HW#0 - Homework Report
Sofia Huang
09/02/2022 (due 09/13/2022)

Part 1, Question 1

You may copy the question into your report, but make sure that you make it clear where the question ends and your answer begins.

Answer

All figures must have a caption and must be referenced in the text. Example below.

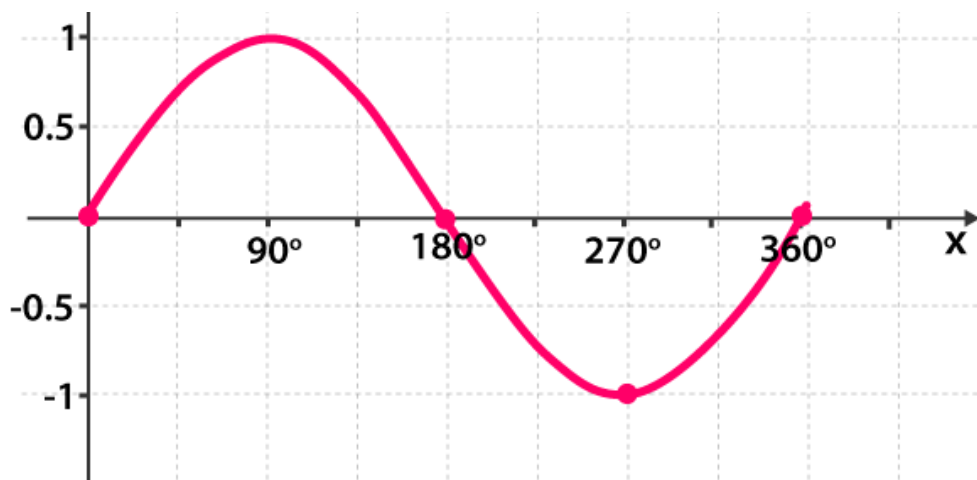Figure 1 shows the graph of the trigonometry function $y = sinx$.

Figure 1: Sine Graph

If you want to include code in your report, you can insert a screenshot (if it's legible), or you can copy/paste the code into a listings environment. There are examples below and more information is available at $https://www.overleaf.com/learn/latex/code\_listing$.

Listing 7 is an example of directly copying code into the LaTeX document and having the listings package perform syntax highlighting. Listing 2 is an example of importing the code from a file rather than copying it in.

```
1 a = 1
2
3 try:
4     b = int(input("Please enter a number to divide a"))
```

```
 5      a = a/b
 6      print("Success a=",a)
 7 except ZeroDivisionError:
 8      print("The number you provided cant divide 1 because it is 0")
 9 except ValueError:
10      print("You did not provide a number")
11 except:
12      print("Something went wrong")
```

Listing 1: Python example copied into the LaTeX

```
 1 a = 1
 2
 3 try:
 4      b = int(input("Please enter a number to divide a"))
 5      a = a/b
 6      print("Success a=",a)
 7 except ZeroDivisionError:
 8      print("The number you provided cant divide 1 because it is 0")
 9 except ValueError:
10      print("You did not provide a number")
11 except:
12      print("Something went wrong")
```

Listing 2: Python sample code loaded from file

Table 1 shows a simple example table. Table 2 shows an example confusion matrix (you'll see this term later) from `https://en.wikipedia.org/wiki/Confusion_matrix`. This employs rows that span multiple columns (multicol) and columns that span multiple rows (multirow).

Table 1: Simple Table

Week	Date	Topic
1	Sep 1, 6	Introduction to Web Science and Web Architecture
2	Sep 8, 13	Introduction to Python
3	Sep 15, 20	Introduction to Info Vis with R, Python
4	Sep 22, 27	Measuring the Web

Table 2: Example Confusion Matrix from Wikipedia

		Actual	
		Cat	Dog
Predicted	Cat	5 (TP)	3 (FP)
	Dog	2 (FN)	3 (TN)

Discussion

For the first part of the homework assignment I followed the instructions given to replace various figures and code blocks to get familiar with Latex.

Part 2, Question 1

Answer

```
1 (base) sofiahuang@Sofias-MacBook-Pro Desktop % chmod 700 data440
2 (base) sofiahuang@Sofias-MacBook-Pro Desktop % ls -l data440
3 total 8
4 -rwx------@ 1 sofiahuang  staff  45 Sep  6 10:26 test.txt
```

Listing 3: Permissions on directory

Discussion

I googled how to change the permissions on a directory using the Terminal and I came across chmod. The 700 after the command denotes what the permissions are being changed to. The first digit placeholder represents the user permissions, the second is for groups, and the third is for others. Then, the actual number represents the sum of the type of permissions: 4 - read, 2 - write, 1 - execute, 0 - none. So, to give the user permission to read, write, and execute, we add $4 + 2 + 1 = 7$. Then, everyone else (groups and others) is assigned 0 for no permissions.

Part 2, Question 2

Answer

```
1 (base) sofiahuang@Sofias-MacBook-Pro data440 % wc -l test.txt
2        5 test.txt
3
4 (base) sofiahuang@Sofias-MacBook-Pro data440 % echo "CS 800" >> test.
    txt; cat test.txt
5 CS 800
6 CS 432
7 CS 725
8 MATH 212
9 MATH 32
```

```
10 CS 800
11
12 (base) sofiahuang@Sofias-MacBook-Pro data440 % grep CS test.txt
13 CS 800
14 CS 432
15 CS 725
16 CS 800
17
18 (base) sofiahuang@Sofias-MacBook-Pro data440 % grep -c CS test.txt
19 4
20
21 (base) sofiahuang@Sofias-MacBook-Pro data440 % sort test.txt
22 CS 432
23 CS 725
24 CS 800
25 CS 800
26 MATH 212
27 MATH 32
28
29 (base) sofiahuang@Sofias-MacBook-Pro data440 % sort -k2 test.txt
30 MATH 212
31 MATH 32
32 CS 432
33 CS 725
34 CS 800
35 CS 800
36
37 (base) sofiahuang@Sofias-MacBook-Pro data440 % sort -k2 -n test.txt
38
39 MATH 32
40 MATH 212
41 CS 432
42 CS 725
43 CS 800
44 CS 800
45
46 (base) sofiahuang@Sofias-MacBook-Pro data440 % sort test.txt | uniq -c
47
48    1 CS 432
49    1 CS 725
50    2 CS 800
51    1 MATH 212
52    1 MATH 32
```

Listing 4: Unix commands

Discussion

For each command, I used the `man` command in the Terminal to find out its purpose.

The `wc -l` command returns the line count (specified by the `-l` parameter) of the input file, which was 5.

The `echo` command writes arguments to the standard output, which is what you see returned in the Terminal after executing commands that have an output. The `>>` symbol is used for output redirection, which is how the output, "CS 800", from the `echo` command is able to be appended to the input file. Then, the semicolon is used to run multiple Linux commands in one line (using a semicolon, all commands are run, regardless if the previous executed successfully or not). The `cat` command is used to read and print the specified file to the standard output so we can check and see that "CS 800" was indeed appended to the text file.

The `grep` command searches the file and returns the lines that match the given patterns/regular expressions. In this example, we looked for "CS" and all of the lines in the test file with that expression were returned.

When you add the `-c` parameter to the `grep` command, the output returns the count of matches from the input file, which in this case was 4.

The `sort` command sorts the input file by line, lexicographically.

When the `-k2` parameter is added, the input file is sorted by the second column. However, still in lexicographical order, which is why "212" is before "32".

The `-n` parameter tells the command to sort numerically. Combined with the `-k2` parameter, we see that the file is sorted by the second column in ascending numerical order.

Lastly, the pipe command combine multiple commands so that the output of one is the input of the next. The `uniq` command compares each adjacent line and filters out duplicates. It is best to sort the input file first since this command compares adjacent lines, so for example, if the first and last line (not adjacent) are duplicates of each other, this will not be detected and both lines will be outputted. By adding the `-c` parameter, the count of each line is added to the output. Since "CS 800" appears twice in the file, there is a "2" in front of it.

Part 3, Question 1

Answer

```
1 (base) sofiahuang@Sofias-MacBook-Pro data440 % python3 get_tweets.py "
    national park"
2 25 tweets written to tweets.jsonl for query "national park lang:en has:
    links -is:retweet"
```

Listing 5: Running get_tweets.py

```
1 (base) sofiahuang@Sofias-MacBook-Pro data440 % jq . tweets.jsonl
2 {
3   "text": "Come tour with us!\nThe Mole National Park is a place to
    visit and experience wildlife in the Savanna Region of Ghana.\
    nNature is beautiful, Yaa Asantewaa Tours is here for you!!\nContact
     us anytime for adventurous and exciting tours.\n0209079288\
    nyaaahendea@gmail.com\n#pulseofafrica https://t.co/YhUHPGdUTO",
4   "id": "1567876436031062018",
5   "entities": {
6     "annotations": [
7       {
8         "start": 18,
9         "end": 40,
10        "probability": 0.2616,
11        "type": "Place",
12        "normalized_text": "The Mole National Park"
13      },
14      {
15        "start": 93,
16        "end": 115,
17        "probability": 0.8115,
18        "type": "Place",
19        "normalized_text": "Savanna Region of Ghana"
20      },
21      {
22        "start": 139,
23        "end": 157,
24        "probability": 0.5972,
25        "type": "Place",
26        "normalized_text": "Yaa Asantewaa Tours"
27      }
28    ],
29  ...
```

Listing 6: tweets.jsonl output snippet

Discussion

After I created a Developer Account and installed twarc, I had to find where the config file was loacted on my computer that contained my keys in order to run `get_tweets.py`. Once I found it and changed the file path in the python script, I was able to run and get tweets (I chose to use the keywords "national park"). In order to display the JSONL file in an organized manner, I installed and used the `jq` package, which is a command-line JSON file processor.

Part 3, Question 2

Answer

```
 1 (base) sofiahuang@Sofias-MacBook-Pro data440 % python3 process_tweets.
     py < tweets.jsonl > tweets_info.txt
 2 (base) sofiahuang@Sofias-MacBook-Pro data440 % cat tweets_info.txt
 3 1567876436031062018 2022-09-08T14:04:06.000Z  yaaasantewaa_ad 0
 4   Brand Category: Sports & Fitness Business
 5   Brand Category: Travel & Transportation Business
 6   Entities [Entity Service]: Travel
 7   Interests and Hobbies Category: Destinations
 8   Interests and Hobbies Category: Outdoors
 9   Interests and Hobbies: National parks
10   Unified Twitter Taxonomy: National parks
11   https://twitter.com/yaaasantewaa_ad/status/1567876436031062018/photo
     /1
```

Listing 7: tweets_info.txt output snippet

Discussion

For this question I ran the line of code provided in the instructions and then used the `cat` command to show the .txt file that was created. Above is a snippet of the output showing the tweet information.

References

Every report must list the references that you consulted while completing the assignment. If you consulted a webpage, you must include the URL.

- StackExchange - Can't edit a downloaded template from Overleaf on TexMaker `https://tex.stackexchange.com/questions/593363/cant-edit-a-downloaded-template-from-overleaf-on-texmaker`

- How to Create a Folder in Github Repos in 4 Simple Steps `https://www.alpharithms.com/how-to-create-a-folder-in-github-repos-463022/`

- Linux Commands `https://linuxconfig.org/linux-commands-cheat-sheet`

- Latex Inline Code `https://tex.stackexchange.com/questions/286094/insert-code-keywords-inline`

- Linux chmod `https://www.computerhope.com/unix/uchmod.htm`