In [194]:

```
import pandas as pd
import wordninja
import re
import nltk
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
nltk.download('averaged_perceptron_tagger')
nltk.download('punkt')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/summerai/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

Out[194]:

```
True
```

In [210]:

```
data = pd.read_csv('fulltext_cleaned_0122.csv')
```

In [211]:

```
data
```

Out[211]:

	Index_Unknown	uid	url	key
0	5	380117	https://www.westernjournal.com/trump-right-jud...	coronavirus|covid|qanon|biden|voterfraud
1	13	448712	https://pjmedia.com/news-and-politics/victoria...	antilatinx|whitesupremacy|an
2	15	256646	https://www.breitbart.com/politics/2020/12/23/...	biden|antilatinx|antiimmigrant|coror
3	15	256646	https://www.breitbart.com/politics/2020/12/23/...	biden|antilatinx|antiimmigrant|coror
4	21	406930	https://www.theepochtimes.com/supreme-court-ju...	bigtech|qanon|
...	...	...	...	
119919	911692	911560	https://hannity.com/media-room/not-so-fast-ari...	qanon|biden|voterfraud|biden|biden|qanon|t
119920	911739	911656	https://www.nytimes.com/2021/03/26/opinion/ezr...	coronavirus|disinformation|whitesupremacy
119921	911739	911656	https://www.nytimes.com/2021/03/26/opinion/ezr...	coronavirus|disinformation|whitesupremacy
119922	911789	911757	https://www.theguardian.com/us-news/2021/apr/0...	coronavirus|covid|antiblack|qanon|whitesupı
119923	911807	911797	https://www.newsweek.com/2024-pence-can-he-mak...	coronavirus|qanon|covid|biden|voterfraud

119924 rows × 6 columns

Assumption

1. Irrelevant text for categories and sublinks are removed
2. Links in the text are removed, hence words like jpg doesn't appear unless it is part of the article
3. White space is approperiate reserved between each two sentences. Eg. no case like sentence1.sentence2

Solution

0. remove null
1. get rid of all special characters; replace them with white space, delete >1 white space
2. use wordninja; consider cases that do not pass to wordninja split eg. n95
3. lemmatize
4. covert all full text to lower case
5. clean stop words

In [212]:

```
# step 0
data = data[data['full_text'].notnull()].reset_index(drop=True)
```

Demo

This demo shows how each step affects a single article

In [262]:

```
example = data.iloc[2,-2]
example
```

Out[262]:

```
'   PoliticsEntertainmentMediaEconomyWorldLondon / EuropeBorder / Cartel ChroniclesIsrael
/ Middle EastAfricaAsiaLatin AmericaAll WorldVideoTechSportsOn the HillOn the Hill Articl
esOn The Hill Exclusive VideoWiresB Inspired    BREITBART    PoliticsEntertainmentMediaEc
onomyWorldLondon / EuropeBorder / Cartel ChroniclesIsrael / Middle EastAfricaAsiaLatin Am
ericaWorld NewsVideoTechSportsOn the HillOn the Hill ArticlesOn The Hill Exclusive VideoW
iresPodcastsBreitbart News DailyB InspiredAbout UsPeopleNewsletters  As Biden suggestedhe
is satisfied with Congress allocating just $600 stimulus checks for each American out of
a $900 billion coronavirus relief package, his advisers said he will push to fund better
housing and coronavirus tests for foreign nationals in Mexico.Biden said, Congress did it
s job this week in reference to the Democrat-controlled House and Republican-controlled S
enate passing the package. President Trump, on the other hand, has demandedthe package be
reworked to include $2,000 stimulus checks for each American.On a call with reporters thi
s week, Biden transition team officials said, they plan on providing funding to improve s
helter and humanitarian assistance to immigrants waiting in northern Mexico, as well as p
rovide COVID-19 testing to ensure people presenting at POE have a negative test before be
ing processed, according to a Buzzfeed News report.Apparently the Biden administration pl
ans to offer shelter and assistance to visa-less migrants even before they try to cross t
he border, including COVID testing.  As Howie Carr would say, "I just want to be treated
like an illegal alien." https://t.co/prrKQBP41aMOST POPULARManchin: Voter Obstruction Is
\'Not Going to Happen\'White House Pursuing Joe Biden Reboot as Polls CollapseKari Lake:
Once Elected I\'ll Declare \'Invasion\' and Close BorderTexas Army National Guardsman Fir
es Upon Smugglers VehicleOregon Officials Float Making Indoor Mask Mandate PermanentCNN O
p-Ed Claims everyone Excited by Prospect of Hillarys ComebackAs U.S. Athletes Face Mounti
ng Dangers in Beijing, Blinken Is in KyivBidens Test Website Prevents Apartment Residents
from Placing OrdersReport Undermines NPR\'s SCOTUS Mask Drama StoryMark Kelly Backs Break
ing Filibuster to Pass Voting Rights Reform FROM THE HOMEPAGEPoll: Joe Biden Approval Rat
ing Hits Record Low as He Reaches One Year as PresidentNolte: Joe Bidens Job Approval Fal
ls to Record Low in RCP Poll AveragePoll: Majority Believe Joe Biden Incapable of Leading
the CountryEstablishment Reporters Say They Will Ask Joe Biden Serious Questions at Year-
One Presser After Softball CampaignBreitbart News Daily Podcast Ep. 54: Dem Infighting an
d Failure Presents the GOP an Opportunity; Guest: Rep. Michael WaltzPoll: GOP Leads Democ
rats by Double Digits on Economy, National Security,JobsNolte: NPR Story About Neil Gorsu
ch Refusing to Wear Mask Debunked as Fake NewsBiden Administration to Distribute 400 Mill
ion N95 MasksRepublicans in Congress Praise Parent Revolution Fighting Back Against Schoo
l MandatesVP Kamala Harris Claims Attending Honduran President Inauguration Will Deter Il
legal ImmigrationKari Lake: As Governor Ill Issue a Declaration of Invasion and Close the
BorderTexas Army National Guardsman Fires Upon Smugglers Vehicle near BorderNolte: Texas
Rabbis Account of Hostage Escape Contradicts FBI Spin  '
```

In [263]:

```
# step 1

# repalce all characters with a white space
# except - between letters and (, or .) between digits or letters;
```

```
# Eg. keep COVID-19, one-year, 2,000, 3.00 and U.S
step1_remove_char = re.sub(r"(?!(?<=[a-zA-Z0-9])[\,\.\-](?=[a-zA-Z0-9]))[^a-zA-Z0-9 \n]",
" ", example + '3.00, 5g')
# if there are more than one white spaces between words, reduce to one
step1_remove_spaces = re.sub('\s+', " ", step1_remove_char).strip()
```

In [264]:

```
step1_remove_char
```

Out[264]:

```
'   PoliticsEntertainmentMediaEconomyWorldLondon   EuropeBorder   Cartel ChroniclesIsrael
Middle EastAfricaAsiaLatin AmericaAll WorldVideoTechSportsOn the HillOn the Hill Articles
On The Hill Exclusive VideoWiresB Inspired    BREITBART    PoliticsEntertainmentMediaEcon
omyWorldLondon   EuropeBorder   Cartel ChroniclesIsrael   Middle EastAfricaAsiaLatin Amer
icaWorld NewsVideoTechSportsOn the HillOn the Hill ArticlesOn The Hill Exclusive VideoWir
esPodcastsBreitbart News DailyB InspiredAbout UsPeopleNewsletters  As Biden suggestedhe i
s satisfied with Congress allocating just  600 stimulus checks for each American out of a
900 billion coronavirus relief package  his advisers said he will push to fund better hou
sing and coronavirus tests for foreign nationals in Mexico.Biden said  Congress did its j
ob this week in reference to the Democrat-controlled House and Republican-controlled Sena
te passing the package  President Trump  on the other hand  has demandedthe package be re
worked to include  2,000 stimulus checks for each American.On a call with reporters this
week  Biden transition team officials said  they plan on providing funding to improve she
lter and humanitarian assistance to immigrants waiting in northern Mexico  as well as pro
vide COVID-19 testing to ensure people presenting at POE have a negative test before bein
g processed  according to a Buzzfeed News report.Apparently the Biden administration plan
s to offer shelter and assistance to visa-less migrants even before they try to cross the
border  including COVID testing   As Howie Carr would say   I just want to be treated lik
e an illegal alien   https   t.co prrKQBP41aMOST POPULARManchin  Voter Obstruction Is  No
t Going to Happen White House Pursuing Joe Biden Reboot as Polls CollapseKari Lake  Once
Elected I ll Declare  Invasion  and Close BorderTexas Army National Guardsman Fires Upon
Smugglers VehicleOregon Officials Float Making Indoor Mask Mandate PermanentCNN Op-Ed Cla
ims everyone Excited by Prospect of Hillarys ComebackAs U.S  Athletes Face Mounting Dange
rs in Beijing  Blinken Is in KyivBidens Test Website Prevents Apartment Residents from Pl
acing OrdersReport Undermines NPR s SCOTUS Mask Drama StoryMark Kelly Backs Breaking Fili
buster to Pass Voting Rights Reform FROM THE HOMEPAGEPoll  Joe Biden Approval Rating Hits
Record Low as He Reaches One Year as PresidentNolte  Joe Bidens Job Approval Falls to Rec
ord Low in RCP Poll AveragePoll  Majority Believe Joe Biden Incapable of Leading the Coun
tryEstablishment Reporters Say They Will Ask Joe Biden Serious Questions at Year-One Pres
ser After Softball CampaignBreitbart News Daily Podcast Ep  54  Dem Infighting and Failur
e Presents the GOP an Opportunity  Guest  Rep  Michael WaltzPoll  GOP Leads Democrats by
Double Digits on Economy  National Security,JobsNolte  NPR Story About Neil Gorsuch Refus
ing to Wear Mask Debunked as Fake NewsBiden Administration to Distribute 400 Million N95
MasksRepublicans in Congress Praise Parent Revolution Fighting Back Against School Mandat
esVP Kamala Harris Claims Attending Honduran President Inauguration Will Deter Illegal Im
migrationKari Lake  As Governor Ill Issue a Declaration of Invasion and Close the BorderT
exas Army National Guardsman Fires Upon Smugglers Vehicle near BorderNolte  Texas Rabbis
Account of Hostage Escape Contradicts FBI Spin  3.00  5g'
```

In [265]:

```
step1_remove_spaces
```

Out[265]:

```
'PoliticsEntertainmentMediaEconomyWorldLondon EuropeBorder Cartel ChroniclesIsrael Middle
EastAfricaAsiaLatin AmericaAll WorldVideoTechSportsOn the HillOn the Hill ArticlesOn The
Hill Exclusive VideoWiresB Inspired BREITBART PoliticsEntertainmentMediaEconomyWorldLondo
n EuropeBorder Cartel ChroniclesIsrael Middle EastAfricaAsiaLatin AmericaWorld NewsVideoT
echSportsOn the HillOn the Hill ArticlesOn The Hill Exclusive VideoWiresPodcastsBreitbart
News DailyB InspiredAbout UsPeopleNewsletters As Biden suggestedhe is satisfied with Cong
ress allocating just 600 stimulus checks for each American out of a 900 billion coronavir
us relief package his advisers said he will push to fund better housing and coronavirus t
ests for foreign nationals in Mexico.Biden said Congress did its job this week in referen
ce to the Democrat-controlled House and Republican-controlled Senate passing the package
President Trump on the other hand has demandedthe package be reworked to include 2,000 st
imulus checks for each American.On a call with reporters this week Biden transition team
officials said they plan on providing funding to improve shelter and humanitarian assista
nce to immigrants waiting in northern Mexico as well as provide COVID-19 testing to ensur
e people presenting at POE have a negative test before being processed according to a Buz
zfeed News report.Apparently the Biden administration plans to offer shelter and assistan
```

ce to visa-less migrants even before they try to cross the border including COVID testing As Howie Carr would say I just want to be treated like an illegal alien https t.co prrKQB P41aMOST POPULARManchin Voter Obstruction Is Not Going to Happen White House Pursuing Joe Biden Reboot as Polls CollapseKari Lake Once Elected I ll Declare Invasion and Close Bord erTexas Army National Guardsman Fires Upon Smugglers VehicleOregon Officials Float Making Indoor Mask Mandate PermanentCNN Op-Ed Claims everyone Excited by Prospect of Hillarys Co mebackAs U.S Athletes Face Mounting Dangers in Beijing Blinken Is in KyivBidens Test Webs ite Prevents Apartment Residents from Placing OrdersReport Undermines NPR s SCOTUS Mask D rama StoryMark Kelly Backs Breaking Filibuster to Pass Voting Rights Reform FROM THE HOME PAGEPoll Joe Biden Approval Rating Hits Record Low as He Reaches One Year as PresidentNol te Joe Bidens Job Approval Falls to Record Low in RCP Poll AveragePoll Majority Believe J oe Biden Incapable of Leading the CountryEstablishment Reporters Say They Will Ask Joe Bi den Serious Questions at Year-One Presser After Softball CampaignBreitbart News Daily Pod cast Ep 54 Dem Infighting and Failure Presents the GOP an Opportunity Guest Rep Michael W altzPoll GOP Leads Democrats by Double Digits on Economy National Security,JobsNolte NPR Story About Neil Gorsuch Refusing to Wear Mask Debunked as Fake NewsBiden Administration to Distribute 400 Million N95 MasksRepublicans in Congress Praise Parent Revolution Fight ing Back Against School MandatesVP Kamala Harris Claims Attending Honduran President Inau guration Will Deter Illegal ImmigrationKari Lake As Governor Ill Issue a Declaration of I nvasion and Close the BorderTexas Army National Guardsman Fires Upon Smugglers Vehicle ne ar BorderNolte Texas Rabbis Account of Hostage Escape Contradicts FBI Spin 3.00 5g'

In [266]:

```
# step 2

# if a word matches this pattern or is in the list then we don't want to pass it to wordn
inja
# if there is hyphen, combination of letters and digits and pure capitalized letters, don
't pass
wordninja_filter = re.compile(r"-|([A-Za-z]+\d+\w*|\d+[A-Za-z]+\w*)|^[^a-z]*$")
# if a word is in the list, don't pass it to wordninja because it can't handle well
words_pass = ['qanon', 'covid', 'vaxx']

# split the string by a white space
string_isolated = step1_remove_spaces.split()

# check word by word to detect split
step2_words_split = ''
for el in string_isolated:
    # if the word matches the pattern or is in the list, then we don't pass it to wordnij
ia to split
    if wordninja_filter.search(el) or el.lower() in words_pass:
        temp = el
    # all the other words will be checked if be split if necessary
    else:
        temp = ' '.join(wordninja.split(el))
    step2_words_split += ' ' + temp
    step2_words_split = step2_words_split.strip()
```

In [268]:

```
step2_words_split
```

Out[268]:

'Politics Entertainment Media Economy World London Europe Border Cartel Chronicles Israel Middle East Africa Asia Latin America All World Video Tech Sports On the Hill On the Hill Articles On The Hill Exclusive Video Wires B Inspired BREITBART Politics Entertainment Me dia Economy World London Europe Border Cartel Chronicles Israel Middle East Africa Asia L atin America World News Video Tech Sports On the Hill On the Hill Articles On The Hill Ex clusive Video Wires Podcasts Bre it bart News Daily B Inspired About Us People Newsletter s As Biden suggested he is satisfied with Congress allocating just 600 stimulus checks fo r each American out of a 900 billion coronavirus relief package his advisers said he will push to fund better housing and coronavirus tests for foreign nationals in Mexico Biden s aid Congress did its job this week in reference to the Democrat-controlled House and Repu blican-controlled Senate passing the package President Trump on the other hand has demand ed the package be reworked to include 2,000 stimulus checks for each American On a call w ith reporters this week Biden transition team officials said they plan on providing fundi ng to improve shelter and humanitarian assistance to immigrants waiting in northern Mexic o as well as provide COVID-19 testing to ensure people presenting at POE have a negative test before being processed according to a Buzz feed News report Apparently the Biden adm

inistration plans to offer shelter and assistance to visa-less migrants even before they try to cross the border including COVID testing As Howie Carr would say I just want to be treated like an illegal alien https t co prrKQBP41aMOST POPULAR Manchin Voter Obstruction Is Not Going to Happen White House Pursuing Joe Biden Reboot as Polls Collapse Kari Lake Once Elected I ll Declare Invasion and Close Border Texas Army National Guardsman Fires U pon Smugglers Vehicle Oregon Officials Float Making Indoor Mask Mandate Permanent CNN Op-Ed Claims everyone Excited by Prospect of Hillary s Comeback As U.S Athletes Face Mountin g Dangers in Beijing Blink en Is in Kyiv Biden s Test Website Prevents Apartment Resident s from Placing Orders Report Undermines NPR s SCOTUS Mask Drama Story Mark Kelly Backs Br eaking Filibuster to Pass Voting Rights Reform FROM THE HOMEPAGE Poll Joe Biden Approval Rating Hits Record Low as He Reaches One Year as President Nolte Joe Biden s Job Approval Falls to Record Low in RCP Poll Average Poll Majority Believe Joe Biden Incapable of Lead ing the Country Establishment Reporters Say They Will Ask Joe Biden Serious Questions at Year-One Presser After Softball Campaign Bre it bart News Daily Podcast Ep 54 Dem In figh ting and Failure Presents the GOP an Opportunity Guest Rep Michael Waltz Poll GOP Leads D emocrats by Double Digits on Economy National Security Jobs Nolte NPR Story About Neil Go r such Refusing to Wear Mask Debunked as Fake News Biden Administration to Distribute 400 Million N95 Masks Republicans in Congress Praise Parent Revolution Fighting Back Against School Mandates VP Kamala Harris Claims Attending Honduran President Inauguration Will De ter Illegal Immigration Kari Lake As Governor Ill Issue a Declaration of Invasion and Clo se the Border Texas Army National Guardsman Fires Upon Smugglers Vehicle near Border Nolt e Texas Rabbis Account of Hostage Escape Contradicts FBI Spin 3.00 5g'

In [269]:

```
# step 3
def step3_get_wordnet_pos(word):
    """Map POS tag to first character lemmatize() accepts"""
    tag = nltk.pos_tag([word])[0][1][0].upper()
    tag_dict = {"J": wordnet.ADJ,
                "N": wordnet.NOUN,
                "V": wordnet.VERB,
                "R": wordnet.ADV}
    return tag_dict.get(tag, wordnet.NOUN)

step3_lemmatizer = WordNetLemmatizer()
```

In [270]:

```
temp_2 = ''
for word in step2_words_split.split():
    words_lemmatized = step3_lemmatizer.lemmatize(word, step3_get_wordnet_pos(word))
    temp_2 += ' ' + words_lemmatized
```

In [271]:

```
temp_2
```

Out[271]:

' Politics Entertainment Media Economy World London Europe Border Cartel Chronicles Israe l Middle East Africa Asia Latin America All World Video Tech Sports On the Hill On the Hi ll Articles On The Hill Exclusive Video Wires B Inspired BREITBART Politics Entertainment Media Economy World London Europe Border Cartel Chronicles Israel Middle East Africa Asia Latin America World News Video Tech Sports On the Hill On the Hill Articles On The Hill E xclusive Video Wires Podcasts Bre it bart News Daily B Inspired About Us People Newslette rs As Biden suggest he be satisfied with Congress allocate just 600 stimulus check for ea ch American out of a 900 billion coronavirus relief package his adviser say he will push to fund well housing and coronavirus test for foreign national in Mexico Biden say Congre ss do it job this week in reference to the Democrat-controlled House and Republican-contr olled Senate passing the package President Trump on the other hand have demand the packag e be rework to include 2,000 stimulus check for each American On a call with reporter thi s week Biden transition team official say they plan on provide funding to improve shelter and humanitarian assistance to immigrant wait in northern Mexico a well a provide COVID-1 9 test to ensure people present at POE have a negative test before be process accord to a Buzz feed News report Apparently the Biden administration plan to offer shelter and assis tance to visa-less migrant even before they try to cross the border include COVID test As Howie Carr would say I just want to be treat like an illegal alien http t co prrKQBP41aMO ST POPULAR Manchin Voter Obstruction Is Not Going to Happen White House Pursuing Joe Bide n Reboot a Polls Collapse Kari Lake Once Elected I ll Declare Invasion and Close Border T exas Army National Guardsman Fires Upon Smugglers Vehicle Oregon Officials Float Making I ndoor Mask Mandate Permanent CNN Op-Ed Claims everyone Excited by Prospect of Hillary s C

omeback As U.S Athletes Face Mounting Dangers in Beijing Blink en Is in Kyiv Biden s Test Website Prevents Apartment Residents from Placing Orders Report Undermines NPR s SCOTUS M ask Drama Story Mark Kelly Backs Breaking Filibuster to Pass Voting Rights Reform FROM TH E HOMEPAGE Poll Joe Biden Approval Rating Hits Record Low a He Reaches One Year a Preside nt Nolte Joe Biden s Job Approval Falls to Record Low in RCP Poll Average Poll Majority B elieve Joe Biden Incapable of Leading the Country Establishment Reporters Say They Will A sk Joe Biden Serious Questions at Year-One Presser After Softball Campaign Bre it bart Ne ws Daily Podcast Ep 54 Dem In fight and Failure Presents the GOP an Opportunity Guest Rep Michael Waltz Poll GOP Leads Democrats by Double Digits on Economy National Security Jobs Nolte NPR Story About Neil Gor such Refusing to Wear Mask Debunked a Fake News Biden Admi nistration to Distribute 400 Million N95 Masks Republicans in Congress Praise Parent Revo lution Fighting Back Against School Mandates VP Kamala Harris Claims Attending Honduran P resident Inauguration Will Deter Illegal Immigration Kari Lake As Governor Ill Issue a De claration of Invasion and Close the Border Texas Army National Guardsman Fires Upon Smugg lers Vehicle near Border Nolte Texas Rabbis Account of Hostage Escape Contradicts FBI Spi n 3.00 5g'

In [272]:

```
# step 4 & step 5
step4_stop_words = set(stopwords.words('english'))
result = ''
for word in temp_2.split():
    if word.lower() not in step4_stop_words:
        result += ' ' + word.lower()
print(result.strip())
```

politics entertainment media economy world london europe border cartel chronicles israel middle east africa asia latin america world video tech sports hill hill articles hill exc lusive video wires b inspired breitbart politics entertainment media economy world london europe border cartel chronicles israel middle east africa asia latin america world news v ideo tech sports hill hill articles hill exclusive video wires podcasts bre bart news dai ly b inspired us people newsletters biden suggest satisfied congress allocate 600 stimulu s check american 900 billion coronavirus relief package adviser say push fund well housin g coronavirus test foreign national mexico biden say congress job week reference democrat -controlled house republican-controlled senate passing package president trump hand deman d package rework include 2,000 stimulus check american call reporter week biden transitio n team official say plan provide funding improve shelter humanitarian assistance immigran t wait northern mexico well provide covid-19 test ensure people present poe negative test process accord buzz feed news report apparently biden administration plan offer shelter a ssistance visa-less migrant even try cross border include covid test howie carr would say want treat like illegal alien http co prrkqbp41amost popular manchin voter obstruction go ing happen white house pursuing joe biden reboot polls collapse kari lake elected declare invasion close border texas army national guardsman fires upon smugglers vehicle oregon o fficials float making indoor mask mandate permanent cnn op-ed claims everyone excited pro spect hillary comeback u.s athletes face mounting dangers beijing blink en kyiv biden tes t website prevents apartment residents placing orders report undermines npr scotus mask d rama story mark kelly backs breaking filibuster pass voting rights reform homepage poll j oe biden approval rating hits record low reaches one year president nolte joe biden job a pproval falls record low rcp poll average poll majority believe joe biden incapable leadi ng country establishment reporters say ask joe biden serious questions year-one presser s oftball campaign bre bart news daily podcast ep 54 dem fight failure presents gop opportu nity guest rep michael waltz poll gop leads democrats double digits economy national secu rity jobs nolte npr story neil gor refusing wear mask debunked fake news biden administra tion distribute 400 million n95 masks republicans congress praise parent revolution fight ing back school mandates vp kamala harris claims attending honduran president inauguratio n deter illegal immigration kari lake governor ill issue declaration invasion close borde r texas army national guardsman fires upon smugglers vehicle near border nolte texas rabb is account hostage escape contradicts fbi spin 3.00 5g

Putting everything together

In [273]:

```
def fulltext_clean(string):
    #PREPARATION
    # step 1
    # repalce all characters with a white space except these three char - , . among di
gits/letters;
    # Eg. keep 2,000, 3.00, covid-19
```

```
    remove_char = re.sub(r"(?!(?<=[a-zA-Z0-9])[\,\.\-](?=[a-zA-Z0-9]))[^a-zA-Z0-9 \n]","
", string)
    # if there are more than one white spaces between words, reduce to one
    remove_spaces = re.sub('\s+', " ", remove_char).strip()

    # step 2
    # if a word matches this pattern or is in the list then we don't want to pass it to w
ordninja
    # if there is hyphen, combination of letters and digits or pure capitalized letters,
don't pass
    wordninja_filter = re.compile(r"-|([A-Za-z]+\d+\w*|\d+[A-Za-z]+\w*)|^[^a-z]*$")
    # if a word is in the list, don't pass it to wordninja because it can't handle the wo
rd well
    words_pass = ['qanon', 'covid']

    # step 3
    # set up for lemmatize
    def get_wordnet_pos(word):
        """Map POS tag to first character lemmatize() accepts"""
        tag = nltk.pos_tag([word])[0][1][0].upper()
        tag_dict = {"J": wordnet.ADJ,
                    "N": wordnet.NOUN,
                    "V": wordnet.VERB,
                    "R": wordnet.ADV}
        return tag_dict.get(tag, wordnet.NOUN)

    lemmatizer = WordNetLemmatizer()

    # step4
    # prepare stop words
    stop_words = set(stopwords.words('english'))

    # CLEANING
    # split the string by a white space
    string_isolated = remove_spaces.split()

    # check the string word by word to detect necessary split, lemmatize and remove stop
word
    words_split = ''
    for el in string_isolated:
        # step 2
        # if the word matches the pattern or is in the list, then we don't pass it to wor
dnijia to split
        if wordninja_filter.search(el) or el.lower() in words_pass:
            temp = el
        # all the other words will be checked and be split if necessary
        else:
            temp = ' '.join(wordninja.split(el))

        # step 3: lemmatize the word
        words_lemmatized = lemmatizer.lemmatize(temp, get_wordnet_pos(temp))

        # step 4 & step 5
        if words_lemmatized.lower() not in stop_words:
             words_split += ' ' + words_lemmatized.lower()

    words_split = words_split.strip()

    return words_split
```

In [274]:

```
# original article
single_article = data.iloc[2000,-2]
single_article
```

Out[274]:

'PopularFully Vaxxed ESPN Host Stephen Smith Says He Nearly Died After Contracting CovidA
rmy Conducting Two-Week \'Guerrilla Warfare Exercise\' in Rural North Carolina Focused On
Battling \'Freedom Fighters\'\'Shut The F**k Up!\' Fans Heckle NYC Mayor Eric Adams At Ne
w York Knicks GamePoll: Nearly Half of Democrats Support Fining or Imprisoning Americans

```
Who \'Question Efficacy\' of Covid ShotsTrain Derails in Garbage-Strewn Area Trashed by L
ooters in Los Angeles BREAKING: Coca-Cola is forcing employees to complete online trainin
g telling them to "try to be less white." These images are from an internal whistleblower
: pic.twitter.com/gRi4N20esZ Karlyn supports banning critical race theory in NH (@DrKarly
nB) February 19, 2021 Tucker Carlson\'s Review Of Robin DiAngelo\'s Book \'White Fragilit
y\'"The real point of her book is to defeat & demoralize you.""Everything about \'White F
ragility\' is poisonous garbage." pic.twitter.com/4lkyz57ROQ The Columbia Bugle  (@Columb
iaBugle) June 25, 2020 '
```

In [275]:

```
# aritcle after cleaning
fulltext_clean(single_article)
```

Out[275]:

```
'popular fully vax x ed espn host stephen smith says nearly died contracting c ovid army
conducting two-week guerrilla warfare exercise rural north carolina focused battling free
dom fighters shut f k fans heckle nyc mayor eric adams new york knicks game poll nearly h
alf democrats support fining imprisoning americans question efficacy covid shots train de
rails garbage-strewn area trashed looters los angeles breaking coca-cola force employee c
omplete online training tell try less white image internal whistle blower pic twitter com
gri4n20esz karl yn support ban critical race theory nh dr karl yn b february 19 2021 tuck
er carlson review robin di angelo book white fragility real point book defeat demoralize
everything white fragility poisonous garbage pic twitter com 4lkyz57roq columbia bugle co
lumbia bugle june 25 2020'
```

Discussion

1. Should we deal with words that look like shit and fk? Eg. see data.iloc[150, -2] and the case above
2. If a word in 'word_pass' is concatenated together with another word, we can't separate it correctly. Eg. see the word 'CovidArmy' above - words like this will be sent to wordninja but since it can't separate covid well we end up getting c ovid army; if we don't pass it to wordninja we end up getting covidarmy

In [ ]:

```

```