

Jonathan Soma  
Knight Chair in Data Journalism  
Columbia Journalism School  
js4571@columbia.edu

***Visit [bit.ly/ds-dojo2024](https://bit.ly/ds-dojo2024)***

# ***Basic concepts***

**replace and split are  
99% of the game**

66 years old

replace "years old"  
with... nothing!

~~66 years old~~

convert to number

66

j soma, js4571@columbia.edu, @dangerscarf

split on “,”

j soma | js4571@columbia.edu | @dangerscarf

take second piece

js4571@columbia.edu

j soma, js4571@columbia.edu, @dangerscarf

j soma | js4571@columbia.edu | @dangerscarf

split

take second

js4571@columbia.edu

split on @

replace with nothing

~~js4571@columbia.edu~~

take second

js4571 | columbia.edu

columbia.edu

I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a Google Sheet and I want to split it by ", ", take the second part, then replace the "@columbia.edu" with nothing. Just give me the formula.

**just ask!  
then TEST IT**



You can use the following Google Sheets formula to achieve this:

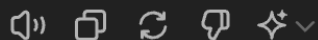
plaintext

Copy code

```
=SUBSTITUTE(INDEX(SPLIT(A1, ", "), 2), "@columbia.edu", "")
```

Explanation:

1. ``SPLIT(A1, ", ")`` splits the text in cell ``A1`` by ", ".
2. ``INDEX(SPLIT(A1, ", "), 2)`` selects the second part of the split text.
3. ``SUBSTITUTE(INDEX(SPLIT(A1, ", "), 2), "@columbia.edu", "")`` removes the "@columbia.edu" part from the selected text.




I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a Google Sheet and I want to extract the "@columbia.edu" part. Write me a formula to do this. Only show me the formula.

**just ask!  
then TEST IT**



excel

 Copy code

```
=RIGHT(A1,LEN(A1)-FIND("@",A1)+1)
```



Feel free to reach out at 555-555-5555



take the final 12 letters

555-555-5555



**just ask!  
then TEST IT**



I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a pandas dataframe column named "details" and I want to split it by ",", take the second part, then replace the "@columbia.edu" with nothing. Try not to use .apply. Don't create the df, just give me the code.

< 6/8 >



Certainly! Here's the code to achieve that:

python

Copy code

```
# Split the 'details' column by ", " and take the second part
df['email'] = df['details'].str.split(', ').str[1]

# Replace "@columbia.edu" with nothing
df['email'] = df['email'].str.replace('@columbia.edu', '', regex=False)
```

This will split the column and then replace the domain part as required.




just ask!  
then TEST IT

I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a pandas dataframe column named "details" and I want to extract the "@columbia.edu" part. Try not to use .apply. Don't create the df, just give me the code.



You can use the `str.extract` method to extract the "@columbia.edu" part from the "details" column. Here's the code to do that:

python

 Copy code

```
df['extracted'] = df['details'].str.extract(r'(\b@\w+\b\.edu\b)')
```



# ***Regular expressions***

**regex is the other  
other 1%**

66 years old

hug what we want  
(capture group)



(66) years old

match anything



(.\*) years old

I am 66 years old

hug what we want



I am (66) years old

match anything



I am (.\*?) years old

Contact me at 555-555-5555

hug what we want

Contact me at (.\*)

but maybe not!

Contact me at 555-555-5555 or  
js4571@columbia.edu

Contact me at (\d\d\d-\d\d\d-\d\d\d\d)

just numbers

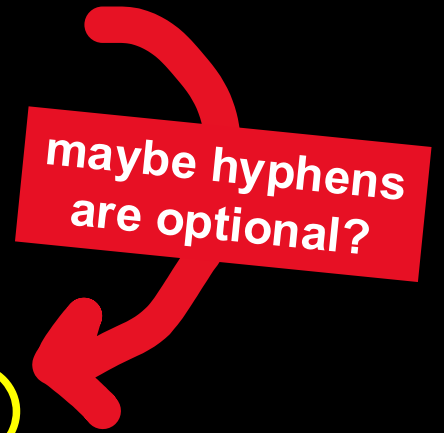
Contact me at 555-555-5555



maybe we don't  
need an anchor?

(\d\d\d-\d\d\d-\d\d\d\d)

Contact me at 5555555555



(\d\d\d-?\d\d\d-?\d\d\d\d)



Contact me at 555+555+5555



any character is fine!  
(and also optional)

(\d\d\d.\?\d\d\d.\?\d\d\d\d)

I have a table that looks like this. Write the SIMPLEST POSSIBLE regex to just extract the values for "Deforested area" in km. Maybe anchor around the year at the start of each line?

just ask!  
then TEST IT

Year	Deforestation of the Amazon rainforest	Deforested area	Accumulated loss
2005	10,311 km2 (7,341 sq mi)	332,470 km2 (128,370 sq mi)	
2006	10,311 km2 (5,516 sq mi)	346,756 km2 (133,883 sq mi)	
2007	10,311 km2 (4,498 sq mi)	358,407 km2 (138,382 sq mi)	
2008	10,311 km2 (4,985 sq mi)	371,318 km2 (143,367 sq mi)	
2009	7,464 km2 (2,882 sq mi)	378,782 km2 (146,249 sq mi)	
2010	7,000 km2 (2,700 sq mi)	385,782 km2 (148,951 sq mi)	
2011	6,418 km2 (2,478 sq mi)	392,200 km2 (151,400 sq mi)	
2012	4,571 km2 (1,765 sq mi)	396,771 km2 (153,194 sq mi)	
2013	5,891 km2 (2,275 sq mi)	402,662 km2 (155,469 sq mi)	
2014	4,848 km2 (1,872 sq mi)	407,510 km2 (157,340 sq mi)	

< 5/5 >



To extract the "Deforested area" values in km<sup>2</sup> using a regex, you can use the following simple pattern:

SCSS

Copy code

```
^\d{4}\s+(\d{1,5})\s+km2
```

Here's a breakdown of the regex:

- `^` asserts the position at the start of the line.`
- `\d{4}` matches exactly four digits (the year).`
- `\s+` matches one or more whitespace characters.`
- `(\d{1,5})` captures between 1 and 5 digits (the deforested area value) in a capturing group.`
- `\s+` matches one or more whitespace characters.`
- `km2` matches the literal string "km2".`

This pattern will capture the deforested area in km<sup>2</sup> from each line.



Jonathan Soma  
Knight Chair in Data Journalism  
Columbia Journalism School  
js4571@columbia.edu

***Visit [bit.ly/ds-dojo2024](https://bit.ly/ds-dojo2024)***