Homework 1

Write a bash / shell script to accomplish the following tasks. You can just break the tasks into separate sections in your code, use comments to indicate which parts answer which of the questions. You can use the echo command to print out a message with the result if needed.

1. Getting data
    - Use wget or curl to download the text file *The Variation of Animals and Plants Under Domestication, Vol. I.* originally downloaded from https://www.gutenberg.org/ebooks/24923
    - How big is this file (in kilobytes)?
2. Compressing and uncompressing
    - Compress the File pg24923.txt you downloaded with gzip, how big is it in kilobytes?
    - Uncompress it, then compress it with bzip2, how big is it in kilobytes?
    - Uncompress it again
3. Counting
    - How many total words are in the file presenting Darwin's *The Variation of Animals and Plants Under Domestication, Vol. I.*
    - How many rows are in this data file
4. Sorting
    - Sort the data file based on the FPKM column (gene expression) (write out to a new file called Nc20H.expr.sorted.tab)
    - How many exon features are there in the rice chromosome 6 gff file (this file is compressed). You can do this in several different ways.
5. Finding and Counting
    - Count the number of gene features in this genbank file - see for example this explanation of a genbank file if you are not familiar.
6. Column combining
    - Take these files Nc20H Nc3H and combine columns 1-6 (gene_id,bundle_id,chr,left,right,FPKM) from Nc20H and column 6 from Nc3H to make a new file with 7 columns the last 2 are gene expression values for the two experiments. You should use the cut and paste UNIX commands to accomplist this, though use of other tools is permitted. You may want to validate the gene counts in both files, using sort and uniq