This is adapted from Homework_0b from the Environmental Bioinformatics course at Woods Hole Oceanographic Institution (Fall 2021)

You will be working with nucleotide sequences in the fasta files `Griffin.fa` and `Unicorn.fa` which are inside the sequence folder. Take a look at both your files. Notice that for each sequence there is a header that starts with > and contains a unique identifier and other information. To store your results and the other requested files, please make a new directory inside `sequences` called `results`.

1. Write a loop that will count the **number of sequences** of each file and write the output to a the file `num_seqs` . Also save your command in a separate file `num_seqs_command`. *Hint: the number of lines doesn't equal to the number of sequences in these files. Think if there are elements that you can count.*
2. Count how many sequences contain the codon AGT at the **beginning** (of the sequence) in `Griffin.fa`. Copy the command you used and the nember of sequences in a file called `AGT_Griffin`.*Hint: check the grep command and the options it offers.*
3. Write an one-liner (commands one after the other using pipe) that extract all unique "starting" codons (triplets of nucleotides at the beginning of the sequence) and how many times they occur in descenting order. Save the command in the file `extract_starting_triplet` and the output in `counts_starting_triplet`. *Hint: check the commands grep, cut, sort and uniq*
4. Transform the command above to a script that takes as input the name of a sequence file and produces a `counts_starting_triplet` file. Save the script as `extract_starting_triplet.sh`. Test it at the `Unicorn.fa`.
5. The Unicorn.fa file appears to have a space after the > in the sequence header, which causes the software we use to process it to fail. Write a command to get rid of the space and write the fixed fasta to a new file called `unicorn-nospace.fa`. *Don't delete all the spaces in the fasta header!* Report the command you used in the file `remove_space`.
6. Even after the removal of the space there are still problems with the files caused by the non-alpha numeric characters (e.g. spaces, , =). Create a new file for each of the Unicorn and Griffin fasta files that saves all the information from the headers as `{}-headers.txt`. Then create a new cleaned fasta file that retains only the sequence name and removes all the descriptors. Save these new files as `{}_cleaned.fa`. Report the commands you used in the file `clean_headers`.
7. Save all your command history from this homework with: `history > homework.log`

For the successful completion of your homework, you should have the following files in your `results` directory you create.

- num_seqs
- num_seqs_command
- AGT_Griffin
- extract_starting_triplet
- extract_starting_triplet.sh
- unicorn-nospace.fa
- remove_space
- griffin-headers.txt and unicorn-headers.txt
- griffin-cleaned.fa and unicorn-cleaned.fa
- clean_headers
- homework.log

In most of the problems, there are more than one correct answers :)