Homework Assignment – Regex Practice with Notepad++ and Linux Terminal

Report by Riina Kikkas

Dataset (mydata.txt)

This is line with the the duplicate word.

Today is 11-09-2025 and the weather is sunny.

Visit https://www.google.ee for search.

Käisin eile metsas ja nägin ühte põtra.

The price is 25 euros for the the service.

Another website: http://example.com/test.

Kuupäev on 05-12-2024 ja kellaaeg oli 14:35.

Õunad, pirnid ja ploomid on kõik laual.

Duplicate words are useful: is is, was was.

Check the Estonian portal: https://postimees.ee.

Number sequence: 12345.1, 67890.7.

Öösel paistis kuu eredalt taevas.

Another address: https://news.bbc.com.

The meeting is on 03-07-20234.

Vihm sadas, aga lapsed mängisid õues.

The meeting is on 03-07-2023.

Visit http://tartu.ee for more info.

Sometimes the the repetition is accidental.

Kägu kukkus metsas kolm korda järjest.

The time is 26:59 before the new year.

Õpilased õppisid matemaatikat ja kirjandust.

Date of birth: 29-02-2000.

A website with org: https://example.org.

The the mistake appears again in this line.

Üks väike rõõmus tüdruk jooksis mööda teed.

Another date: 12-25-2022 (Christmas).

The URL https://ttu.ee leads to TalTech.

Lõoke laulis varahommikul rõõmsalt.

Duplicate again: word word in a row.

The price is 99.99 euros.

Check this blog: http://myblog.net.

Pühapäeval käisime vanaema juures külas.

The deadline is 01-01-2026 for submission.

Visit https://haridus.ee for education news.

Tähed särasid taevas nagu väikesed tuled.

This line has the the phrase twice.

Another link: https://openai.com.

The time is 235:902 before the new year.

Jõgi voolas rahulikult ja linnud laulsid.

Meeting date: 10-10-2020.

The portal https://riigikogu.ee has official info.

Ära unusta oma vihikut ja pastakat kooli.

Duplicate sentence with with words.

Another price: 150 euros.

https://example.ee/testpage is another Estonian URL.

Päike tõusis idast ja loojus läänest.

The time is 23:59 before the new year.

Another date: 08-03-2019.

Ülikoolis õppisin programmeerimist ja matemaatikat.

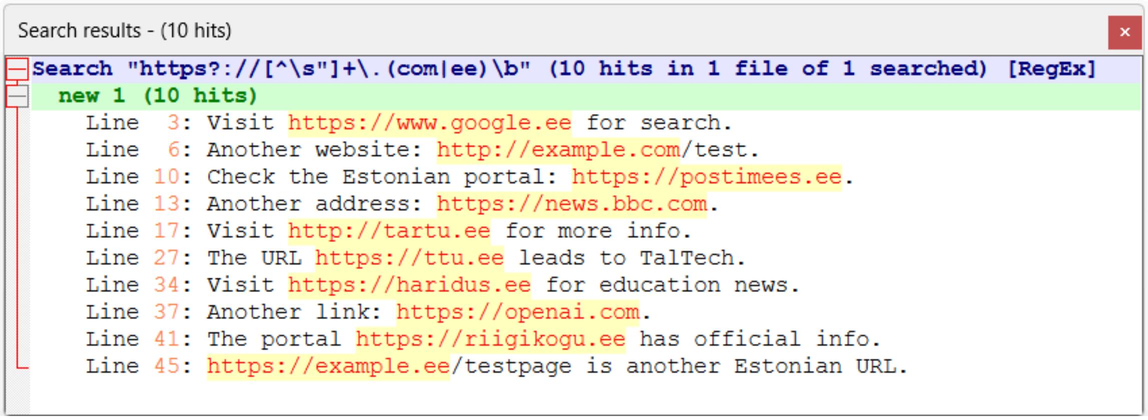
This is the the last line of the dataset.

Tasks in Notepad++ and Linux

In this assignment I practiced using regular expressions both in Notepad++ and in the Linux terminal. I tested different patterns: detecting duplicate words, dates, times, Estonian special characters. I also experimented with replacements such as changing date formats and replacing .com with .ee.

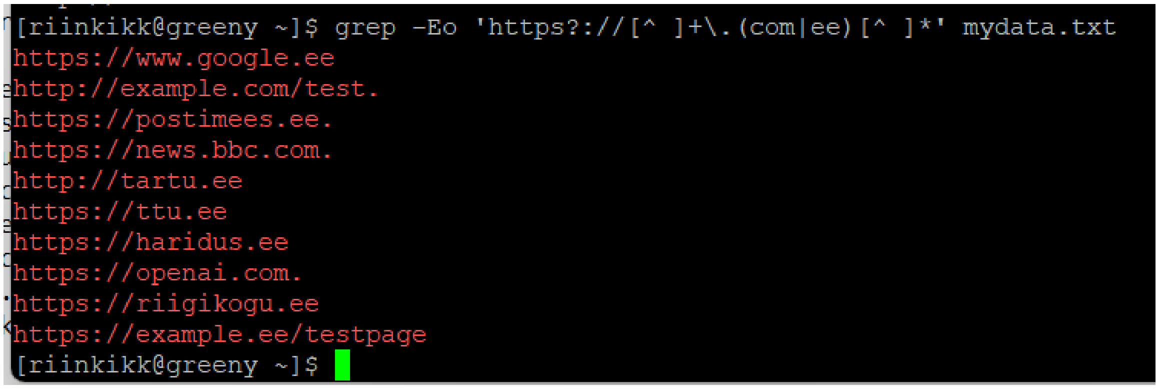
1. **Task: replacing .com with .ee.**

First, I searched for all the web addresses ending with .com or .ee.



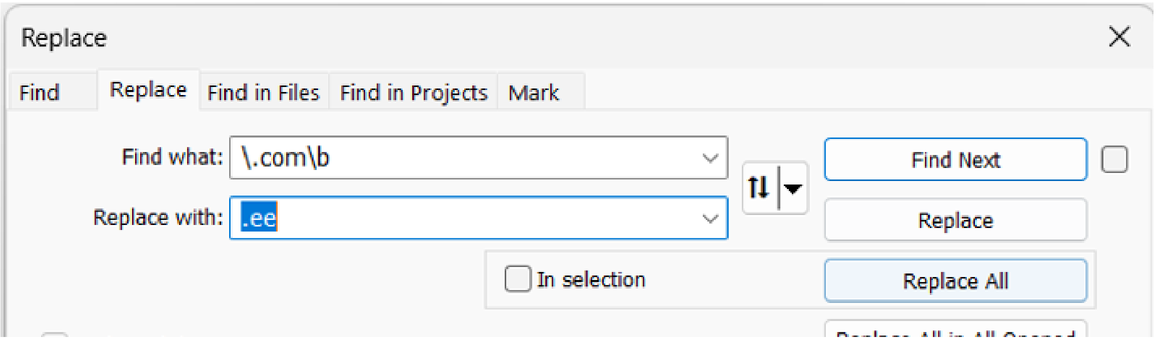
Search results - (10 hits)

```
Search "https?://[^\s"]+\.(com|ee)\b" (10 hits in 1 file of 1 searched) [RegEx]
  new 1 (10 hits)
    Line  3: Visit https://www.google.ee for search.
    Line  6: Another website: http://example.com/test.
    Line 10: Check the Estonian portal: https://postimees.ee.
    Line 13: Another address: https://news.bbc.com.
    Line 17: Visit http://tartu.ee for more info.
    Line 27: The URL https://ttu.ee leads to TalTech.
    Line 34: Visit https://haridus.ee for education news.
    Line 37: Another link: https://openai.com.
    Line 41: The portal https://riigikogu.ee has official info.
    Line 45: https://example.ee/testpage is another Estonian URL.
```



```
[riinkikk@greeny ~]$ grep -Eo 'https?://[^ ]+\.(com|ee)[^ ]*' mydata.txt
https://www.google.ee
http://example.com/test.
https://postimees.ee.
https://news.bbc.com.
http://tartu.ee
https://ttu.ee
https://haridus.ee
https://openai.com.
https://riigikogu.ee
https://example.ee/testpage
[riinkikk@greeny ~]$
```

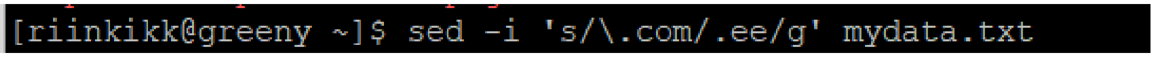
Then I replaced the .com addresses with .ee.



Find what: \.com\b

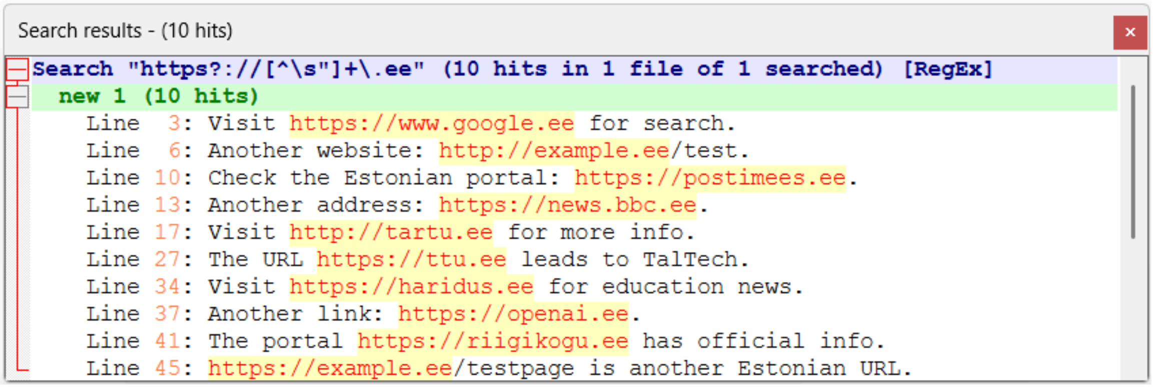
Replace with: .ee

☐ In selection

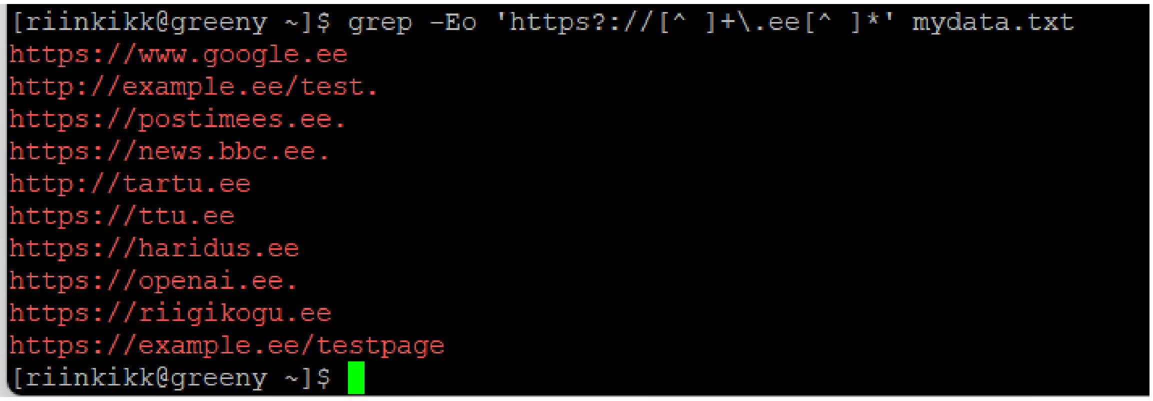


```
[riinkikk@greeny ~]$ sed -i 's/\.com/.ee/g' mydata.txt
```

Finally, I checked again to confirm that all addresses now ended with .ee.



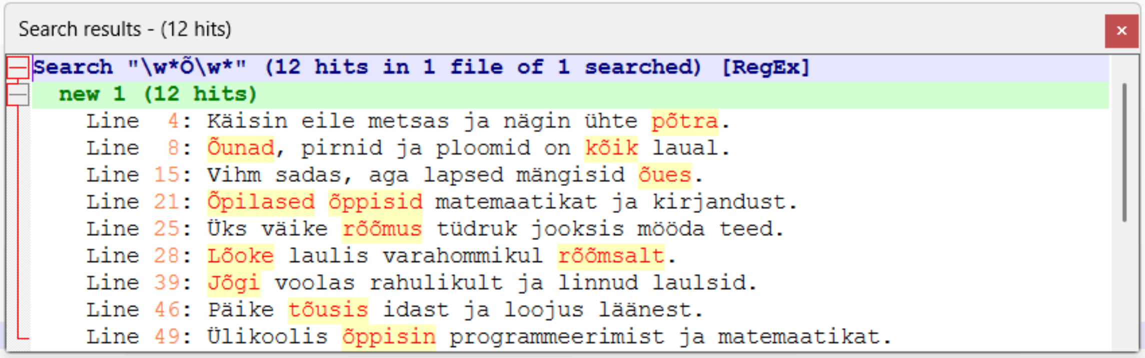
```
Search results - (10 hits)
Search "https?://[^\s"]+\.ee" (10 hits in 1 file of 1 searched) [RegEx]
  new 1 (10 hits)
    Line  3: Visit https://www.google.ee for search.
    Line  6: Another website: http://example.ee/test.
    Line 10: Check the Estonian portal: https://postimees.ee.
    Line 13: Another address: https://news.bbc.ee.
    Line 17: Visit http://tartu.ee for more info.
    Line 27: The URL https://ttu.ee leads to TalTech.
    Line 34: Visit https://haridus.ee for education news.
    Line 37: Another link: https://openai.ee.
    Line 41: The portal https://riigikogu.ee has official info.
    Line 45: https://example.ee/testpage is another Estonian URL.
```



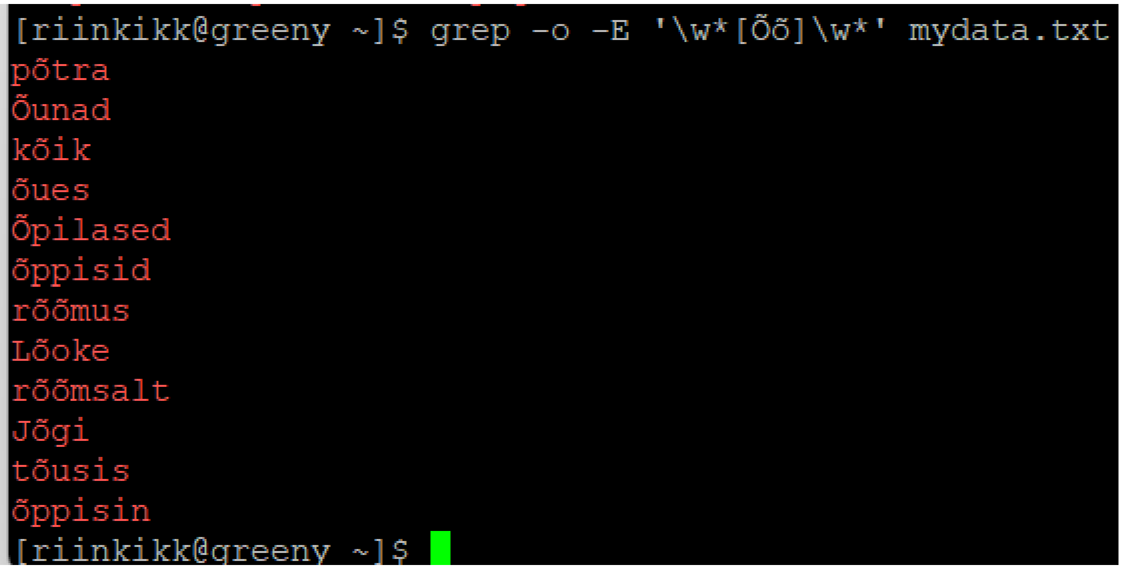
```
[riinkikk@greeny ~]$ grep -Eo 'https?://[^ ]+\.ee[^ ]*' mydata.txt
https://www.google.ee
http://example.ee/test.
https://postimees.ee.
https://news.bbc.ee.
http://tartu.ee
https://ttu.ee
https://haridus.ee
https://openai.ee.
https://riigikogu.ee
https://example.ee/testpage
[riinkikk@greeny ~]$
```

2. Task: deleting all the lines with words containing Õ or õ.

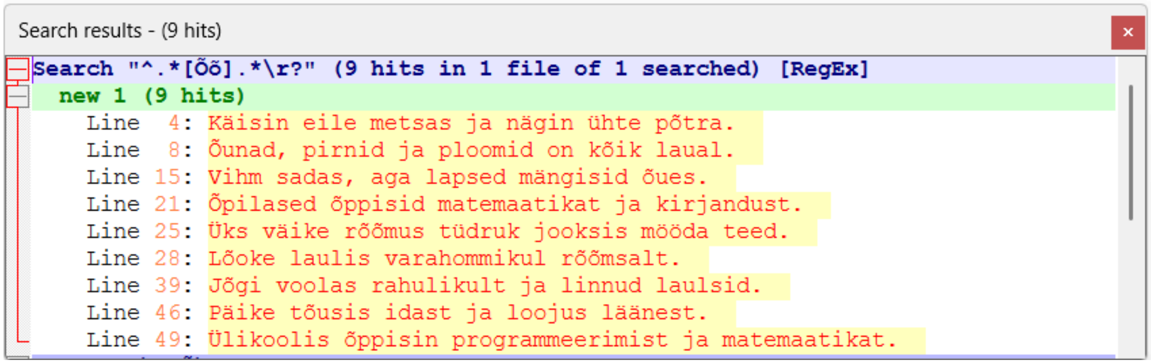
First, I searched for all words and lines containing the letters Õ or õ.



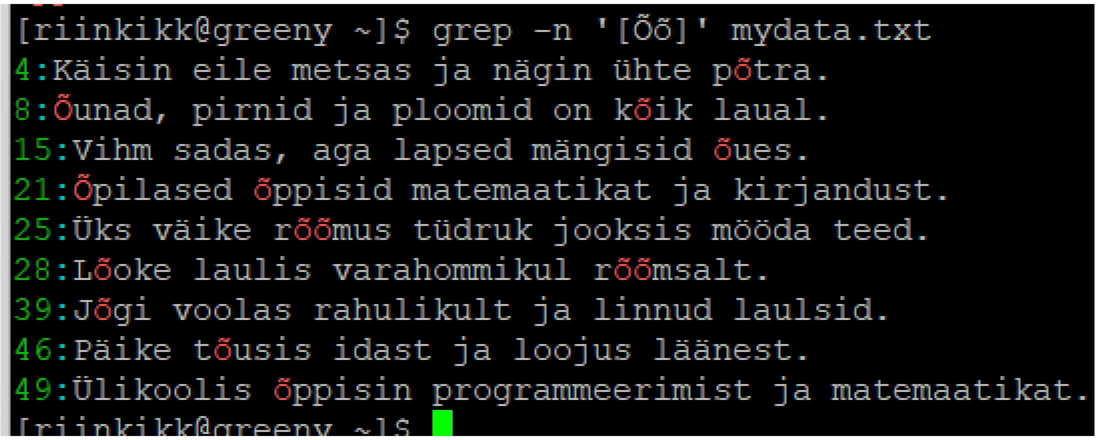
```
Search results - (12 hits)
Search "\w*Õ\w*" (12 hits in 1 file of 1 searched) [RegEx]
  new 1 (12 hits)
    Line  4: Käisin eile metsas ja nägin ühte põtra.
    Line  8: Õunad, pirnid ja ploomid on kõik laual.
    Line 15: Vihm sadas, aga lapsed mängisid õues.
    Line 21: Õpilased õppisid matemaatikat ja kirjandust.
    Line 25: Üks väike rõõmus tüdruk jooksis mööda teed.
    Line 28: Lõoke laulis varahommikul rõõmsalt.
    Line 39: Jõgi voolas rahulikult ja linnud laulsid.
    Line 46: Päike tõusis idast ja loojus läänest.
    Line 49: Ülikoolis õppisin programmeerimist ja matemaatikat.
```



```
[riinkikk@greeny ~]$ grep -o -E '\w*[Õõ]\w*' mydata.txt
põtra
Õunad
kõik
õues
Õpilased
õppisid
rõõmus
Lõoke
rõõmsalt
Jõgi
tõusis
õppisin
[riinkikk@greeny ~]$
```

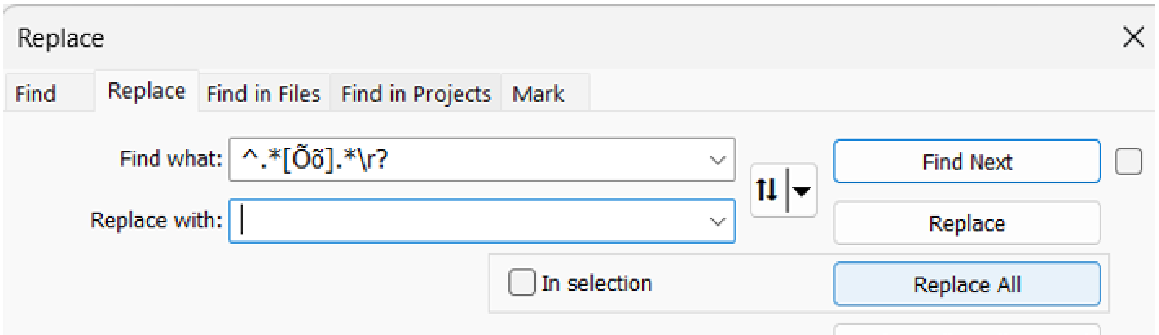


```
Search results - (9 hits)
Search "^.*[Õõ].*\r?" (9 hits in 1 file of 1 searched) [RegEx]
  new 1 (9 hits)
    Line  4: Käisin eile metsas ja nägin ühte põtra.
    Line  8: Õunad, pirnid ja ploomid on kõik laual.
    Line 15: Vihm sadas, aga lapsed mängisid õues.
    Line 21: Õpilased õppisid matemaatikat ja kirjandust.
    Line 25: Üks väike rõõmus tüdruk jooksis mööda teed.
    Line 28: Lõoke laulis varahommikul rõõmsalt.
    Line 39: Jõgi voolas rahulikult ja linnud laulsid.
    Line 46: Päike tõusis idast ja loojus läänest.
    Line 49: Ülikoolis õppisin programmeerimist ja matemaatikat.
```



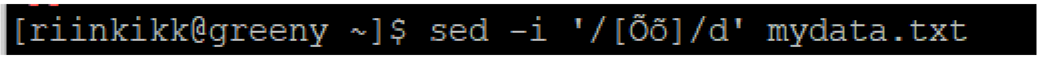
```
[riinkikk@greeny ~]$ grep -n '[Õõ]' mydata.txt
4:Käisin eile metsas ja nägin ühte põtra.
8:Õunad, pirnid ja ploomid on kõik laual.
15:Vihm sadas, aga lapsed mängisid õues.
21:Õpilased õppisid matemaatikat ja kirjandust.
25:Üks väike rõõmus tüdruk jooksis mööda teed.
28:Lõoke laulis varahommikul rõõmsalt.
39:Jõgi voolas rahulikult ja linnud laulsid.
46:Päike tõusis idast ja loojus läänest.
49:Ülikoolis õppisin programmeerimist ja matemaatikat.
[riinkikk@greeny ~]$
```

After identifying these lines, I deleted them by leaving the replace field empty in Notepad++.



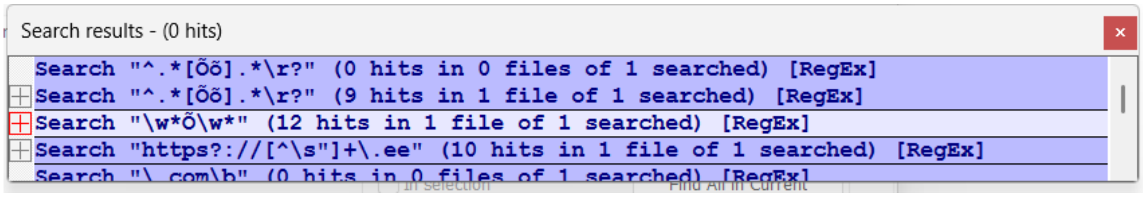
Find what: ^.*[Õõ].*\r?

Replace with:



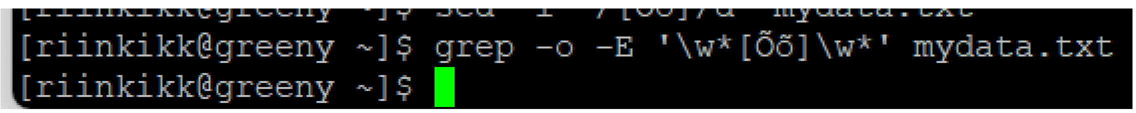
```
[riinkikk@greeny ~]$ sed -i '/[Õõ]/d' mydata.txt
```

Finally, I searched for the letter Õ again to confirm that no matches remained.



Search results - (0 hits)

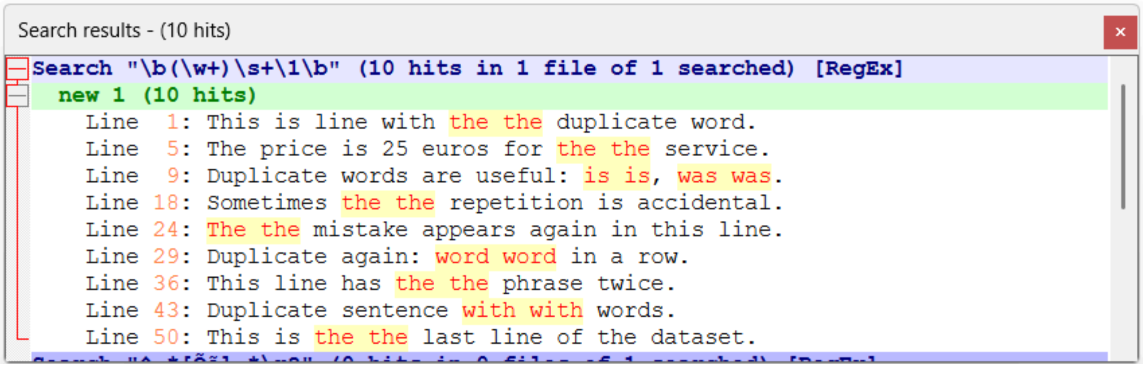
```
Search "^.*[Õõ].*\r?" (0 hits in 0 files of 1 searched) [RegEx]
Search "^.*[Õõ].*\r?" (9 hits in 1 file of 1 searched) [RegEx]
Search "\w*Õ\w*" (12 hits in 1 file of 1 searched) [RegEx]
Search "https?://[^\s"]+\.ee" (10 hits in 1 file of 1 searched) [RegEx]
Search "\.com\b" (0 hits in 0 files of 1 searched) [RegEx]
```



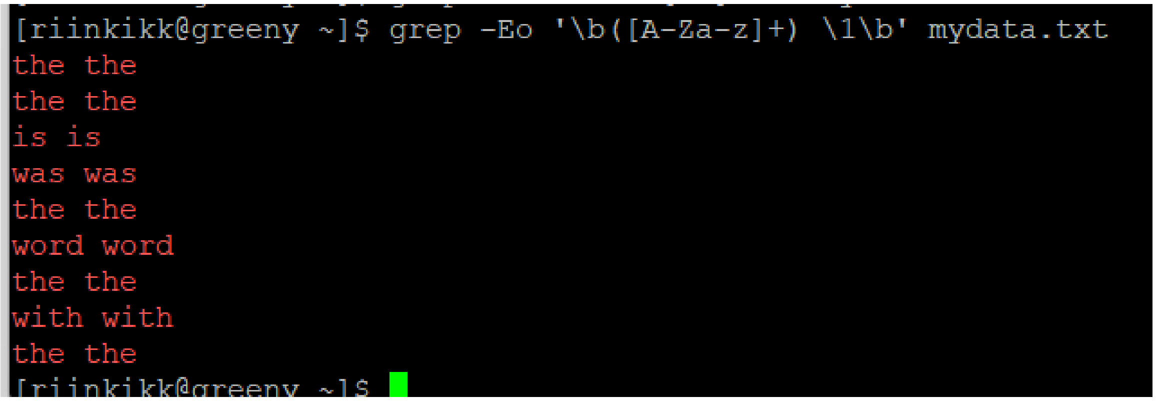
```
[riinkikk@greeny ~]$ sed -i '/[Õõ]/d' mydata.txt
[riinkikk@greeny ~]$ grep -o -E '\w*[Õõ]\w*' mydata.txt
[riinkikk@greeny ~]$
```

3. **Task: Finding all duplicate words and replacing them.**

There were many duplicate words in the file. First, I searched for all duplicate words.

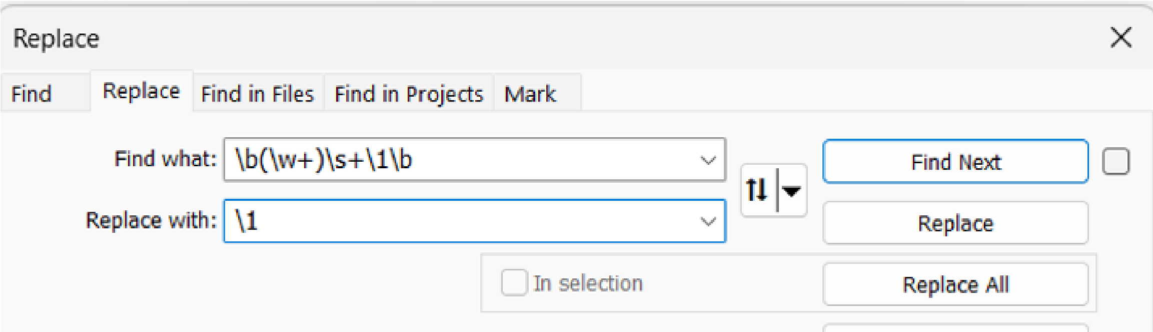


```
Search "\b(\w+)\s+\1\b" (10 hits in 1 file of 1 searched) [RegEx]
  new 1 (10 hits)
    Line  1: This is line with the the duplicate word.
    Line  5: The price is 25 euros for the the service.
    Line  9: Duplicate words are useful: is is, was was.
    Line 18: Sometimes the the repetition is accidental.
    Line 24: The the mistake appears again in this line.
    Line 29: Duplicate again: word word in a row.
    Line 36: This line has the the phrase twice.
    Line 43: Duplicate sentence with with words.
    Line 50: This is the the last line of the dataset.
```



```
[riinkikk@greeny ~]$ grep -Eo '\b([A-Za-z]+) \1\b' mydata.txt
the the
the the
is is
was was
the the
word word
the the
with with
the the
[riinkikk@greeny ~]$
```

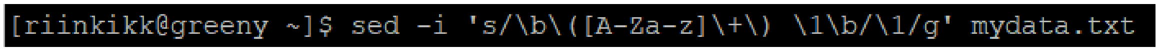
After that, I replaced them by leaving only one.



Find what: \b(\w+)\s+\1\b

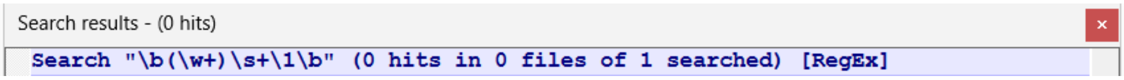
Replace with: \1

□ In selection

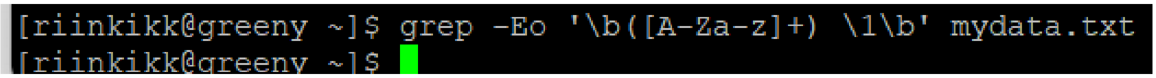


```
[riinkikk@greeny ~]$ sed -i 's/\b\([A-Za-z]\+\) \1\b/\1/g' mydata.txt
```

Finally, I checked again to make sure there were no duplicate words left.



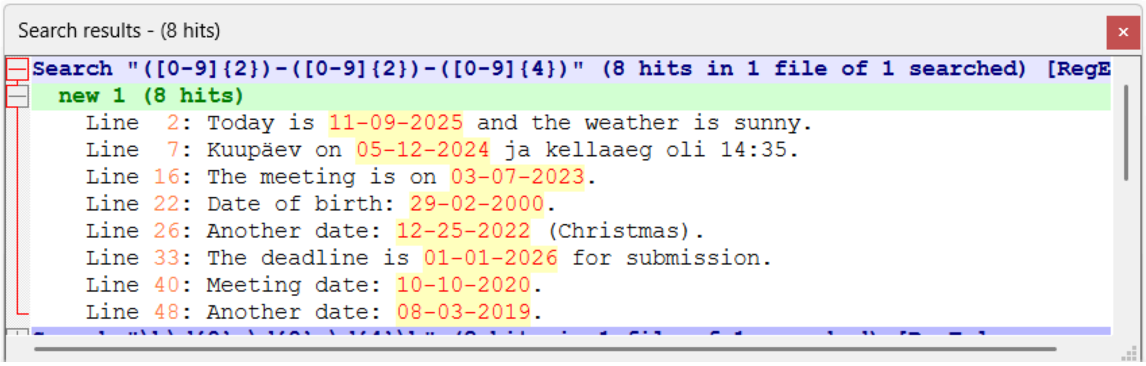
```
Search "\b(\w+)\s+\1\b" (0 hits in 0 files of 1 searched) [RegEx]
```



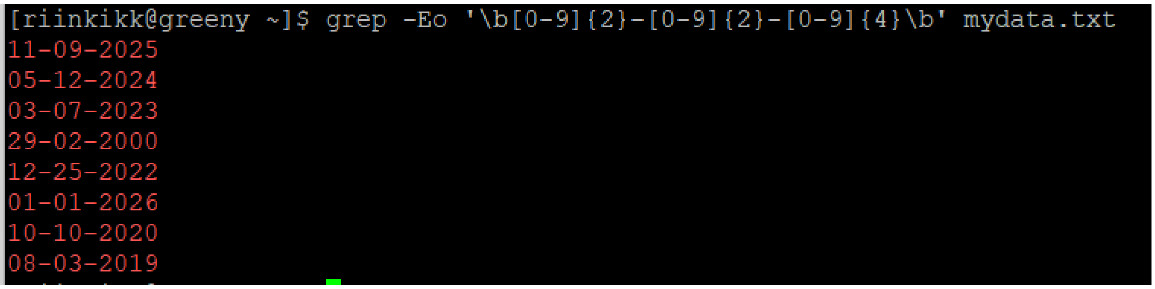
```
[riinkikk@greeny ~]$ grep -Eo '\b([A-Za-z]+) \1\b' mydata.txt
[riinkikk@greeny ~]$
```

4. **Task: Finding dates with format (DD-MM-YYYY) and changing date format (DD-MM-YYYY) to (YYYY-MM-DD).**

First, I found all dates in the format (DD-MM-YYYY).

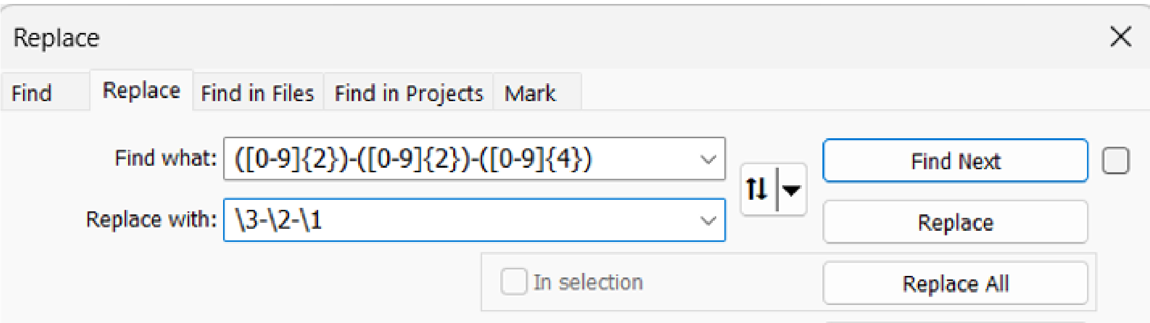


```
Search "([0-9]{2})-([0-9]{2})-([0-9]{4})" (8 hits in 1 file of 1 searched) [RegE
  new 1 (8 hits)
    Line  2: Today is 11-09-2025 and the weather is sunny.
    Line  7: Kuupäev on 05-12-2024 ja kellaaeg oli 14:35.
    Line 16: The meeting is on 03-07-2023.
    Line 22: Date of birth: 29-02-2000.
    Line 26: Another date: 12-25-2022 (Christmas).
    Line 33: The deadline is 01-01-2026 for submission.
    Line 40: Meeting date: 10-10-2020.
    Line 48: Another date: 08-03-2019.
```



```
[riinkikk@greeny ~]$ grep -Eo '\b[0-9]{2}-[0-9]{2}-[0-9]{4}\b' mydata.txt
11-09-2025
05-12-2024
03-07-2023
29-02-2000
12-25-2022
01-01-2026
10-10-2020
08-03-2019
```

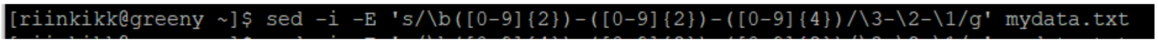
Then I changed their format to (YYYY-MM-DD).



Find what: ([0-9]{2})-([0-9]{2})-([0-9]{4})

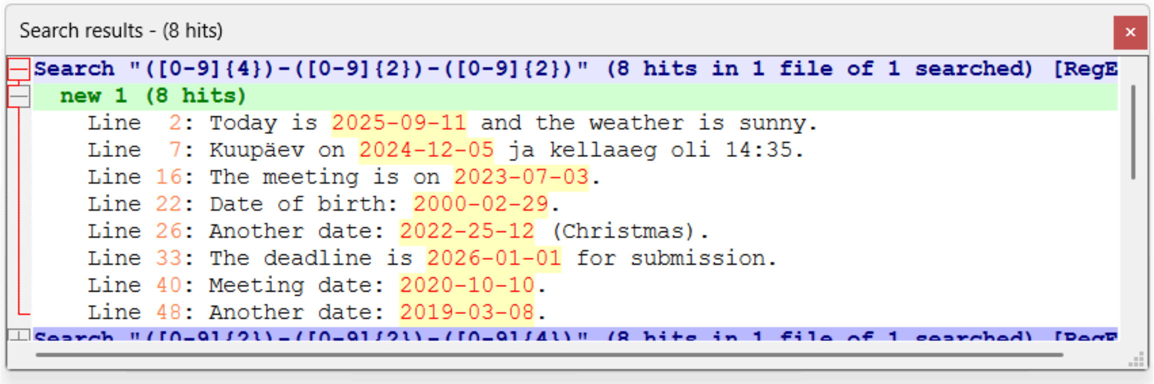
Replace with: \3-\2-\1

☐ In selection

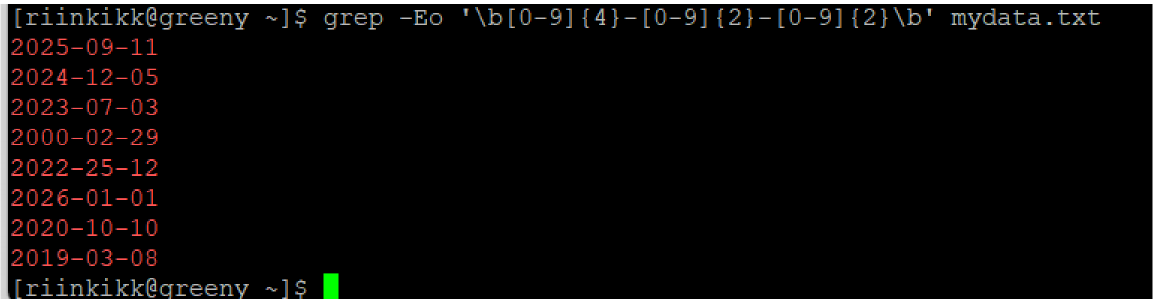


```
[riinkikk@greeny ~]$ sed -i -E 's/\b([0-9]{2})-([0-9]{2})-([0-9]{4})/\3-\2-\1/g' mydata.txt
```

Finally, I checked again to confirm that the dates were now in the new format.



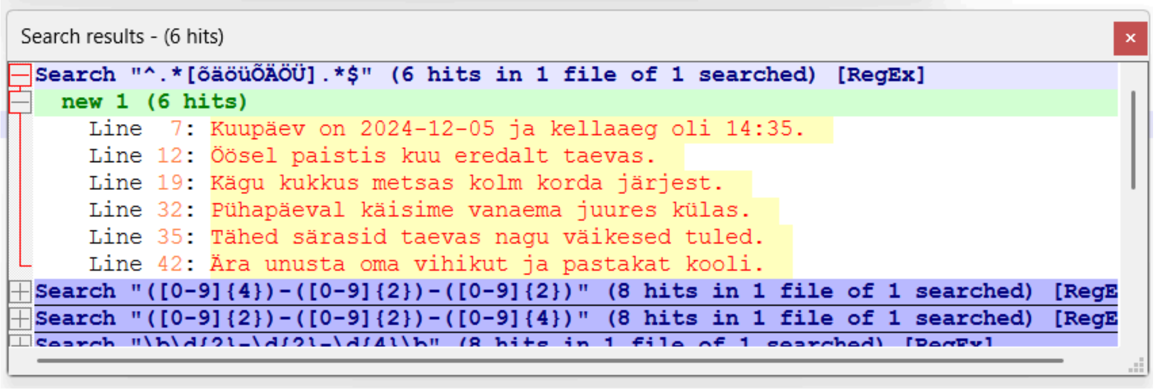
```
Search results - (8 hits)
Search "([0-9]{4})-([0-9]{2})-([0-9]{2})" (8 hits in 1 file of 1 searched) [RegE
  new 1 (8 hits)
    Line  2: Today is 2025-09-11 and the weather is sunny.
    Line  7: Kuupäev on 2024-12-05 ja kellaaeg oli 14:35.
    Line 16: The meeting is on 2023-07-03.
    Line 22: Date of birth: 2000-02-29.
    Line 26: Another date: 2022-25-12 (Christmas).
    Line 33: The deadline is 2026-01-01 for submission.
    Line 40: Meeting date: 2020-10-10.
    Line 48: Another date: 2019-03-08.
Search "([0-9]{2})-([0-9]{2})-([0-9]{4})" (8 hits in 1 file of 1 searched) [RegE
```



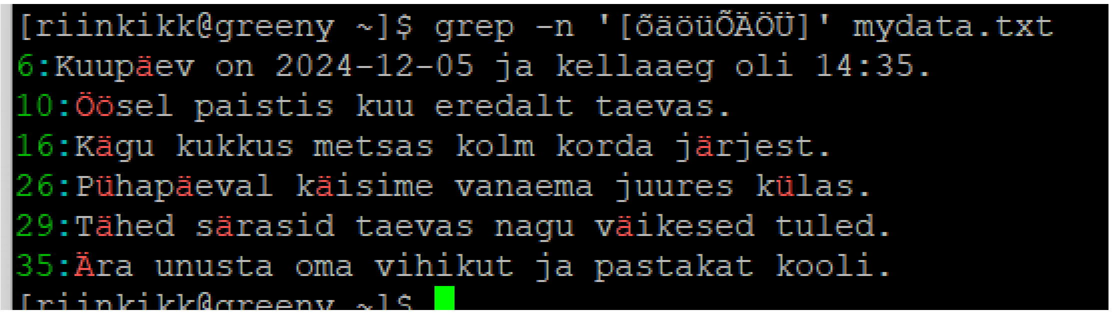
```
[riinkikk@greeny ~]$ grep -Eo '\b[0-9]{4}-[0-9]{2}-[0-9]{2}\b' mydata.txt
2025-09-11
2024-12-05
2023-07-03
2000-02-29
2022-25-12
2026-01-01
2020-10-10
2019-03-08
[riinkikk@greeny ~]$
```

5. **Task: Find all lines with Estonian special letters (õäöüÕÄÖÜ) and mark these with [EST] at the beginning of lines.**

First, I searched for all lines containing Estonian special characters (õäöüÕÄÖÜ).

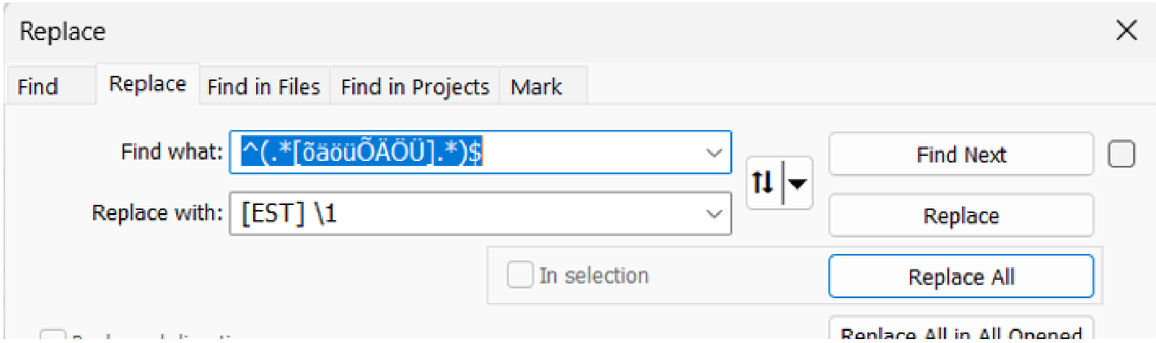


```
Search results - (6 hits)
Search "^.*[õäöüÕÄÖÜ].*$" (6 hits in 1 file of 1 searched) [RegEx]
  new 1 (6 hits)
    Line  7: Kuupäev on 2024-12-05 ja kellaaeg oli 14:35.
    Line 12: Öösel paistis kuu eredalt taevas.
    Line 19: Kägu kukkus metsas kolm korda järjest.
    Line 32: Pühapäeval käisime vanaema juures külas.
    Line 35: Tähed särasid taevas nagu väikesed tuled.
    Line 42: Ära unusta oma vihikut ja pastakat kooli.
Search "([0-9]{4})-([0-9]{2})-([0-9]{2})" (8 hits in 1 file of 1 searched) [RegE
Search "([0-9]{2})-([0-9]{2})-([0-9]{4})" (8 hits in 1 file of 1 searched) [RegE
Search "\b\d{2}-\d{2}-\d{4}\b" (8 hits in 1 file of 1 searched) [RegEx]
```



```
[riinkikk@greeny ~]$ grep -n '[õäöüÕÄÖÜ]' mydata.txt
6:Kuupäev on 2024-12-05 ja kellaaeg oli 14:35.
10:Öösel paistis kuu eredalt taevas.
16:Kägu kukkus metsas kolm korda järjest.
26:Pühapäeval käisime vanaema juures külas.
29:Tähed särasid taevas nagu väikesed tuled.
35:Ära unusta oma vihikut ja pastakat kooli.
[riinkikk@greeny ~]$
```

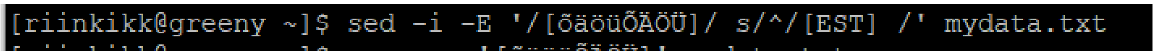
Then I marked these lines by adding [EST] at the beginning.



Find what: ^(.*[õäöüÕÄÖÜ].*)$

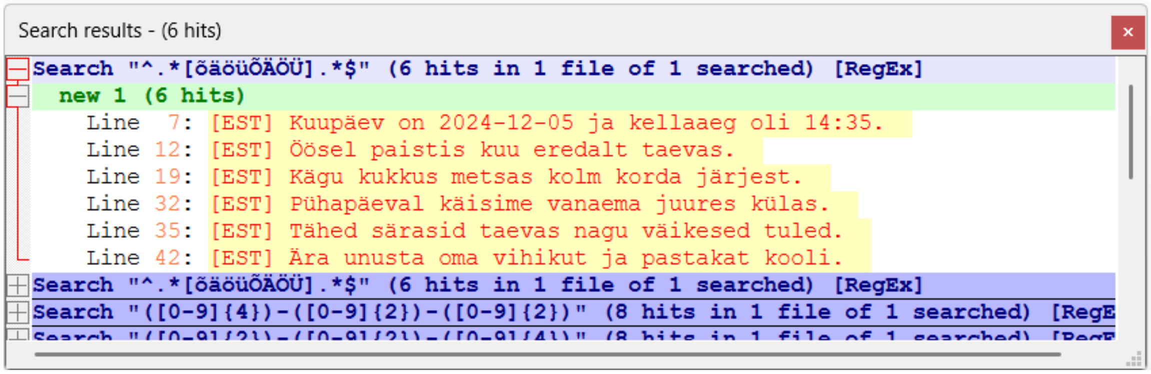
Replace with: [EST] \1

□ In selection

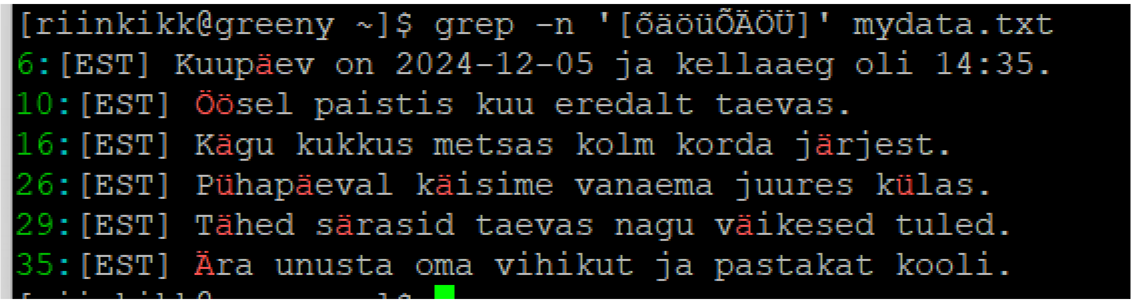


```
[riinkikk@greeny ~]$ sed -i -E '/[õäöüÕÄÖÜ]/ s/^/[EST] /' mydata.txt
```

Finally, I checked again to make sure the markings were applied correctly.



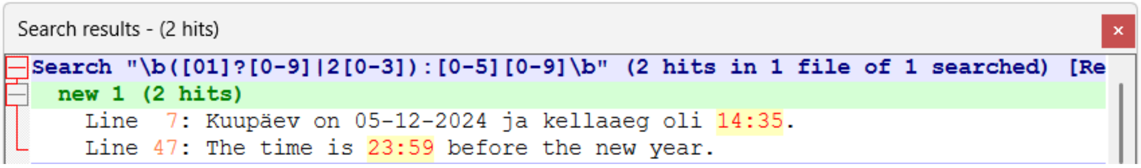
```
Search "^.*[õäöüÕÄÖÜ].*$" (6 hits in 1 file of 1 searched) [RegEx]
  new 1 (6 hits)
    Line  7: [EST] Kuupäev on 2024-12-05 ja kellaaeg oli 14:35.
    Line 12: [EST] Öösel paistis kuu eredalt taevas.
    Line 19: [EST] Kägu kukkus metsas kolm korda järjest.
    Line 32: [EST] Pühapäeval käisime vanaema juures külas.
    Line 35: [EST] Tähed särasid taevas nagu väikesed tuled.
    Line 42: [EST] Ära unusta oma vihikut ja pastakat kooli.
Search "^.*[õäöüÕÄÖÜ].*$" (6 hits in 1 file of 1 searched) [RegEx]
Search "([0-9]{4})-([0-9]{2})-([0-9]{2})" (8 hits in 1 file of 1 searched) [RegE
```



```
[riinkikk@greeny ~]$ grep -n '[õäöüÕÄÖÜ]' mydata.txt
6:[EST] Kuupäev on 2024-12-05 ja kellaaeg oli 14:35.
10:[EST] Öösel paistis kuu eredalt taevas.
16:[EST] Kägu kukkus metsas kolm korda järjest.
26:[EST] Pühapäeval käisime vanaema juures külas.
29:[EST] Tähed särasid taevas nagu väikesed tuled.
35:[EST] Ära unusta oma vihikut ja pastakat kooli.
```

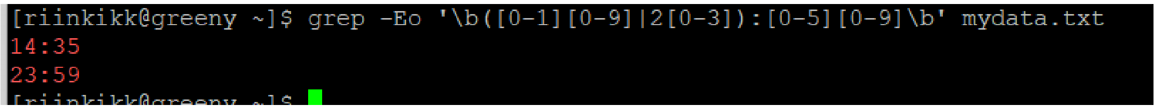
6. **Task: Find times (HH:MM).**

There were some invalid time values in the dataset, such as 26:59 and 235:902. Therefore, I had to create a regex that only matched valid times in the format HH:MM, ranging from 00:00 to 23:59.



Search results - (2 hits)

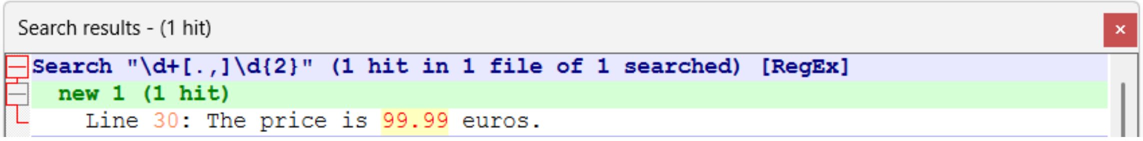
```
Search "\b([01]?[0-9]|2[0-3]):[0-5][0-9]\b" (2 hits in 1 file of 1 searched) [Re
  new 1 (2 hits)
    Line  7: Kuupäev on 05-12-2024 ja kellaaeg oli 14:35.
    Line 47: The time is 23:59 before the new year.
```



```
[riinkikk@greeny ~]$ grep -Eo '\b([0-1][0-9]|2[0-3]):[0-5][0-9]\b' mydata.txt
14:35
23:59
[riinkikk@greeny ~]$
```

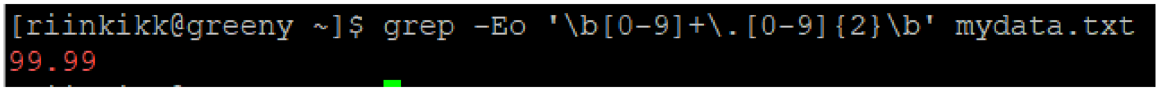
7. **Task: Find amounts with exactly two decimals.**

The dataset also included numbers with other decimal places (for example, 12345.1 or 67890.7). However, I only searched for numbers with exactly two decimal places, such as 99.99.



Search results - (1 hit)

```
Search "\d+[.,]\d{2}" (1 hit in 1 file of 1 searched) [RegEx]
  new 1 (1 hit)
    Line 30: The price is 99.99 euros.
```



```
[riinkikk@greeny ~]$ grep -Eo '\b[0-9]+\.[0-9]{2}\b' mydata.txt
99.99
```