

Word Analytics

You should implement a tool that gives the user some details on common word usage, letter usage, and some other analytics for a given document. More specifically, you must read a given text that is served by a web service in UTF-8 format and print off the following details:

1. Number of words
2. Number of letters
3. Number of symbols (any non-letter and non-digit character, excluding white spaces)
4. Top three most common words (you may count "small words", such as "it" or "the")
5. Top three most common letters
6. Most common first word of a paragraph (paragraph being defined as a block of text with an empty line above it) (Optional bonus)
7. Words only used once (Optional bonus)
8. All letters not used in the document (Optional bonus)

Please note that your tool does not have to be case sensitive, meaning the word "Hello" is the same as "hello" and "HELLO".

Formal Inputs & Outputs

Input Description

As an argument to your program on the command line, you will be given a an HTTPS endpoint from which the input text can be loaded. This text may be empty, but will be guaranteed well-formed (all valid UTF-8 characters). You can assume that line endings will follow the UNIX-style new-line ending (unlike the Windows carriage-return & new-line format).

For testing purpose use <https://vmsensorlog.westeurope.cloudapp.azure.com:1880/huckleberry-finn>

Output Description

For each analytic feature, you must print the results in a special string format. Simply you will print off 6 to 8 sentences with the following format:

"A words", where A is the number of words in the given document

"B letters", where B is the number of letters in the given document

"C symbols", where C is the number of non-letter and non-digit character, excluding white spaces, in the document

"Top three most common words: D, E, F", where D, E, and F are the top three most common words

"Top three most common letters: G, H, I", where G, H, and I are the top three most common letters

"J is the most common first word of all paragraphs", where J is the most common word at the start of all paragraphs in the document (paragraph being defined as a block of text with an empty line above it) (*Optional bonus*)

"Words only used once: K", where K is a comma-delimited list of all words only used once (*Optional bonus*)

"Letters not used in the document: L", where L is a comma-delimited list of all alphabetic characters not in the document (*Optional bonus*)

If there are certain lines that have no answers (such as the situation in which a given document has no paragraph structures), simply do not print that line of text.