

Regular Expressions

```
In [2]: text = """03-02-2025 10:22pm: My email address is testuser@gmail.com and I check
        it often. You should send me an email!"""
```

We want to extract the email address from this string.

But let's start small.

```
In [3]: haystack = "Date: 03-02-2025"
```

```
In [4]: needle = "Da"
```

```
In [5]: import re
```

```
In [6]: re.match(needle, haystack)
```

```
Out[6]: <_sre.SRE_Match object; span=(0, 2), match='Da'>
```

```
In [7]: match = re.match(needle, haystack)
```

```
In [8]: match.group(0)
```

```
Out[8]: 'Da'
```

```
In [12]: needle = "Date: \d\d"
```

```
In [13]: res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

```
Found!
Date: 03
```

```
In [14]: needle = ".....\d"
res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

```
Found!
Date: 0
```

```
In [17]: needle = ".....\s\d\d"
res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

```
Found!
Date: 03
```

Beyond Basic Programming - Intermediate Python

recluze.net/learn

```
In [18]: needle = ".{5}\\s\\d{2}"
res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

Found!
Date: 03

```
In [19]: needle = ".*\\d{2}"
res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

Found!
Date: 03-02-2025

But that's not what we expected!

The problem is that `.*` is matching everything and `d{2}` matches the last two digits!

```
In [20]: needle = ".*?\\d{2}"
res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

Found!
Date: 03

The `?` means that it should not be **greedy**. As soon as `\\d{2}` can be satisfied, the effect of `.*` should be stopped.

Problem: This is already getting too complex!

Oh, but regular expressions look **ugly**!

That's because you need to approach them in a modular way, just as we break down our whole program into functions.

```
In [21]: str_date = ".*?"          # You can also use: ".*\\s"
str_day  = "\\d{2}"

# needle = ".*?\\d{2}"
needle = str_date + str_day

res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

Found!
Date: 03

Let's take the more complicated string.

Beyond Basic Programming - Intermediate Python

recluze.net/learn

```
In [22]: haystack = text
         print(haystack)
```

```
03-02-2025 10:22pm: My email address is testuser@gmail.com and I check
                  it often. You should send me an email!
```

```
In [23]: str_day = "\d{2}-\d{2}-\d{4}"

         str_time = "\d{2}:\d{2}pm"

         needle = str_day + "\s" + str_time

         res = re.match(needle, haystack)
         if res:
             print("Found!")
             print(res.group(0))
         else:
             print("Not found!")
```

```
Found!
03-02-2025 10:22pm
```

But what if the time is in the morning!

```
In [28]: haystack = """03-02-2025 10:22PM: My email address is testuser@gmail.com and I check
                  it often. You should send me an email!"""

         str_day = "\d{2}-\d{2}-\d{4}"

         str_time = "\d{2}:\d{2}[apAP][mM]"

         needle = str_day + "\s" + str_time

         res = re.match(needle, haystack)
         if res:
             print("Found!")
             print(res.group(0))
         else:
             print("Not found!")
```

```
Found!
03-02-2025 10:22PM
```

Let's find that email!

```
In [29]: haystack = """03-02-2025 10:22am: My email address is testuser@gmail.com and I check
                  it often. You should send me an email!"""
```

Beyond Basic Programming - Intermediate Python

recluze.net/learn

```
In [30]: str_prefix = ".*"

str_username = "[a-zA-Z0-9.]*"
str_domain = ".*\..*?" # "[a-zA-Z0-9_]*"
# ".*\..*?"

str_email = str_username + "@" + str_domain

needle = str_prefix + "\s" + str_email + "\s"

res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

Found!

03-02-2025 10:22am: My email address is testuser@gmail.com

```
In [31]: # If we wrote this as one singular RegExp:

print(needle) # Ouch!
```

.*\s[a-zA-Z0-9.]*@\..*?\s

That's the stuff of nightmares!

Extraction

```
In [50]: str_prefix = ".*"

str_username = "([a-zA-Z0-9_].*)"
str_domain = "(.*\..*?)"

str_email = "(" + str_username + "@" + str_domain + ")" # Add ( ) around stuff you need

needle = str_prefix + "\s" + str_email + "\s"

res = re.match(needle, haystack)
if res:
    print("Found!")
    print(res.group(0))
else:
    print("Not found!")
```

Found!

03-02-2025 10:22am: My email address is testuser@gmail.com

```
In [54]: res.group(3)
```

```
Out[54]: 'gmail.com'
```

```
In [55]: print(needle)
```

.*\s(([a-zA-Z0-9_].*)@(.*\..*?))\s

```
In [ ]:
```