

```

import time
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
import re
import string

```

```

df1 = pd.read_csv("white_fungus.csv").iloc[:, 1:]
df2 = pd.read_csv("yellow_fungus.csv").iloc[:, 1:]

```

```

df = df1.append(df2, ignore_index = True)


```

```

df

```

		Tweets \
0	The Real White FUNGUS □ <a href="https://t.co/Kr79gJiowr">https://t.co/Kr79gJiowr</a>	
1	@Naseebkhan7223 White and black fungus is bett...	
2	Dear PMO Situation is getting worse & can...	
3	Corona , black , white fungus now aspergillosi...	
4	Are 0 12th walo; Covid, Black, White, Yellow f...	
...		...
18537	"Yellow-white fungus grows inside the cicadas,...	
18538	@gayatri008_16 Nooo.. in telugu states it's YE...	
18539	Yellow fungus is deadlier than both white &...	
18540	Fungus is clearly a racist Virus.\nWhy only Wh...	
18541	before- black fungus\nnow- white fungus\n\nnaft...	

	User	User_statuses_count
user_followers \		
0	LoGaN	1753
99		
1	Srinivas Bajju	364
10		
2	Ishu NSUI YAMUNANAGAR ( 2.3k)	58251
2334		
3	sagarriyahanhai	84
13		
4	Ankit Sinha 	200
71		
...	...	...
...		
18537	Diana Hussein	92840
10388		
18538	CHARAN	7462

```

941
18539          Mohammed Mahmood          26092
117
18540      KADI NINDA Dilse_Secular          3017
33
18541          prayag sonar          57426
11236

```

```

      User_location  User_verified  fav_count  rt_count  \
0      Tiruppur, India          NaN          0          0
1      Nirmal Telangana          NaN          0          0
2      Yamunanagar, India          NaN          1          0
3          NaN          NaN          0          0
4      Patna, India          NaN          0          0
...
18537      Washington, DC          1.0          6          2
18538          NaN          NaN          0          0
18539      Aligarh, UP, India          NaN          0          0
18540          NaN          NaN          2          0
18541          NaN          NaN          9          0

```

```

      tweet_date
0      2021-05-28 14:17:55
1      2021-05-28 14:16:56
2      2021-05-28 14:15:56
3      2021-05-28 14:12:51
4      2021-05-28 14:11:00
...
18537      2021-05-20 15:29:23
18538      2021-05-20 15:27:00
18539      2021-05-20 15:04:44
18540      2021-05-20 14:50:12
18541      2021-05-20 13:56:39

```

```
[18542 rows x 9 columns]
```

```
# df.to_csv('white_plus_yellow.csv')
```

```
df.drop_duplicates(['Tweets'], keep = 'first', inplace = True)
```

```
tweets = df.copy()
```

```
kk=0;
```

```
def clean_text(txt):
```

```
    global kk
```

```
    kk=kk+1
```

```
    ...
```

```
    cleans the input text in the following steps:
```

```
    1 - replace contractions
```

```
    2 - removing punctuation
```

```
    3 - splitting into words
```

```
4 - removing stopwords
5 - removing leftover punctuations
6 - lower-case everything
'''
```

```
contraction_dict = {
    "ain't": "is_not", "aren't": "are_not", "can't": "cannot",
    "'cause": "because", "could've": "could_have",
    "couldn't": "could_not", "didn't": "did_not", "doesn't":
    "does_not", "don't": "do_not",
    "hadn't": "had_not", "hasn't": "has_not", "haven't":
    "have_not", "he'd": "he_would", "he'll": "he_will",
    "he's": "he_is", "how'd": "how_did", "how'd'y": "how_do_you",
    "how'll": "how_will", "how's": "how_is",
    "I'd": "I_would", "I'd've": "I_would_have", "I'll": "I_will",
    "I'll've": "I_will_have", "I'm": "I_am",
    "I've": "I_have", "i'd": "i_would", "i'd've": "i_would_have",
    "i'll": "i_will", "i'll've": "i_will_have",
    "i'm": "i_am", "i've": "i_have", "isn't": "is_not", "it'd":
    "it_would", "it'd've": "it_would_have",
    "it'll": "it_will", "it'll've": "it_will_have", "it's":
    "it_is", "let's": "let_us", "ma'am": "madam",
    "mayn't": "may_not", "might've": "might_have", "mightn't":
    "might_not", "mightn't've": "might_not_have",
    "must've": "must_have", "mustn't": "must_not", "mustn't've":
    "must_not_have", "needn't": "need_not",
    "needn't've": "need_not_have", "o'clock": "of_the_clock",
    "oughtn't": "ought_not",
    "oughtn't've": "ought_not_have", "shan't": "shall_not",
    "sha'n't": "shall_not", "shan't've": "shall_not_have",
    "she'd": "she_would", "she'd've": "she_would_have", "she'll":
    "she_will", "she'll've": "she_will_have",
    "she's": "she_is", "should've": "should_have", "shouldn't":
    "should_not", "shouldn't've": "should_not_have",
    "so've": "so_have", "so's": "so_as", "this's":
    "this_is", "that'd": "that_would", "that'd've": "that_would_have",
    "that's": "that_is", "there'd": "there_would", "there'd've":
    "there_would_have", "there's": "there_is",
    "here's": "here_is", "they'd": "they_would", "they'd've":
    "they_would_have", "they'll": "they_will",
    "they'll've": "they_will_have", "they're": "they_are",
    "they've": "they_have", "to've": "to_have",
    "wasn't": "was_not", "we'd": "we_would", "we'd've":
    "we_would_have", "we'll": "we_will",
    "we'll've": "we_will_have", "we're": "we_are", "we've": "we
    have", "weren't": "were_not", "what'll": "what_will",
    "what'll've": "what_will_have", "what're": "what_are",
    "what's": "what_is", "what've": "what_have",
    "when's": "when_is", "when've": "when_have", "where'd": "where
    did", "where's": "where_is",
    "where've": "where_have", "who'll": "who_will", "who'll've":
```

```

"who will have", "who's": "who is",
    "who've": "who have", "why's": "why is", "why've": "why have",
"will've": "will have", "won't": "will_not",
    "won't've": "will_not_have", "would've": "would have",
"wouldn't": "would_not", "wouldn't've": "would_not_have",
    "y'all": "you all", "y'all'd": "you all would", "y'all'd've":
"you all would have", "y'all're": "you all are",
    "y'all've": "you all have", "you'd": "you would", "you'd've":
"you would have", "you'll": "you will",
    "you'll've": "you will have", "you're": "you are", "you've":
"you have"}

```

```

def _get_contractions(contraction_dict):
    contraction_re = re.compile('%s' %
'|'.join(contraction_dict.keys()))
    return contraction_dict, contraction_re

def replace_contractions(text):
    contractions, contractions_re =
_get_contractions(contraction_dict)
    def replace(match):
        return contractions[match.group(0)]
    return str(contractions_re.sub(replace, text))

# replace contractions
txt = txt.lower()
txt = replace_contractions(txt)

# print(kk)

txt = re.sub('(@[A-Za-z0-9]+)|(\w+:\/\/\S+)', ' ', txt)

#remove punctuations
txt = "".join([char for char in txt if char not in
string.punctuation])
txt = re.sub('[0-9]+', ' ', txt)
txt = re.sub('[^0-9A-Za-z \t]', ' ', txt)
# split into words
words = word_tokenize(txt)

# remove stopwords
stop_words = set(stopwords.words('english'))
words = [w for w in words if not w in stop_words]

# removing leftover punctuations
#words = [word for word in words if word.isalpha()]

# lower-case everything

```

```

# stem the words
lem=WordNetLemmatizer()
words = [lem.lemmatize(w) for w in words]

cleaned_text = ' '.join(words)

return cleaned_text

tweets['Tweets'] = tweets['Tweets'].apply(lambda txt:
clean_text(str(txt)))

tweets.head()

```

	Tweets \	User	User_statuses_count	user_followers
0	real white fungus	LoGaN	1753	99
1	white black fungus better aid know	Srinivas Bajju	364	10
2	dear pmo situation getting worse amp cancellat...	Ishu NSUI YAMUNANAGAR ( 2.3k)	58251	2334
3	corona black white fungus aspergillosis india ...	sagarryanhai	84	13
4	th walo covid black white yellow fungus dengue...	Ankit Sinha	200	71

tweet_date	User_location	User_verified	fav_count	rt_count
0 14:17:55	Tiruppur, India	NaN	0	0
1 14:16:56	Nirmal Telangana	NaN	0	0
2 14:15:56	Yamunanagar, India	NaN	1	0
3 14:12:51	NaN	NaN	0	0
4 14:11:00	Patna, India	NaN	0	0

```

tweet = tweets['Tweets']
tweet.head()

```

```

0                                real white fungus
1                white black fungus better aid know
2    dear pmo situation getting worse amp cancellat...
3    corona black white fungus aspergillosis india ...
4    th walo covid black white yellow fungus dengue...
Name: Tweets, dtype: object

```

```

def f(line):
    count = 0
    target_words = ['black',
                    'white',
                    'yellow',
                    'mucormycosis',
                    'mucor',
                    'zygomycosis',
                    'mucormycetes',
                    'candidiasis',
                    'candida',
                    'aspergillosis',
                    'covid',
                    'covid19',
                    'corona',
                    'flu',
                    'infection',
                    'spores',
                    'environment',
                    'skin',
                    'cut',
                    'scrape',
                    'burn',
                    'trauma',
                    'mould'
    ]
    target_words = [str(target_word).strip().lower() for target_word
in target_words]

    for w in word_tokenize(line.strip().lower()):
        if w in target_words:
            count += 1

    return count

```

```

counts = tweet.apply(f)
tweets.insert(0, 'counts', counts)
tweets


```

	counts	Tweets \
0	1	real white fungus
1	2	white black fungus better aid know

```

2          4 dear pmo situation getting worse amp cancellat...
3          4 corona black white fungus aspergillosis india ...
4          4 th walo covid black white yellow fungus dengue...
...      ...
18537      0 yellowwhite fungus grows inside cicada filling...
18538      1          nooo telugu state itis yellow fungus
18539      3          yellow fungus deadlier white amp black one
18540      3 fungus clearly racist virus white amp black br...
18541      3 black fungus white fungus day pink yellow gree...

```

	User	User_statuses_count
user_followers \		
0	LoGaN	1753
99		
1	Srinivas Bajju	364
10		
2	Ishu NSUI YAMUNANAGAR ( 2.3k)	58251
2334		
3	sagarryahanhai	84
13		
4	Ankit Sinha 	200
71		
...	...	...
...		
18537	Diana Hussein	92840
10388		
18538	CHARAN	7462
941		
18539	Mohammed Mahmood	26092
117		
18540	KADI NINDA Dilse_Secular	3017
33		
18541	prayag sonar	57426
11236		

	User_location	User_verified	fav_count	rt_count	\
0	Tiruppur, India	NaN	0	0	
1	Nirmal Telangana	NaN	0	0	
2	Yamunanagar, India	NaN	1	0	
3	NaN	NaN	0	0	
4	Patna, India	NaN	0	0	
...	...	...	...	...	
18537	Washington, DC	1.0	6	2	
18538	NaN	NaN	0	0	
18539	Aligarh, UP, India	NaN	0	0	
18540	NaN	NaN	2	0	
18541	NaN	NaN	9	0	

	tweet_date
0	2021-05-28 14:17:55

```

1      2021-05-28 14:16:56
2      2021-05-28 14:15:56
3      2021-05-28 14:12:51
4      2021-05-28 14:11:00
...
18537  2021-05-20 15:29:23
18538  2021-05-20 15:27:00
18539  2021-05-20 15:04:44
18540  2021-05-20 14:50:12
18541  2021-05-20 13:56:39

```

[13401 rows x 10 columns]

tweets[tweets['counts'] == 0] # which dont want

	counts	Tweets	\
118	0	yeh lo bhai another contender fungal world	
299	0	breaking news latest update may live	
333	0	blackorangewhite fungus meme nonsense shit log...	
340	0	hopefully	
341	0	without understand reason blackyellowwhite fungus	
...	...		
18494	0	hindu party colour used vent orange fungus als...	
18530	0	yellowwhite fungus grows inside cicada filling...	
18531	0	yellowwhite fungus grows inside cicada ring co...	
18534	0	yellowwhite fungus infecting cicada sad year u...	
18537	0	yellowwhite fungus grows inside cicada filling...	

	User	User_statuses_count	user_followers	\
118	On the way to Mars 🚀	5779	111	
299	SikhNews247.Com	48905	267	
333	Subhan Kausar	56	11	
340	Rajarshi Dasgupta	2028	105	
341	Tripti	184	0	
...	...	...	...	
18494	srihari krishna	4738	30	
18530	Laura Miers	100678	11019	
18531	davisadele ♀	98179	532	
18534	Ann Schurman	73878	1160	
18537	Diana Hussein	92840	10388	

	User_location	User_verified	fav_count	
rt_count \				
118	Somewhere among the stars	NaN	0	0
299	NaN	NaN	0	0
333	Muzaffarnagar	NaN	0	0
340	New Delhi,India	NaN	0	0



341		NaN	NaN	0	0
...		...	...	...	...
18494	Bangalore	NaN	0	0	
18530	New York, USA	NaN	20	1	
18531	NY+/-	NaN	0	0	
18534	Mid Atlantic , USA	NaN	6	0	
18537	Washington, DC	1.0	6	2	

	tweet_date
118	2021-05-28 10:35:26
299	2021-05-28 09:15:01
333	2021-05-28 08:36:49
340	2021-05-28 08:13:11
341	2021-05-28 08:11:52
...	...
18494	2021-05-23 05:42:21
18530	2021-05-20 18:30:10
18531	2021-05-20 17:42:23
18534	2021-05-20 16:34:35
18537	2021-05-20 15:29:23

[456 rows x 10 columns]

tweets[tweets['counts'] != 0] # which we want

	counts	Tweets \
0	1	real white fungus
1	2	white black fungus better aid know
2	4	dear pmo situation getting worse amp cancellat...
3	4	corona black white fungus aspergillosis india ...
4	4	th walo covid black white yellow fungus dengue...
...	...	...
18536	2	chahal black fungus yellow fungo
18538	1	nooo telugu state itis yellow fungus
18539	3	yellow fungus deadlier white amp black one
18540	3	fungus clearly racist virus white amp black br...
18541	3	black fungus white fungus day pink yellow gree...

	User	User_statuses_count
user_followers \		
0	LoGaN	1753

99			
1	Srinivas Bajju	364	
10			
2	Ishu NSUI YAMUNANAGAR ( 2.3k)	58251	
2334			
3	sagarriyahanhai	84	
13			
4	Ankit Sinha [IN]	200	
71			
...	...	...	
...			
18536	विक्रम सिंह....	19086	713
18538	CHARAN	7462	
941			
18539	Mohammed Mahmood	26092	
117			
18540	KADI NINDA Dilse_Secular	3017	
33			
18541	prayag sonar	57426	
11236			

	User_location	User_verified	fav_count	rt_count	\
0	Tiruppur, India	NaN	0	0	
1	Nirmal Telangana	NaN	0	0	
2	Yamunanagar, India	NaN	1	0	
3	NaN	NaN	0	0	
4	Patna, India	NaN	0	0	
...	...	...	...	...	
18536	India	NaN	0	0	
18538	NaN	NaN	0	0	
18539	Aligarh, UP, India	NaN	0	0	
18540	NaN	NaN	2	0	
18541	NaN	NaN	9	0	

	tweet_date
0	2021-05-28 14:17:55
1	2021-05-28 14:16:56
2	2021-05-28 14:15:56
3	2021-05-28 14:12:51
4	2021-05-28 14:11:00
...	...
18536	2021-05-20 16:01:18
18538	2021-05-20 15:27:00
18539	2021-05-20 15:04:44
18540	2021-05-20 14:50:12
18541	2021-05-20 13:56:39

[12945 rows x 10 columns]

her's the total runtime (optimized one) with preprocssing -> prev its was around 20 sec

```
tweets = df.copy()
```

```
st = time.time()
```

```
tweets['Tweets'] = tweets['Tweets'].apply(lambda txt:
```

```
clean_text(str(txt)))
```

```
tweet = tweets['Tweets']
```

```
counts = tweet.apply(f)
```

```
print("total time in sec", time.time() - st)
```

```
tweets.insert(0, 'counts', counts)
```

```
total time in sec 10.924993515014648
```