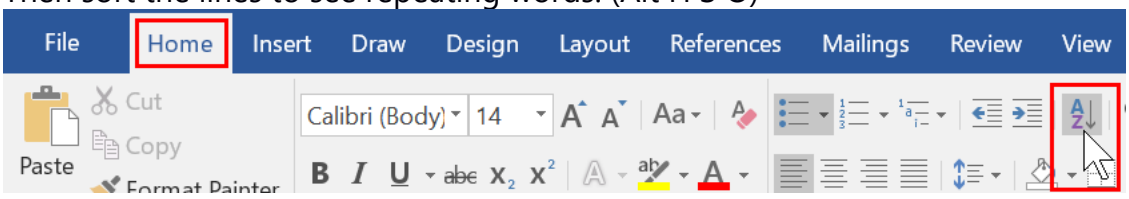
(Alt+H,8 will toggle the display of whitespace characters)

This will help with your substitution analysis: separate the words in the text so each is on its own line, then sort the lines to see repeating patterns of individual words.

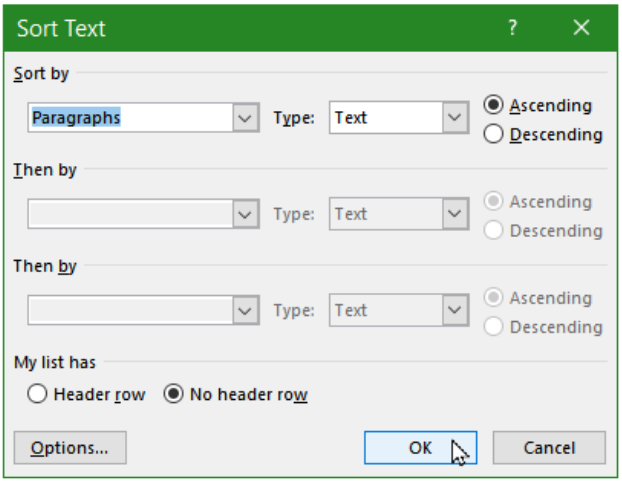
- copy the lyrics to another new document (Ctrl-N) used only for analysis
- Find and Replace a **space** with a **space + paragraph marker ^p** (Ctrl+H)
  Find what: □ Replace with: □^p



- Then sort the lines to see repeating words. (Alt H S O)



Sorting by Paragraphs results in one word per line, in alphabetical order;
this makes it easy to see repeating words:



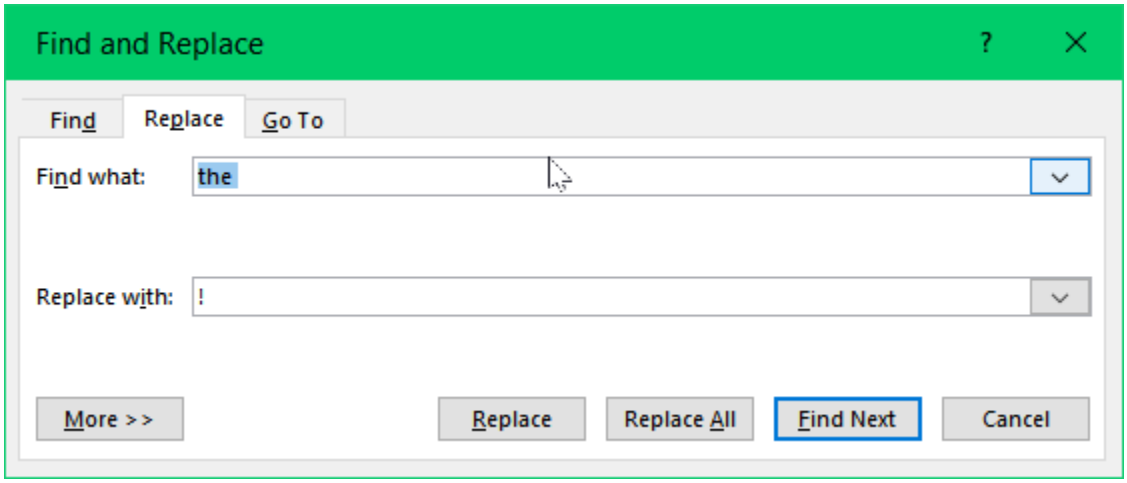
→ How much can you compress the lyrics?

Use the CP4P_Compression_Activity_calculator.xlsx Excel file (in the archive) to keep your compression dictionary and to help with optimisation. Sample data from the above example is in the green filled cells; replace it with yours. Be careful that the text you paste into the "compressed text" cell matches exactly to the substitutions made in the Word document.

You must use the locally installed application of Excel, not the web app, to see red dogeared cells where you can hover cursor for more info.

In Excel, hover your mouse pointer over cells with red dogear for instructions.

Anything occurring only once is not worth substituting with a token and including in the dictionary; you will be adding two characters (the tokens) to the file. A string with a length ≤ 3 and occurring only twice is similarly not worth it. A string ≤ 2 and occurring only three times is also not worth it.

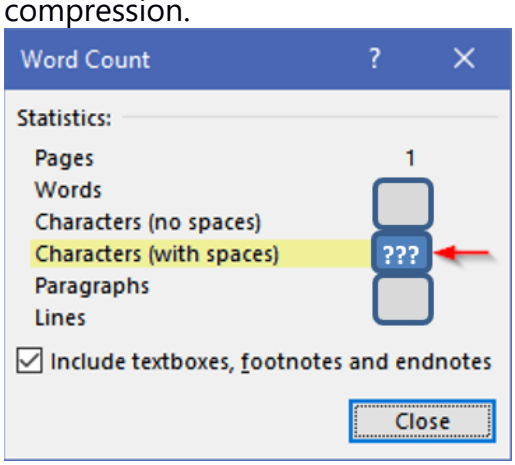


- a space is a *character* that can be compressed together with most word strings (Spaces*are*characters,itishardtoreadwordswithoutthem.)
- a mindless replacement of recurring words will not result in the best compression.

    - Consider compressing phrases before compressing individual words
    - Consider whether a leading and/or trailing space should be compressed with a word
    - Consider using portions of a word which may result in more substitutions. E.g. "transmit" and "technique" in the above example.
- Use unique tokens – symbols that do not appear in the lyrics. E.g. the special characters and digits on the keyboard's top row.
  N.B. do not use the **^ carat** symbol, it is a Microsoft escape character which will confuse its Find & Replace process.
- Decompression reads the token in the dictionary and the next characters to end-of-line as the original string to replace tokens with.

➔ 1. Paste the Suno Song Description you used to generate the song,
and the resulting lyrics into this question number in the _Activity_Answers.docx

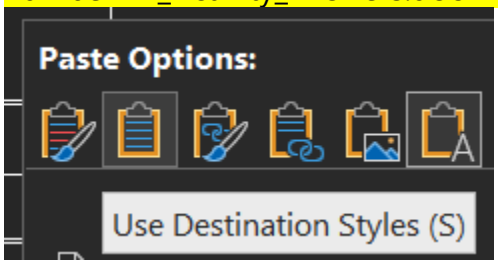
→ Update Excel "Total original Characters (with spaces) count from Word doc" **before** compression.



→ Update Excel rows with the character strings replaced by tokens

→ Update Excel "Compressed Characters (with spaces) count from Word doc" **after** compression.

➔ 2. After your compression efforts, select Excel cells A1 – E29 and paste into this question number in _Activity_Answers.docx with Paste Option: Use Destination Styles



➔ 3. Select the compressed text and paste into this question number in _Activity_Answers.docx

➔ 4. **Test your compression dictionary by decompressing.** Process dictionary items from the bottom up: find the compression character in the compressed data and replace it with the original string. **Paste the decompressed version** into this question number in _Activity_Answers.docx – *even if it is not perfect*.

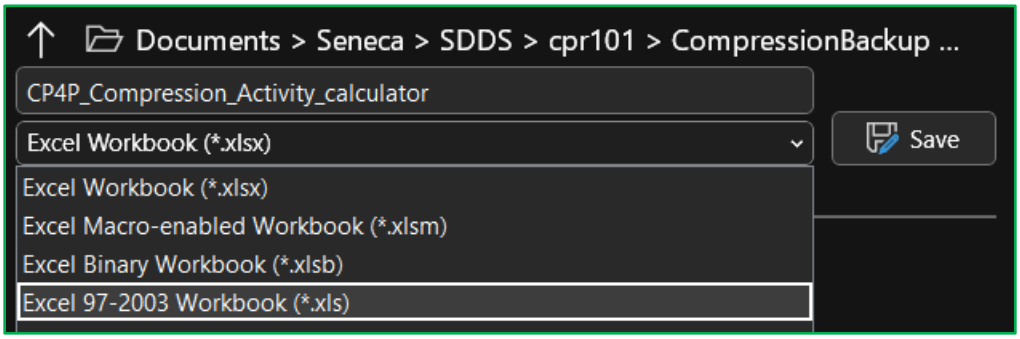
➔ 5. **What modifications, if any, does the compression dictionary need to return the compressed data back into its original state? (If none, then well done! but please make a note.)**

Part B: File formats with built-in compression…or not. (20 points)

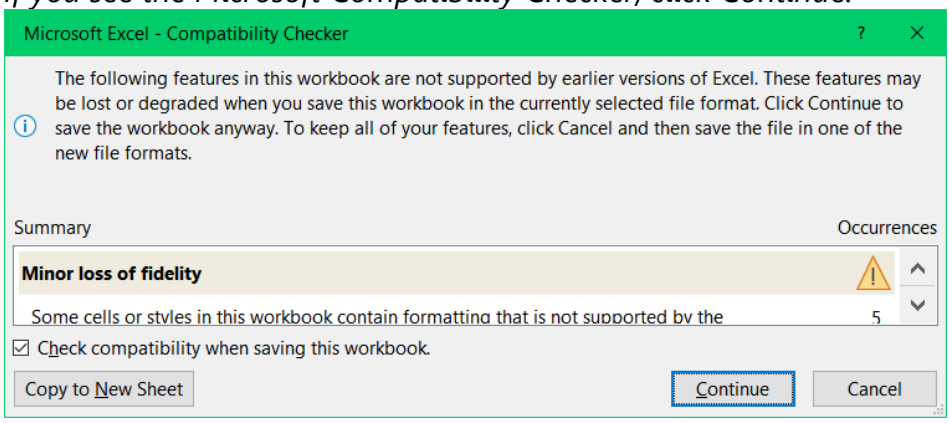
- Regarding this week's activity .zip archive downloaded to your folder.
  - Remember that compressed files must be decompressed before they can be opened.
  - Windows does this automatically into the %temp% folder if you open a file directly from a .zip archive. This is fine to quickly browse a file's content.
  - However, if the file is to be kept or its content modified,
    first copy/extract it from the .zip archive to your folder.
    - That will be the case for `CP4P_CompressionBackup_Activity_Answers.docx` because you will be adding your answers to it.
      First extract it from the archive to your folder, *then* open it.
      If you open it first – into %temp% – you may never find your work again.

These next steps will open an MS 365 file, then save it as an older type. We'll do that to examine the compression differences when we add them to the archive.

- Open the `CP4P_Compression_Activity_calculator.xlsx` file.
  - File menu > Save As > Excel 97-2003 Workbook (*.xls)



If you see the Microsoft Compatibility Checker, click Continue.



- Add the `CP4P_Compression_Activity_calculator.xls` file from your folder into the zip archive we have been using.
  - select the file, right click, and
    Send to > Compressed (zipped) folder option,

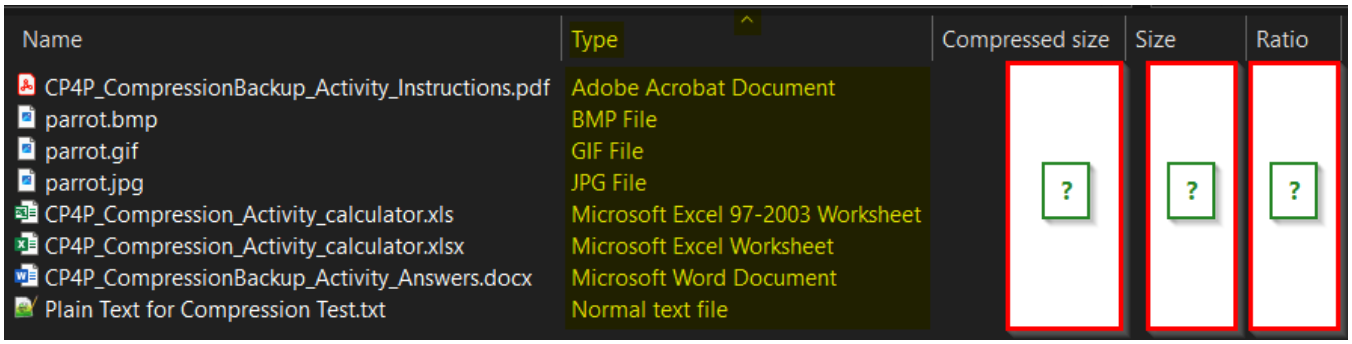
    or use 7zip [on Windows], or your favourite compression utility to drag and drop.

- Open the .zip archive with Windows File Explorer.
  - On macOS, open a terminal window and cd to folder with .zip file

    `$ zipinfo -m archivename.zip`

    `-m [medium]` shows % of file size saved by compression, higher is better.
    `-l [large ]` shows original and compressed file sizes in bytes.

- Use the Snipping Tool or Snip & Sketch (⊞ + "snip") to copy only the information seen below.



The Ratio shows the proportion of space saved. `Ratio = (Size – Compressed) / Size * 100`

"Ratio" is a misnomer because it is not a *ratio* of the sizes shown.
That column indicates % of space saved by compression.

FYI: opening the .zip archive with 7zip will show bytes, not rounded K bytes, for original Size and Packed (compressed) size. It will not show a percentage ratio.

See https://www.noupe.com/design/everything-you-need-to-know-about-image-compression.html

➔ **6.** Paste the image of the Windows [File] Explorer .zip archive information (or equivalent from macOS) into _Activity_Answers.docx.

➔ ***knowing the properties of file formats is essential to answering the questions below.***

Which image format should you use? See this. Reduce the Size of Microsoft Office Documents

➔ 7. Files with the **lowest** ratios were compressed the **least**. Ratio indicates % of space saved. Which file types compressed the least? Why would that be? (10 pts)

➔ 8. Files with the **highest** ratios were compressed the **most**.
Which file types compressed the most? Why would that be? (10 pts)

Part C: Backup (40 points)

The most common cause of data loss is accidental deletion of a file by the end-user on their own PC, or by IT professionals of a great many files on a server. To recover from these inevitable cases of *shooting yourself in the foot*, make a backup just before loading your gun.

> *A good backup software option for ICT people is Duplicati. The downsides should not bother School of Computer Programming & Analysis students. Senecans do not need a storage provider because we already have one: MS 365 OneDrive. See this review.*

A **backup** is a **copy** in a **geographically separate location** on an **independent platform**. A good backup location is Microsoft Office 365 OneDrive, in a folder that is *not* synchronized with any other system.

Microsoft OneDrive has a feature they call "Sync and backup" to "Back up important PC folders to OneDrive." **There is no such thing as both sync *and* backup, it is either / or, it cannot be both.** What **OneDrive** calls backup **is pure synchronization**. See THIS but ignore their software pitch for server backup / replication.

Another option is to collect your data into a zip archive using the password option, save it to a local USB drive (a copy on an independent platform *when ejected*), and also SFTP it to the matrix server (a second copy in a geographically separate location *and* on an independent platform).

a. Create a backup folder/directory on the target system.
b. Copy important files to that folder. e.g. the zip archive you created in Part 2. Because it is already compressed into a single file, it will take a minimum amount of time to upload.
c. Congratulations. You just backed up something.

➔ 9. paste screen shots showing the locations and contents of your backups into _Activity_Answers.docx. (use the Screen Snip tool) **(10 points)**

Imagine your laptop just stopped working and could not be restarted after you completed a great many hours of work today and yesterday. You need a backup & restore strategy.
(30 points total for four answers ~100 words each, 400 in total.)

➔ 10. What is (or what should have been) your backup routine? How do you ensure your backup is current?

➔ 11. How does your backup routine address the three characteristics of a real backup and fulfill the 3-2-1 backup checklist?

➔ 12. How does your backup routine address the three characteristics of a real backup and fulfill the 3-2-1 backup checklist? Now that you have a backup but no computer, how will you access and work with the current version of your backed up files? What is your restore/recovery strategy?

➔ 13. If your active files became completely unavailable tomorrow morning because your computer is somehow unavailable or there is no network connection or the files are corrupted / deleted/gone!, i.e. you are unable to work the way you expected, what would you do then?