2. Text Analysis using CLI

Gaurav Ojha
VisFac @ MPSTME, Mumbai

Why learn Linux CLI?

1. **Data Processing and Analysis:** Big data often requires complex processing and analysis tasks. The Linux CLI provides a powerful environment with numerous tools and utilities that allow you to efficiently process, clean, and transform large datasets.
2. **Scalability and Performance:** Linux is widely used in server environments, including big data clusters. Understanding the CLI enables you to navigate and manage these systems effectively, optimizing performance and ensuring scalability.
3. **Automation and Scripting:** Dealing with big data often involves repetitive tasks. By mastering the CLI, you can create scripts and automate various processes, saving time and reducing the risk of errors.
4. **Flexibility and Customization:** Linux offers great flexibility and customization options. You can tailor your environment, install specific software packages, and configure settings according to your big data project's needs.
5. **Remote Access and Cloud Computing:** Big data projects often involve remote servers and cloud platforms. The Linux CLI allows you to access and manage these resources from anywhere, making it easier to work with distributed data.
6. **Version Control and Collaboration:** Many big data projects involve collaboration among team members. The CLI interfaces seamlessly with version control systems like Git, allowing efficient collaboration and code sharing.
7. **Troubleshooting and Debugging:** Linux provides various debugging and monitoring tools that are invaluable when dealing with big data systems. Understanding the CLI helps you diagnose and resolve issues effectively.
8. **Data Security:** Big data often involves sensitive information. Knowing the Linux CLI allows you to manage user permissions, secure data, and implement various security measures to protect valuable data assets.
9. **Resource Management:** Working with big data can be resource-intensive. The CLI enables you to monitor system resources, such as CPU, memory, and disk usage, ensuring efficient resource management.
10. **In-demand Skillset:** Proficiency in the Linux CLI is a highly sought-after skill in the big data industry. Whether you're pursuing a career as a data engineer, data scientist, or data analyst, knowing how to work with the Linux CLI can significantly enhance your job prospects.

Setting Up Shared Folder

1. Create a shared folder in the host os name it : "Shared Files"
2. Create a folder in the VM, name it: "Windows"
3. In the VM open the terminal add the following commands:
    - su
    - password: cloudera
    - mount -t vboxsf SharedFolder /home/cloudera/Desktop/Windows

Now the shared files is active

- Also in the VM go to devices -> Shared Clipboard -> Bidirectional (This will aloows us to copy files between the 2 OS)
- Download Sample.txt from teams in the shared folder

Text Analysis Exercise using the Terminal:

For this exercise, we will be using the terminal to perform text analysis on the provided "sample.txt" file. The tasks will include word count, line count, finding specific words, and extracting specific lines.

1. Word Count and Line Count:
   a) Print the total word count in the "sample.txt" file.
   b) Print the total line count in the "sample.txt" file.
2. Find Specific Words:
   a) Search for the number of occurrences of the word "Voldemort" in the file.
   b) Search for the number of occurrences of the word "Snape" in the file.
3. Extract Specific Lines:
   a) Extract and print all the lines containing the word "Lily" in the file.
   b) Extract and print all the lines containing the word "Hogwarts" in the file.
4. Advanced Challenge - Character Count:
   a) Print the total character count (including spaces) in the "sample.txt" file.
5. Advanced Challenge - Unique Word Count:
   a) Print the count of unique words in the "sample.txt" file (excluding punctuation and case sensitive).
6. Advanced Challenge - Longest Word:
   a) Find and print the longest word in the "sample.txt" file.
7. Advanced Challenge - Word Frequency Analysis:
   a) Create a list of the 10 most frequently used words in the "sample.txt" file, along with their corresponding frequencies.

8. Advanced Challenge - Replace Text:
    a) Replace all occurrences of the word "Harry" with "Potter" in the "sample.txt" file and save the changes to a new file named "modified_sample.txt".
9. Advanced Challenge - Sort Text:
    a) Sort all the lines in the "sample.txt" file in alphabetical order and save the sorted lines to a new file named "sorted_sample.txt".
10. Clean up:
    a) Remove any intermediate files created during the exercise.

Note: For this exercise, learners will need to use various terminal commands such as `cat`, `grep`, `wc`, `sed`, and `sort`. The advanced challenges will require additional research and experimentation to accomplish.

Solution

---

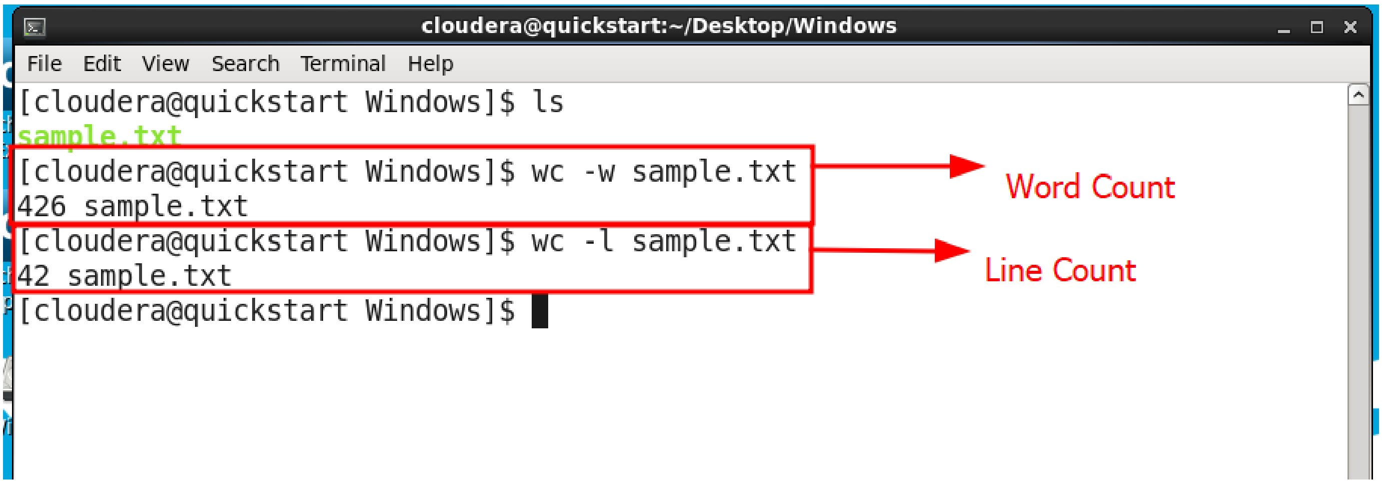
1. Word Count and Line Count:

a) Print the total word count in the "sample.txt" file.

b) Print the total line count in the "sample.txt" file.

```
# Total word count
wc -w sample.txt

# Total line count
wc -l sample.txt
```



```
[cloudera@quickstart Windows]$ ls
sample.txt
[cloudera@quickstart Windows]$ wc -w sample.txt
426 sample.txt
[cloudera@quickstart Windows]$ wc -l sample.txt
42 sample.txt
[cloudera@quickstart Windows]$ 
```

cloudera@quickstart:~/Desktop/Windows
File Edit View Search Terminal Help
Word Count
Line Count

---

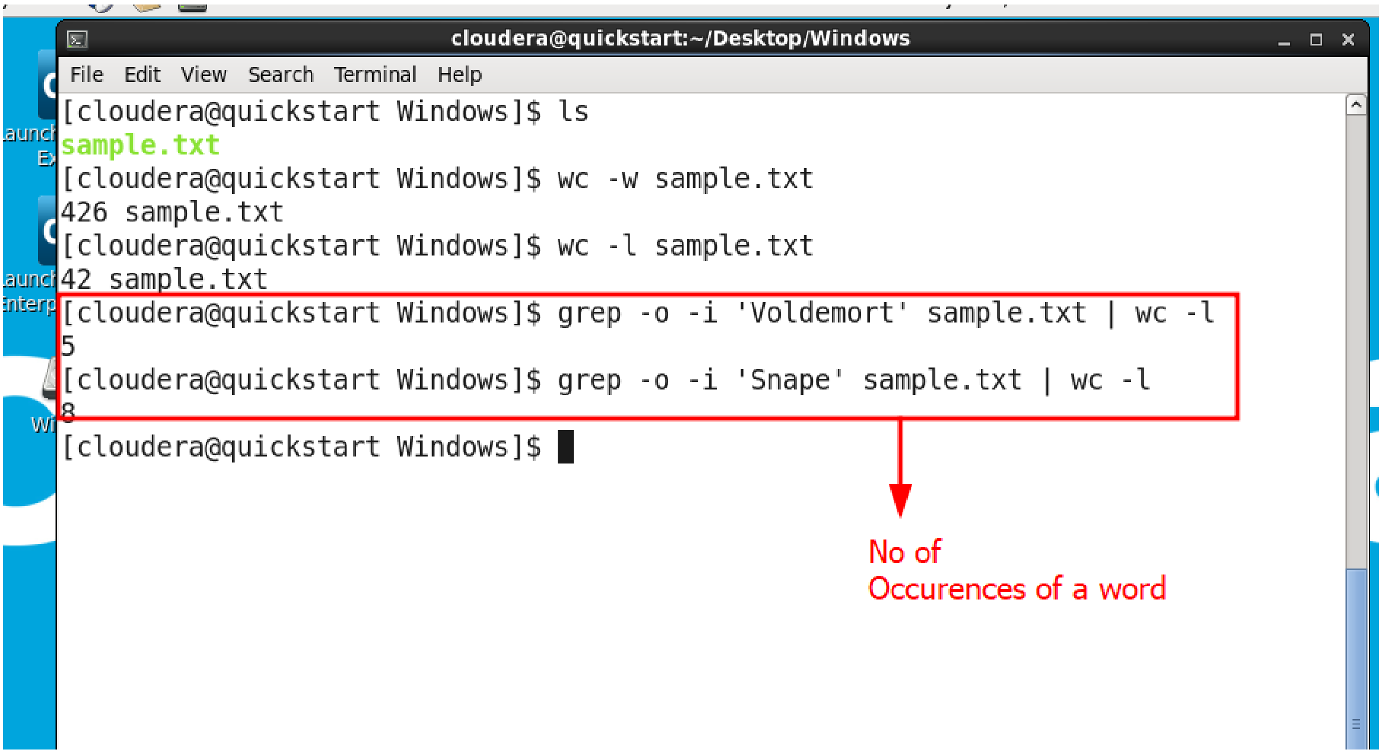
2. Find Specific Words:

a) Search for the number of occurrences of the word "Voldemort" in the file.

b) Search for the number of occurrences of the word "Snape" in the file.

```
# Number of occurrences of "Voldemort"
grep -o -i 'Voldemort' sample.txt | wc -l
```

```
# Number of occurrences of "Snape"
grep -o -i 'Snape' sample.txt | wc -l
```



cloudera@quickstart:~/Desktop/Windows
File Edit View Search Terminal Help

```
[cloudera@quickstart Windows]$ ls
sample.txt
[cloudera@quickstart Windows]$ wc -w sample.txt
426 sample.txt
[cloudera@quickstart Windows]$ wc -l sample.txt
42 sample.txt
[cloudera@quickstart Windows]$ grep -o -i 'Voldemort' sample.txt | wc -l
5
[cloudera@quickstart Windows]$ grep -o -i 'Snape' sample.txt | wc -l
8
[cloudera@quickstart Windows]$ 
```

No of
Occurences of a word

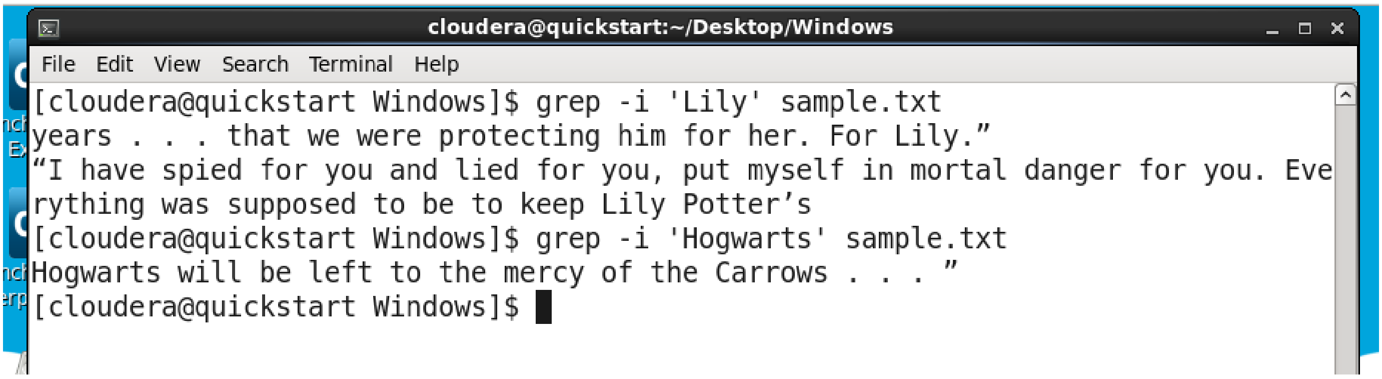
3. Extract Specific Lines:

a) Extract and print all the lines containing the word "Lily" in the file.

b) Extract and print all the lines containing the word "Hogwarts" in the file.

```
# Extract lines containing "Lily"
grep -i 'Lily' sample.txt

# Extract lines containing "Hogwarts"
grep -i 'Hogwarts' sample.txt
```



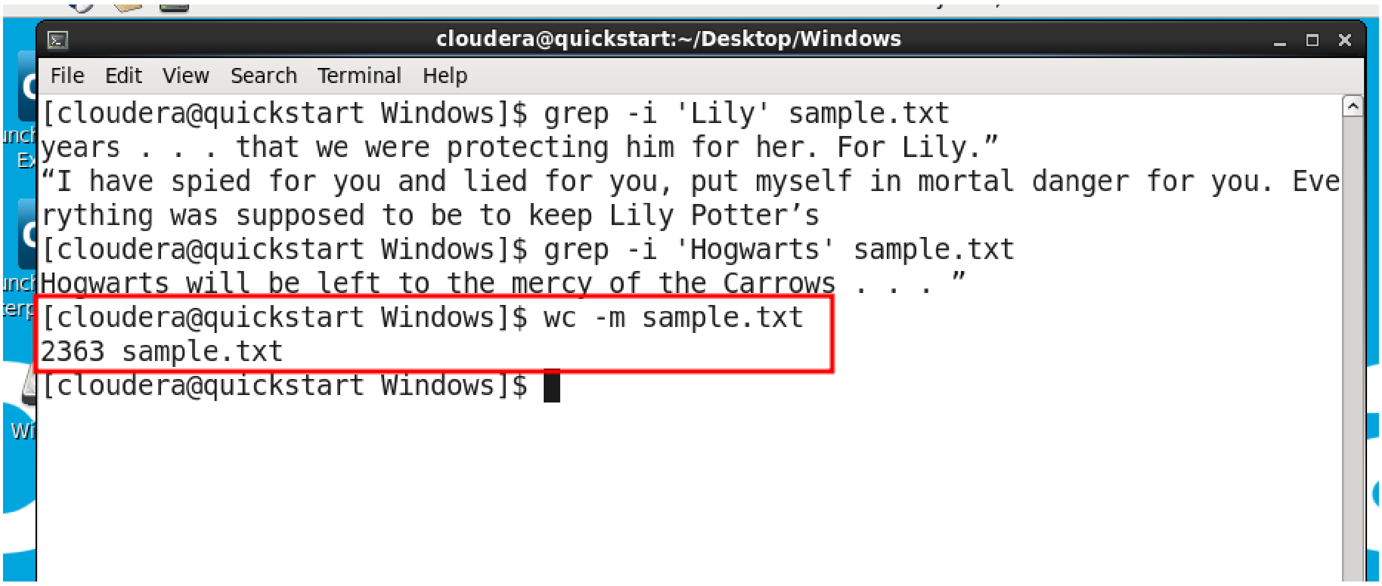
cloudera@quickstart:~/Desktop/Windows
File Edit View Search Terminal Help

```
[cloudera@quickstart Windows]$ grep -i 'Lily' sample.txt
years . . . that we were protecting him for her. For Lily.”
“I have spied for you and lied for you, put myself in mortal danger for you. Eve
rything was supposed to be to keep Lily Potter’s
[cloudera@quickstart Windows]$ grep -i 'Hogwarts' sample.txt
Hogwarts will be left to the mercy of the Carrows . . . ”
[cloudera@quickstart Windows]$ 
```

4. Advanced Challenge - Character Count:

a) Print the total character count (including spaces) in the "sample.txt" file.

```
# Total character count (including spaces)
wc -m sample.txt
```

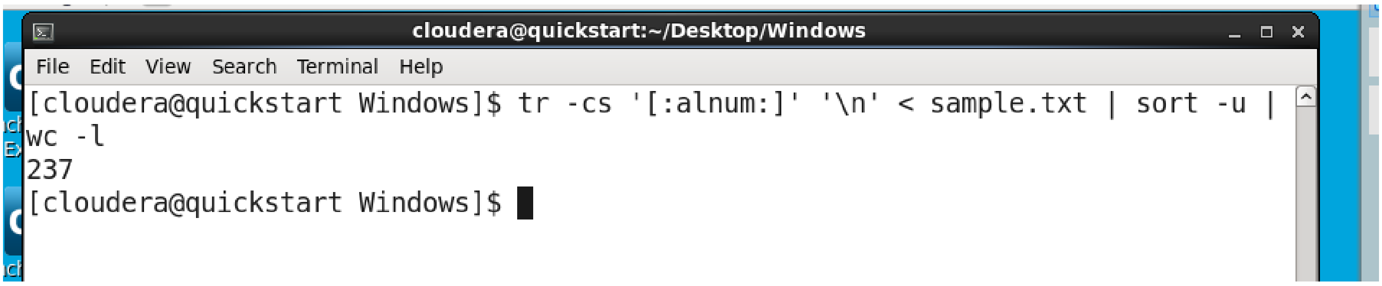


```
[cloudera@quickstart Windows]$ grep -i 'Lily' sample.txt
years . . . that we were protecting him for her. For Lily.”
“I have spied for you and lied for you, put myself in mortal danger for you. Eve
rything was supposed to be to keep Lily Potter’s
[cloudera@quickstart Windows]$ grep -i 'Hogwarts' sample.txt
Hogwarts will be left to the mercy of the Carrows . . . ”
[cloudera@quickstart Windows]$ wc -m sample.txt
2363 sample.txt
[cloudera@quickstart Windows]$
```

5. Advanced Challenge - Unique Word Count:

a) Print the count of unique words in the "sample.txt" file (excluding punctuation and case sensitive).

```
# Count of unique words (excluding punctuation and case sensitive)
tr -cs '[:alnum:]' '\n' < sample.txt | sort -u | wc -l
```

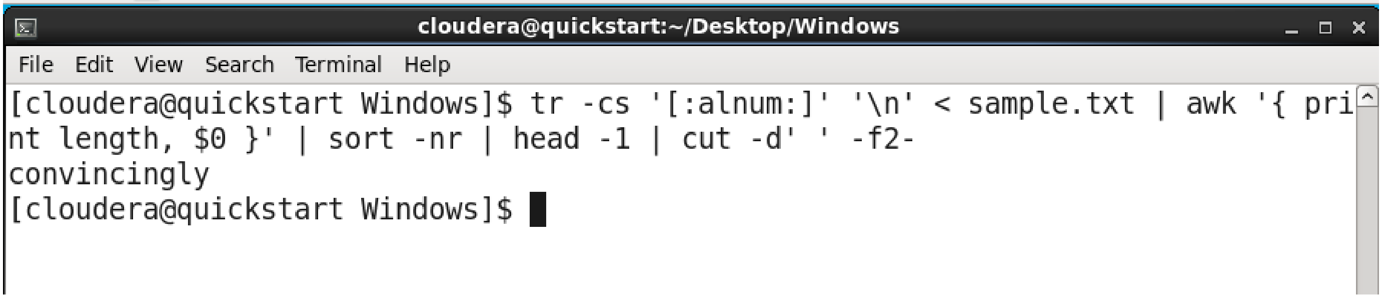


```
[cloudera@quickstart Windows]$ tr -cs '[:alnum:]' '\n' < sample.txt | sort -u |
wc -l
237
[cloudera@quickstart Windows]$
```

6. Advanced Challenge - Longest Word:

a) Find and print the longest word in the "sample.txt" file.

```
# Find and print the longest word
tr -cs '[:alnum:]' '\n' < sample.txt | awk '{ print length, $0 }' | sort -nr
| head -1 | cut -d' ' -f2-
```

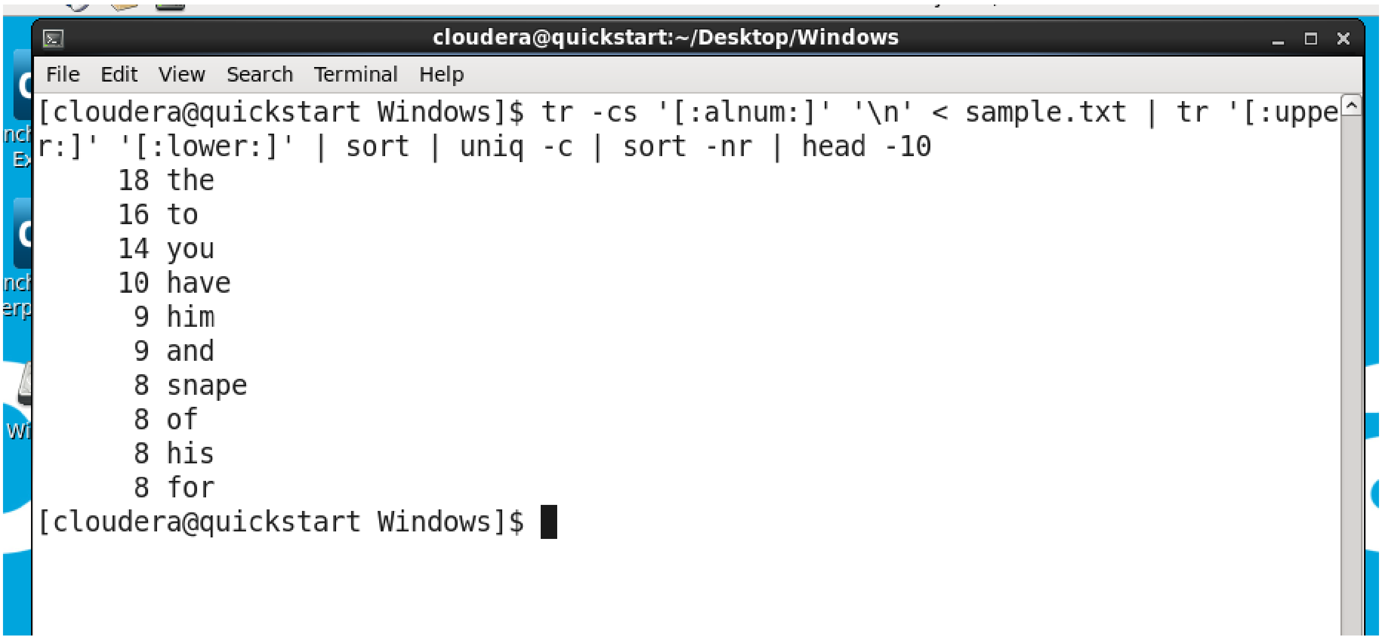


```
[cloudera@quickstart Windows]$ tr -cs '[:alnum:]' '\n' < sample.txt | awk '{ pri
nt length, $0 }' | sort -nr | head -1 | cut -d' ' -f2-
convincingly
[cloudera@quickstart Windows]$
```

7. Advanced Challenge - Word Frequency Analysis:

a) Create a list of the 10 most frequently used words in the "sample.txt" file, along with their corresponding frequencies.

```
# List 10 most frequently used words with their corresponding frequencies
tr -cs '[:alnum:]' '\n' < sample.txt | tr '[:upper:]' '[:lower:]' | sort |
uniq -c | sort -nr | head -10
```

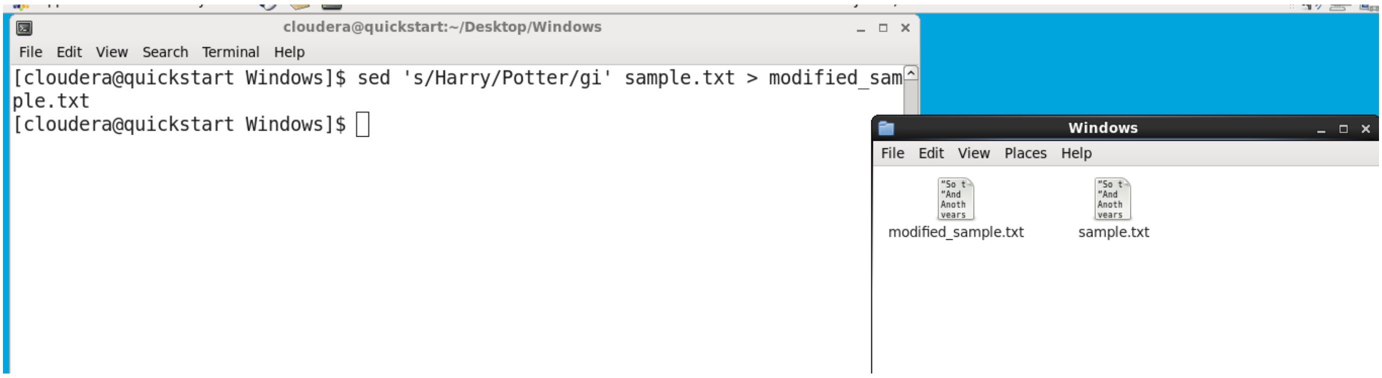

cloudera@quickstart:~/Desktop/Windows
File Edit View Search Terminal Help

```
[cloudera@quickstart Windows]$ tr -cs '[:alnum:]' '\n' < sample.txt | tr '[:uppe
r:]' '[:lower:]' | sort | uniq -c | sort -nr | head -10
     18 the
     16 to
     14 you
     10 have
      9 him
      9 and
      8 snape
      8 of
      8 his
      8 for
[cloudera@quickstart Windows]$ 
```

8. Advanced Challenge - Replace Text:

a) Replace all occurrences of the word "Harry" with "Potter" in the "sample.txt" file and save the changes to a new file named "modified_sample.txt".

```
# Replace "Harry" with "Potter" and save to a new file
sed 's/Harry/Potter/gi' sample.txt > modified_sample.txt
```

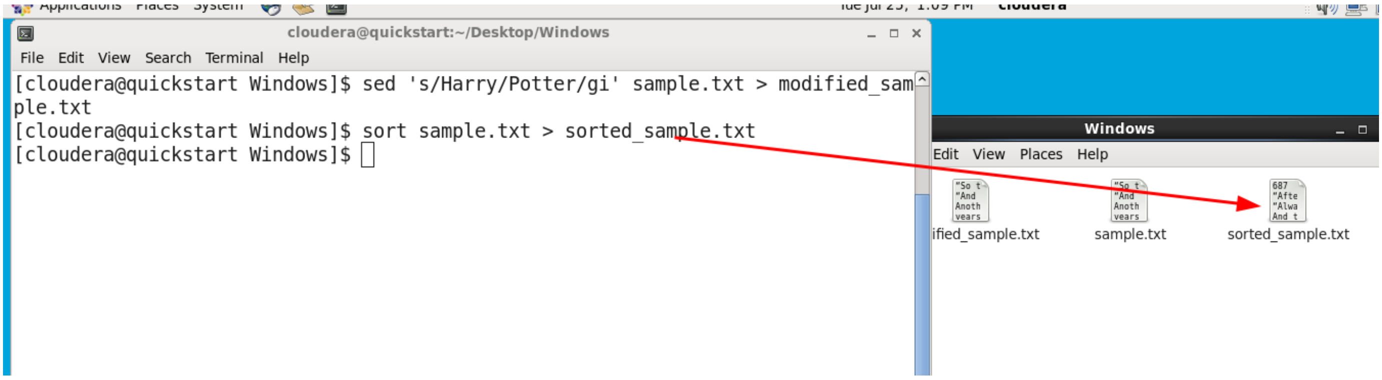

cloudera@quickstart:~/Desktop/Windows
File Edit View Search Terminal Help
Windows
File Edit View Places Help
modified_sample.txt
sample.txt

```
[cloudera@quickstart Windows]$ sed 's/Harry/Potter/gi' sample.txt > modified_sam
ple.txt
[cloudera@quickstart Windows]$ 
```

9. Advanced Challenge - Sort Text:

a) Sort all the lines in the "sample.txt" file in alphabetical order and save the sorted lines to a new file named "sorted_sample.txt".

```
# Sort lines in alphabetical order and save to a new file
sort sample.txt > sorted_sample.txt
```



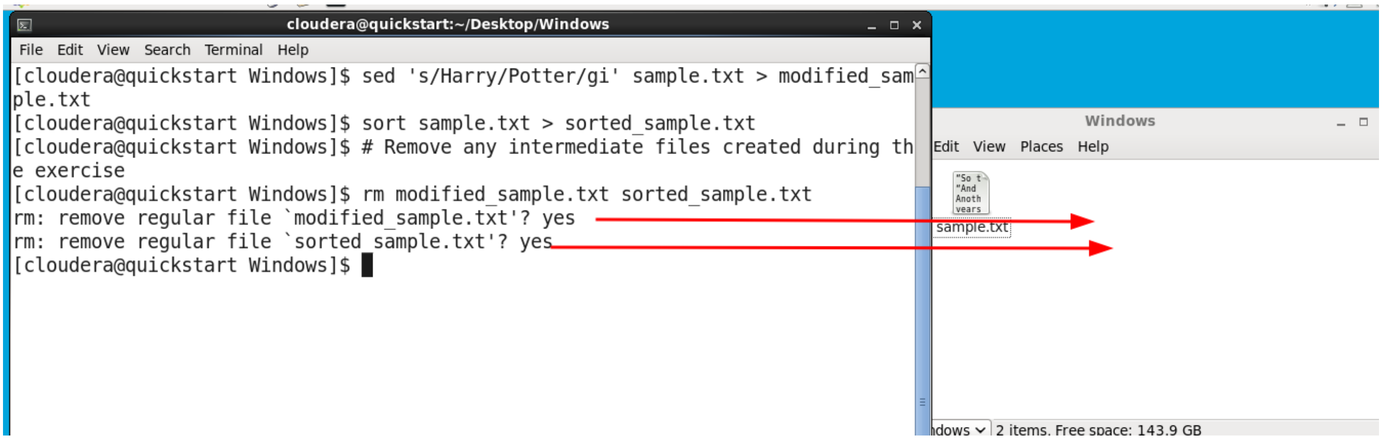
```
[cloudera@quickstart Windows]$ sed 's/Harry/Potter/gi' sample.txt > modified_sample.txt
[cloudera@quickstart Windows]$ sort sample.txt > sorted_sample.txt
[cloudera@quickstart Windows]$ 
```

cloudera@quickstart:~/Desktop/Windows
File Edit View Search Terminal Help
Windows
Edit View Places Help
ified_sample.txt
sample.txt
sorted_sample.txt

10. Clean up:

a) Remove any intermediate files created during the exercise.

```
# Remove any intermediate files created during the exercise
rm modified_sample.txt sorted_sample.txt
```



```
[cloudera@quickstart Windows]$ sed 's/Harry/Potter/gi' sample.txt > modified_sample.txt
[cloudera@quickstart Windows]$ sort sample.txt > sorted_sample.txt
[cloudera@quickstart Windows]$ # Remove any intermediate files created during the exercise
[cloudera@quickstart Windows]$ rm modified_sample.txt sorted_sample.txt
rm: remove regular file `modified_sample.txt'? yes
rm: remove regular file `sorted_sample.txt'? yes
[cloudera@quickstart Windows]$ 
```

cloudera@quickstart:~/Desktop/Windows
File Edit View Search Terminal Help
Windows
Edit View Places Help
sample.txt
ndows
2 items, Free space: 143.9 GB