

Web Scrapping Project - Using BASH

-Ritika Jain

This script is used to scrape daily horoscope data from the website <https://www.vogue.in/horoscope/collection/horoscope> for 12 different sun signs for various dates.

The first line in the bash script creates a directory named sun-signs, and the second line creates 12 directories, one for each sun sign: Aries, Taurus, Gemini, Cancer, Leo, Libra, Scorpio, Sagittarius, Capricorn, Aquarius, Pisces, Virgo.

The script then defines the start and end date of the range as December-05-2022 and December-15-2022, respectively. A **while** loop is then used to create a list of dates in the defined range and save them to a file called **datelist.txt**. The loop starts by **echoing** the current date in the format December-%d-2022 to the terminal and then updating the start value to the next date using the date command. The loop continues until the start is equal to the end. The script then uses a for loop to iterate through the dates in datelist.txt and the **wget** command to download the daily horoscope page for the Aries sun sign on each date. The **-O flag** is used to specify the output file name, which is the date in the format December-%d-2022. Finally, the script again uses a **for loop** in which it uses the **grep** and **sed** commands to extract the horoscope text from each downloaded HTML page and save it to a new file with the same name as the original file.

The **grep** command is used to select the lines of the file that contain the horoscope text, and the **sed** command is used to remove HTML tags and other unwanted text from the selected lines. The resulting horoscope text is saved to a new file with the same name as the original. Then using the **mv command** script, all the text files are moved into the "Aries" directory. The script then repeats the same process for the remaining sun signs: Taurus, Gemini, Cancer, Leo, Libra, Scorpio, Sagittarius, Capricorn, Aquarius, Pisces, and Virgo.

For each sun sign, the script uses a while loop to create a list of dates, a for loop to iterate through the dates and download the corresponding horoscope page for the sun sign, and then again uses a for loop with the **grep** and **sed** commands to extract and save the horoscope text to a new file. Finally, the script uses the **mv** command to move all the processed horoscope files to the appropriate sun sign directory; at the end, all these directories are moved to the "sun-sign" directory created at first using the **mv** command. **Overall, this script automates downloading and processing daily horoscope data for various dates and sun signs. It can be helpful for anyone looking to collect and analyse horoscope data for research or personal use.**

It is worth noting that the script relies on the website to provide the daily horoscope data. If the website changes its structure or the data is unavailable, the script may not work as intended. Additionally, the script uses the **wget** command to download the horoscope pages, which may not work if the website is blocking web scraping or the user's network is blocking the **wget** command. The script also uses the **grep** and **sed** commands to extract and process the horoscope text from the downloaded HTML pages.