

```
import numpy as np
import pandas as pd
import re
import nltk
import spacy
import string
pd.options.mode.chained_assignment
```

```
full_df = pd.read_csv("/content/
df = full_df[["text"]]
df["text"] = df["text"].astype(s
full_df.head()
```



	tweet_id	author_id
0	119237	105834
1	119238	ChaseSupport
2	119239	105835
3	119240	VirginTrains
4	119241	105836

Next
steps:

[code](#) [full_df](#)



Lower Casing

```
df["text_lower"] = df["text"].str.lower()
df.head()
```

LookupError

NameError

error X



You

Please explain the error:

error: nothing to repeat at position 1



Gemini

Hypothesis

The error "nothing to repeat at position 1" usually arises in regular expressions when a quantifier (+, *, ?, etc.) is used without a preceding character or group to quantify.

Looking at the code and the error message, the issue seems to be with the `emot` variable in the for loop. Some of these emojis might be special characters or have special meanings in regex, and when they are directly inserted into the regex pattern, they are interpreted as quantifiers or other regex tokens, leading to the "nothing to repeat" error.

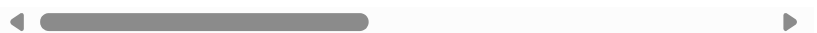
Suggested Changes



```
!pip install emot
import re
from emot.emot_unicode import UNICODE_EMOJI

def convert_emojis(text):
    for emot in UNICODE_EMOJI:
        # Escape special characters in emot using re.escape
        text = re.sub(r'(' + re.escape(emot) + ')', "_".jo
    return text
```

```
text = "I am 🤔"
convert_emojis(text)
```



[Use code with caution](#)

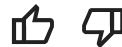
Sources

1. stackoverflow.com/questions/31386552/nothing-to-repeat-from-python-regex
2. stackoverflow.com/questions/3675144/regex-error-nothing-to-repeat



	text	text_1
0	@AppleSupport causing the reply to be disregar...	@applesu causin reply disre
1	@105835 Your business means a lot to us. Pleas...	@105835 bus means a us. pl
2	@76328 I really hope you all change but I'm su...	@76 really hop all chang i'm

Rate this answer



Next steps:

code df

reco

Removal of Punctuations

```
PUNCT_TO_REMOVE = string.punctuation
def remove_punctuation(text):
    """custom function to remove punctuation"""
    return text.translate(str.maketrans('', '', PUNCT_TO_REMOVE))
```

```
df["text_wo_punct"] = df["text"].apply(remove_punctuation)
df.head()
```




	text	text_1
0	@AppleSupport causing the reply to be disregar...	@applesu causin reply disre
1	@105835 Your business means a lot to us. Pleas...	@105835 bus means a us. pl
2	@76328 I really hope you all change but I'm su...	@76 really hop all chang i'm

Next steps:

code df

reco

```
import nltk
nltk.download('stopwords')
```



```
[nltk_data] Downloading package
[nltk_data] Package stop
'i, me, my, myself, we, ou
r, ours, ourselves, you, y
ou're, you've, you'll, yo
u'd, your, yours, yoursel
f, yourselves, he, him, hi
s, himself, she, she's, he
r, hers, herself, it, i
t's, its, itself, they, th
em, their, theirs, themsel
ves, what, which, who, who
m, this, that, that'll, th
ese. those. am. is. are. w
```

```
df["text_wo_stop"] = df["text"].str.replace(stopwords, "", regex=False)
df.head()
```

	text	text_1
0	@AppleSupport causing the reply to be disregar...	@applesu causin reply disre
1	@105835 Your business means a lot to us. Pleas...	@105835 bus means a us. pl
2	@76328 I really hope you all change but I'm su...	@76 really hop all chang i'm
3	@105836 LiveChat is online at the moment - htt...	@10 livec online a moment ·

Next
steps:

code df

 recd

Removal of Frequent words

```
from collections import Counter
cnt = Counter()
for text in df["text_wo_stop"].
    for word in text.split():
        cnt[word] += 1
```

```
cnt.most_common(10)
```

```
→ [('I', 34),
    ('us', 25),
    ('DM', 19),
    ('help', 17),
    ('httpstcoGDrqU22YpT',
     12),
    ('AppleSupport', 11),
    ('Thanks', 11),
    ('phone', 9),
    ('Hi', 8),
    ('get', 8)]
```

```
FREQWORDS = set([w for (w, wc)
def remove_freqwords(text):
    """custom function to remove frequent words
    return " ".join([word for word in text.split() if word not in FREQWORDS])
```

```
df["text_wo_stopfreq"] = df["text_wo_stop"].apply(remove_freqwords)
df.head()
```



	text	text_1
0	@AppleSupport causing the reply to be disregar...	@applesu causin reply disre
1	@105835 Your business means a lot to us. Pleas...	@105835 bus means a us. pl
2	@76328 I really hope you all change but I'm su...	@76 really hop all chang i'm
3	@105836 LiveChat is online at the moment - htt...	@10 livec online : moment ·
4	@VirginTrains see attached error message. I've...	@virgint see atta error mes

Next
steps:

[code](#) [df](#)

[rec](#)

Removal of Rare words

```
df.drop(columns=["text_wo_stop"]
df.head()
```



	text	text_1
0	@AppleSupport causing the reply to be disregar...	@applesu causin reply disre
1	@105835 Your business means a lot to us. Pleas...	@105835 bus means a us. pl
2	@76328 I really hope you all change but I'm su...	@76 really hop all chang i'm
3	@105836 LiveChat is online at the moment - htt...	@10 livec online : moment -

Next
steps:

[code](#) [df](#)

[rec](#)

```
# Drop the two columns which are
df.drop(["text_wo_punct"], axis=1)
```

```
n_rare_words = 10
RAREWORDS = set([w for (w, wc)
def remove_rarewords(text):
    """custom function to remove rare words"""
    return " ".join([word for word in text.split() if word not in RAREWORDS])
```

```
df["text_wo_stopfreqrare"] = df["text"].apply(remove_rarewords)
df.head()
```



	text	text_1
0	@AppleSupport causing the reply to be disregar...	@applesu causin reply disre
1	@105835 Your business means a lot to us. Pleas...	@105835 bus means a us. pl
2	@76328 I really hope you all change but I'm su...	@76 really hop all chang i'm
3	@105836 LiveChat is online at the moment - htt...	@10 livec online : moment ·
4	@VirginTrains see attached	@virgint see atta

Next
steps:

[code](#) [df](#)

[rec](#)

Lemmatization

```
!pip install nltk # Make sure I
import nltk
```

```
nltk.download('wordnet') # Down
```

```
from nltk.stem import WordNetLe
```

```
lemmatizer = WordNetLemmatizer(
def lemmatize_words(text):
    return " ".join([lemmatizer
```

```
df["text_lemmatized"] = df["te)
df.head()
```

Requirement already satisfied:
 Requirement already satisfied:
 Requirement already satisfied:
 Requirement already satisfied:
 Requirement already satisfied:
 [nltk_data] Downloading package

	text	text_1
0	@AppleSupport causing the reply to be disregar...	@applesu causin reply disre
1	@105835 Your business means a lot to us. Pleas...	@105835 bus means a us. pl
2	@76328 I really hope you all change but I'm su...	@76 really hop all chang i'm
3	@105836 LiveChat is online at the moment - htt...	@10 livec online ; moment ·
4	@VirginTrains see attached error message. I've...	@virgint see atta error mes

Next
steps:

code df

rec

```
lemmatizer.lemmatize("swiming")
```

→ 'swim'

```
lemmatizer.lemmatize("swiming")
```

→ 'swiming'

Removal of Emojis

```
def remove_emoji(string):
    emoji_pattern = re.compile(
        u"
        u"
```


u"
u"
u"
u"
"]

return emoji_pattern.sub(

remove_emoji("I am 🍌")

↔ 'I am '

Conversion of Emoticon to Words

Enter a prompt here

▶

0 / 400

Responses may display inaccurate or offensive information that doesn't represent Google's views. [Learn more](#)