localhost:8888/notebooks/Desktop/Ruksana/DataScience/ExcelR_Assignements/18.Text Mining/Elon_musk_Text-Mining_Solution.ipynb

1/49

In [6]:

```
1  !pip install spacy
```

```
Collecting spacy
  Downloading spacy-3.2.1-cp38-cp38-win_amd64.whl (12.2 MB)
Collecting thinc<8.1.0,>=8.0.12
  Downloading thinc-8.0.13-cp38-cp38-win_amd64.whl (1.0 MB)
Collecting blis<0.8.0,>=0.4.0
  Downloading blis-0.7.5-cp38-cp38-win_amd64.whl (6.6 MB)
Collecting srsly<3.0.0,>=2.4.1
  Downloading srsly-2.4.2-cp38-cp38-win_amd64.whl (452 kB)
Collecting pathy>=0.3.5
  Downloading pathy-0.6.1-py3-none-any.whl (42 kB)
Collecting pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4
  Downloading pydantic-1.8.2-cp38-cp38-win_amd64.whl (2.0 MB)
Requirement already satisfied: jinja2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (2.
11.3)
Requirement already satisfied: numpy>=1.15.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spa
cy) (1.20.1)
Collecting wasabi<1.1.0,>=0.8.1
  Downloading wasabi-0.9.0-py3-none-any.whl (25 kB)
Collecting preshed<3.1.0,>=3.0.2
  Downloading preshed-3.0.6-cp38-cp38-win_amd64.whl (113 kB)
Collecting murmurhash<1.1.0,>=0.28.0
  Downloading murmurhash-1.0.6-cp38-cp38-win_amd64.whl (21 kB)
Requirement already satisfied: setuptools in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy)
(52.0.0.post20210125)
Collecting catalogue<2.1.0,>=2.0.6
  Downloading catalogue-2.0.6-py3-none-any.whl (17 kB)
Collecting spacy-legacy<3.1.0,>=3.0.8
  Downloading spacy_legacy-3.0.8-py2.py3-none-any.whl (14 kB)
Collecting typer<0.5.0,>=0.3.0
  Downloading typer-0.4.0-py3-none-any.whl (27 kB)
Collecting langcodes<4.0.0,>=3.2.0
  Downloading langcodes-3.3.0-py3-none-any.whl (181 kB)
Requirement already satisfied: packaging>=20.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from s
pacy) (20.9)
Collecting spacy-loggers<2.0.0,>=1.0.0
  Downloading spacy_loggers-1.0.1-py3-none-any.whl (7.0 kB)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from spacy) (2.25.1)
Collecting cymem<2.1.0,>=2.0.2
  Downloading cymem-2.0.6-cp38-cp38-win_amd64.whl (36 kB)
```

```
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy) (4.59.0)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
packaging>=20.0->spacy) (2.4.7)
Collecting smart-open<6.0.0,>=5.0.0
  Downloading smart_open-5.2.1-py3-none-any.whl (58 kB)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\f safrin kauser\anaconda3\lib\site-packa
ges (from pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4->spacy) (3.7.4.3)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fro
m requests<3.0.0,>=2.13.0->spacy) (2020.12.5)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from requests<3.0.0,>=2.13.0->spacy) (1.26.4)
Requirement already satisfied: chardet<5,>=3.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy) (4.0.0)
Requirement already satisfied: idna<3,>=2.5 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from requ
ests<3.0.0,>=2.13.0->spacy) (2.10)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om typer<0.5.0,>=0.3.0->spacy) (7.1.2)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
jinja2->spacy) (1.1.1)
Installing collected packages: murmurhash, cymem, catalogue, wasabi, typer, srsly, smart-open, pydantic, presh
ed, blis, thinc, spacy-loggers, spacy-legacy, pathy, langcodes, spacy
Successfully installed blis-0.7.5 catalogue-2.0.6 cymem-2.0.6 langcodes-3.3.0 murmurhash-1.0.6 pathy-0.6.1 pre
shed-3.0.6 pydantic-1.8.2 smart-open-5.2.1 spacy-3.2.1 spacy-legacy-3.0.8 spacy-loggers-1.0.1 srsly-2.4.2 thin
c-8.0.13 typer-0.4.0 wasabi-0.9.0
```

In [7]:
```
1  !pip install wordcloud
```

```
Collecting wordcloud
  Downloading wordcloud-1.8.1-cp38-cp38-win_amd64.whl (155 kB)
Requirement already satisfied: matplotlib in c:\users\f safrin kauser\anaconda3\lib\site-packages (from wordcl
oud) (3.3.4)
Requirement already satisfied: numpy>=1.6.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from word
cloud) (1.20.1)
Requirement already satisfied: pillow in c:\users\f safrin kauser\anaconda3\lib\site-packages (from wordcloud)
(8.2.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\f safrin kauser\anaconda3
\lib\site-packages (from matplotlib->wordcloud) (2.4.7)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (f
rom matplotlib->wordcloud) (2.8.1)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
matplotlib->wordcloud) (1.3.1)
Requirement already satisfied: cycler>=0.10 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from matp
lotlib->wordcloud) (0.10.0)
Requirement already satisfied: six in c:\users\f safrin kauser\anaconda3\lib\site-packages (from cycler>=0.10-
>matplotlib->wordcloud) (1.15.0)
Installing collected packages: wordcloud
Successfully installed wordcloud-1.8.1
```

In [38]:
```
1   # Import Libraries
2   import numpy as np
3   import pandas as pd
4   import matplotlib.pyplot as plt
5   import string
6   import spacy
7
8   from matplotlib.pyplot import imread
9   from wordcloud import WordCloud, STOPWORDS
10  %matplotlib inline
```

In [39]:
```python
# load the dataset
tweets=pd.read_csv('Elon_musk.csv',encoding='Latin-1')
tweets.drop(['Unnamed: 0'],inplace=True,axis=1)
tweets
```

Out[39]:

	Text
0	@kunalb11 I m an alien
1	@ID_AA_Carmack Ray tracing on Cyberpunk with H...
2	@joerogan @Spotify Great interview!
3	@gtera27 Doge is underestimated
4	@teslacn Congratulations Tesla China for amazi...
...	...
1994	@flcnhvy True, it sounds so surreal, but the n...
1995	@PPathole Make sure to read ur terms & con...
1996	@TeslaGong @PPathole Samwise Gamgee
1997	@PPathole Altho Dumb and Dumber is <U+0001F525...
1998	Progress update August 28

1999 rows × 1 columns

Text Preprocessing

In [40]:
```python
tweets=[Text.strip() for Text in tweets.Text] # remove both the leading and the trailing characters
tweets=[Text for Text in tweets if Text] # removes empty strings, because they are considered in Python as |
tweets[0:10]
```

Out[40]:
```
['@kunalb11 I\x92m an alien',
 '@ID_AA_Carmack Ray tracing on Cyberpunk with HDR is next-level. Have you tried it?',
 '@joerogan @Spotify Great interview!',
 '@gtera27 Doge is underestimated',
 '@teslacn Congratulations Tesla China for amazing execution last year. Now on to the next for even more!!',
 'Happy New Year of the Ox! https://t.co/9WFKMYu2oj', (https://t.co/9WFKMYu2oj',)
 'Frodo was the underdoge,\nAll thought he would fail,\nHimself most of all. https://t.co/zGxJFDzzrM', (http
s://t.co/zGxJFDzzrM',)
 '@OwenSparks_ @flcnhvy @anonyx10 Haha thanks :)',
 '@flcnhvy @anonyx10 Indeed! Tweets definitely do not represent real-world time allocation.',
 'The most entertaining outcome is the most likely']
```

In [41]:
```python
# Joining the list into one string/text
tweets_text=' '.join(tweets)
tweets_text
```

Out[41]:
```
'@kunalb11 I\x92m an alien @ID_AA_Carmack Ray tracing on Cyberpunk with HDR is next-level. Have you tried i
t? @joerogan @Spotify Great interview! @gtera27 Doge is underestimated @teslacn Congratulations Tesla China
for amazing execution last year. Now on to the next for even more!! Happy New Year of the Ox! https://t.co/
9WFKMYu2oj (https://t.co/9WFKMYu2oj) Frodo was the underdoge,\nAll thought he would fail,\nHimself most of
 all. https://t.co/zGxJFDzzrM (https://t.co/zGxJFDzzrM) @OwenSparks_ @flcnhvy @anonyx10 Haha thanks :) @flc
nhvy @anonyx10 Indeed! Tweets definitely do not represent real-world time allocation. The most entertaining
outcome is the most likely @GiveDirectly Just sent some Just agree to do Clubhouse with @kanyewest https://
t.co/3rWE9uHSTS (https://t.co/3rWE9uHSTS) @geoffkeighley @UnrealEngine It\x92s getting real Bought some Dog
ecoin for lil X, so he can be a toddler hodler @JoshManMode He definitely has issues, but the sentencing se
ems a bit high @freewalletorg Thanks for fixing @freewalletorg Please unlock my account @AstroJordy <U+0001
F923><U+0001F923> This is true power haha https://t.co/Fc9uhQSd7O (https://t.co/Fc9uhQSd7O) @freewalletorg
 Any crypto wallet that won\x92t give you your private keys should be avoided at all costs @freewalletorg Y
our app sucks RT @SpaceX: NASA has selected Falcon Heavy to launch the first two elements of the lunar Gate
way together on one mission! https://t.co/3pWt (https://t.co/3pWt) @ajtourville Yes @BLKMDL3 @RationalEtien
ne @Adamklotz_ Once we can predict cash flow reasonably well, Starlink will IPO @RationalEtienne @Adamklotz
_ Starlink is a staggeringly difficult technical &amp; economic endeavor. However, if we don\x85 https://t.
co/9Z8Ac6skqx (https://t.co/9Z8Ac6skqx) @RationalEtienne @Adamklotz_ SpaceX needs to pass through a deep ch
asm of negative cash flow over the next year or\x85 https://t.co/7J1c92hdjl (https://t.co/7J1c92hdjl) @ID_A
A_Carmack Lowest cost per ton of carbon sequestered, net of value of any product made. \n\nMust be scalable
```

In [42]:

```python
# remove Twitter username handles from a given twitter text. (Removes @usernames)
from nltk.tokenize import TweetTokenizer
tknzr = TweetTokenizer(strip_handles=True)
tweets_tokens=tknzr.tokenize(tweets_text)
print(tweets_tokens)
```

```
['I', '\x92', 'm', 'an', 'alien', 'Ray', 'tracing', 'on', 'Cyberpunk', 'with', 'HDR', 'is', 'next-level',
'.', 'Have', 'you', 'tried', 'it', '?', 'Great', 'interview', '!', 'Doge', 'is', 'underestimated', 'Congrat
ulations', 'Tesla', 'China', 'for', 'amazing', 'execution', 'last', 'year', '.', 'Now', 'on', 'to', 'the',
'next', 'for', 'even', 'more', '!', '!', 'Happy', 'New', 'Year', 'of', 'the', 'Ox', '!', 'https://t.co/9WFK
MYu2oj', 'Frodo', 'was', 'the', 'underdoge', ',', 'All', 'thought', 'he', 'would', 'fail', ',', 'Himself',
'most', 'of', 'all', '.', 'https://t.co/zGxJFDzzrM', 'Haha', 'thanks', ':)', 'Indeed', '!', 'Tweets', 'defi
nitely', 'do', 'not', 'represent', 'real-world', 'time', 'allocation', '.', 'The', 'most', 'entertaining',
'outcome', 'is', 'the', 'most', 'likely', 'Just', 'sent', 'some', 'Just', 'agree', 'to', 'do', 'Clubhouse',
'with', 'https://t.co/3rWE9uHSTS', 'It', '\x92', 's', 'getting', 'real', 'Bought', 'some', 'Dogecoin', 'fo
r', 'lil', 'X', ',', 'so', 'he', 'can', 'be', 'a', 'toddler', 'hodler', 'He', 'definitely', 'has', 'issue
s', ',', 'but', 'the', 'sentencing', 'seems', 'a', 'bit', 'high', 'Thanks', 'for', 'fixing', 'Please', 'unl
ock', 'my', 'account', '<U+0001F923>', '<U+0001F923>', 'This', 'is', 'true', 'power', 'haha', 'https://t.c
o/Fc9uhQSd7O', 'Any', 'crypto', 'wallet', 'that', 'won', '\x92', 't', 'give', 'you', 'your', 'private', 'ke
ys', 'should', 'be', 'avoided', 'at', 'all', 'costs', 'Your', 'app', 'sucks', 'RT', ':', 'NASA', 'has', 'se
lected', 'Falcon', 'Heavy', 'to', 'launch', 'the', 'first', 'two', 'elements', 'of', 'the', 'lunar', 'Gatew
ay', 'together', 'on', 'one', 'mission', '!', 'https://t.co/3pWt', 'Yes', 'Once', 'we', 'can', 'predict',
'cash', 'flow', 'reasonably', 'well', ',', 'Starlink', 'will', 'IPO', 'Starlink', 'is', 'a', 'staggeringl
y', 'difficult', 'technical', '&', 'economic', 'endeavor', '.', 'However', ',', 'if', 'we', 'don', 'http
s://t.co/9Z8Ac6skqx', 'SpaceX', 'needs', 'to', 'pass', 'through', 'a', 'deep', 'chasm', 'of', 'negative',
```

In [43]:
```python
# Again Joining the list into one string/text
tweets_tokens_text=' '.join(tweets_tokens)
tweets_tokens_text
```

tps://t.co/ɔrʊkg2Qɔur (https://t.co/ɔrʊkg2Qɔur) It \x92 s the most fun crypto . Its simplicity is its geniu
s Yup <U+0001F923> <U+0001F923> True <U+0001F923> <U+0001F923> Not that easy . This is two decades of inten
se work . Have to look at old notes , emails , texts . Yes Lessons learned Of Earth and Mars Time to tell t
he story of Tesla & SpaceX Have you read ? It \x92 s great ! ! The Second Last Kingdom https://t.co/Je4EI88
HmV (https://t.co/Je4EI88HmV) Haven \x92 t heard that name in years Dogecake YOLT https://t.co/cnOf9yjpF1
 (https://t.co/cnOf9yjpF1) That \x92 s Damian Yeah Sure The great thing about restaurants is that you get t
o hang out with strangers ! - SJM The future currency of Earth Just a scratch Much wow ! <U+0001F5A4> Desti
ny Franz was essential That said , the ship landing burn has a clear solution . My greate https://t.co/e5Wi
kiugkz (https://t.co/e5Wikiugkz) Will still use hot gas maneuvering (RCS) thrusters , https://t.co/vsO9h4
Ioed (https://t.co/vsO9h4Ioed) Higher Isp too Intuitively , it would seem so , but turbopump-fed Raptors ha
ve mu https://t.co/lBTG1sIBuC (https://t.co/lBTG1sIBuC) Yes RT : Falcon 9 will launch \x92 s SPHEREx missio

n \x96 which will collect data on more than 300 million galaxies and explore how the un RT : Falcon 9 launc
hes 60 Starlink satellites to orbit \x96 mission from pad 39A on deck https://t.co/71cuCBgPNX (https://t.c
o/71cuCBgPNX) It \x92 s true ! Yeah . By default , engine with least lever arm would shut down if all 3 are
good . Yes , but engines have a min throttle point where there is flameout risk , so la https://t.co/ThniYs
SnWn (https://t.co/ThniYsSnWn) Yes It was foolish of us not to start 3 engines & immediately shut down 1 ,
 as 2 are needed to land I am become meme , Destroyer of shorts Might actually happen Sandstorm is a master
piece We were too dumb Next time , we try pull * up * method https://t.co/J1M7WHLRDR (https://t.co/J1M7WHLR
DR) Low-key Loki No highs , no lows , only Doge No need to be a gigachad to own Dogecoin is the people \x92
s crypto ur welcome https://t.co/e2KF57KLxb (https://t.co/e2KF57KLxb) The most entertaining outcome is the

In [44]:
```python
# Remove Punctuations
no_punc_text=tweets_tokens_text.translate(str.maketrans('','',string.punctuation))
no_punc_text
```

Mars too U0001F923 U0001F923 XPrize team will manage the $100M carbon capture prize httpstcoFsw5iahEoF Everyone at Tesla receives stock My comp is all stock options which I do not take off the table That \x92 s what you \x92 re missing Back to work I go Does seem a bit high Doge appears to be inflationary but is not meaningfully so fixed of coins per unit time   whereas httpstco4UH4rARC6T Wow 1 Orbital launch tower that can stack 2 Enough Raptors for orbit booster 3 Improve ship booster mass Back to work tonight Ð is for Ðogecoin Instructional video httpstcoUEEocOfcTb The people have spoken httpstcox41oVMzTGo So cute U0001F495 Extremely misleading image as doesn \x92 t reflect true time cost to people or rain pain U0001F3B6 Who let the Doge out U0001F3B6 Hodl the rainforests   So it \x92 s finally come to this httpstcoGf0Rg2QOaF It \x92 s the most fun crypto Its simplicity is its genius Yup U0001F923 U0001F923 True U0001F923 U0001F923 Not that easy This is two decades of intense work Have to look at old notes emails texts Yes Lessons learned Of Earth and Mars Time to tell the story of Tesla SpaceX Have you read It \x92 s great   The Second Last Kingdom httpstcoJe4EI88HmV Haven \x92 t heard that name in years Dogecake YOLT httpstcocnOf9yjpF1 That \x92 s Damian Yeah Sure The great thing about restaurants is that you get to hang out with strangers SJM The future currency of Earth Just a scratch Much wow U0001F5A4 Destiny Franz was essential That said the ship landing burn has a clear solution My greate httpstcoe5Wikiugkz Will still use hot gas maneuvering RCS thrusters httpstcovsO9h4Ioed Higher Isp too Intuitively it would seem so but turbopumpfed Raptors have mu httpstcolBTG1sIBuC Yes RT Falcon 9 will launch \x92 s SPHEREx mission \x96 which will collect data on more than 300 million galaxies and explore how the un RT Falcon 9 launches 60 Starlink satellites to orbit \x96 mission from pad 39A on deck httpstco71cuCBgPNX It \x92 s true Yeah By default engine with least lever arm would shut down if all 3 are good Yes but engines have a min throttle point where there is fl

In [45]:
```python
# remove https or url within text
import re
no_url_text=re.sub(r'http\S+', '', no_punc_text)
no_url_text
```

and Mars Time to tell the story of Tesla SpaceX have you read It \x92 s great The Second East Kingdom
Haven \x92 t heard that name in years Dogecake YOLT That \x92 s Damian Yeah Sure The great thing about res
taurants is that you get to hang out with strangers   SJM The future currency of Earth Just a scratch Much
wow U0001F5A4 Destiny Franz was essential That said the ship landing burn has a clear solution My greate
Will still use hot gas maneuvering RCS thrusters   Higher Isp too Intuitively it would seem so but turb
opumpfed Raptors have mu Yes RT Falcon 9 will launch \x92 s SPHEREx mission \x96 which will collect data
on more than 300 million galaxies and explore how the un RT Falcon 9 launches 60 Starlink satellites to or
bit \x96 mission from pad 39A on deck It \x92 s true Yeah By default engine with least lever arm would
shut down if all 3 are good Yes but engines have a min throttle point where there is flameout risk so la
Yes It was foolish of us not to start 3 engines immediately shut down 1 as 2 are needed to land I am beco
me meme Destroyer of shorts Might actually happen Sandstorm is a masterpiece We were too dumb Next time w
e try pull up method Lowkey Loki No highs no lows only Doge No need to be a gigachad to own Dogecoin i
s the people \x92 s crypto ur welcome The most entertaining outcome is the most likely Doge High seas wi
nd This was a tough one RT Watch Falcon 9 launch 60 Starlink satellites U2192   Off Twitter for a while I

t will be filled with graffiti art U043C U044B U0441 U0434 U0435 U043B U0430 U0435 U043C Giga Berlin progre
ss Neuralink is working super hard to ensure implant safety is in close communication with the RT This
mission enables access for everyday people who dream of going to space RT Announcing the first commercial
astronaut mission to orbit Earth aboard Dragon U2192   If you \x92 ve worked on advanced wearables phones
or robots those skills are needed Feels weird helping make hopefully good version of Cyberpunk come true
Please consider working at Neuralink Shortterm solve brain spine injuries Longterm human AI symbiosis

In [46]:
```python
!pip install nltk
```

Requirement already satisfied: nltk in c:\users\f safrin kauser\anaconda3\lib\site-packages (3.6.1)
Requirement already satisfied: joblib in c:\users\f safrin kauser\anaconda3\lib\site-packages (from nltk) (1.
0.1)
Requirement already satisfied: tqdm in c:\users\f safrin kauser\anaconda3\lib\site-packages (from nltk) (4.59.
0)
Requirement already satisfied: click in c:\users\f safrin kauser\anaconda3\lib\site-packages (from nltk) (7.1.
2)
Requirement already satisfied: regex in c:\users\f safrin kauser\anaconda3\lib\site-packages (from nltk) (202
1.4.4)

In [47]:
```python
import nltk
nltk.download('punkt')
```

[nltk_data] Downloading package punkt to C:\Users\F SAFRIN
[nltk_data]   KAUSER\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!

Out[47]: True

In [48]:
```python
from nltk.tokenize import word_tokenize
text_tokens=word_tokenize(no_url_text)
print(text_tokens)
```

['I', '\x92', 'm', 'an', 'alien', 'Ray', 'tracing', 'on', 'Cyberpunk', 'with', 'HDR', 'is', 'nextlevel', 'Have', 'you', 'tried', 'it', 'Great', 'interview', 'Doge', 'is', 'underestimated', 'Congratulations', 'Tesla', 'China', 'for', 'amazing', 'execution', 'last', 'year', 'Now', 'on', 'to', 'the', 'next', 'for', 'even', 'more', 'Happy', 'New', 'Year', 'of', 'the', 'Ox', 'Frodo', 'was', 'the', 'underdoge', 'All', 'thought', 'he', 'would', 'fail', 'Himself', 'most', 'of', 'all', 'Haha', 'thanks', 'Indeed', 'Tweets', 'definitely', 'do', 'not', 'represent', 'realworld', 'time', 'allocation', 'The', 'most', 'entertaining', 'outcome', 'is', 'the', 'most', 'likely', 'Just', 'sent', 'some', 'Just', 'agree', 'to', 'do', 'Clubhouse', 'with', 'It', '\x92', 's', 'getting', 'real', 'Bought', 'some', 'Dogecoin', 'for', 'lil', 'X', 'so', 'he', 'can', 'be', 'a', 'toddler', 'hodler', 'He', 'definitely', 'has', 'issues', 'but', 'the', 'sentencing', 'seems', 'a', 'bit', 'high', 'Thanks', 'for', 'fixing', 'Please', 'unlock', 'my', 'account', 'U0001F923', 'U0001F923', 'This', 'is', 'true', 'power', 'haha', 'Any', 'crypto', 'wallet', 'that', 'won', '\x92', 't', 'give', 'you', 'your', 'private', 'keys', 'should', 'be', 'avoided', 'at', 'all', 'costs', 'Your', 'app', 'sucks', 'RT', 'NASA', 'has', 'selected', 'Falcon', 'Heavy', 'to', 'launch', 'the', 'first', 'two', 'elements', 'of', 'the', 'lunar', 'Gateway', 'together', 'on', 'one', 'mission', 'Yes', 'Once', 'we', 'can', 'predict', 'cash', 'flow', 'reasonably', 'well', 'Starlink', 'will', 'IPO', 'Starlink', 'is', 'a', 'staggeringly', 'difficult', 'technical', 'economic', 'endeavor', 'However', 'if', 'we', 'don', 'SpaceX', 'needs', 'to', 'pass', 'through', 'a', 'deep', 'chasm', 'of', 'negative', 'cash', 'flow', 'over', 'the', 'next', 'year', 'or', 'Lowest', 'cost', 'per', 'ton', 'of', 'carbon', 'sequestered', 'net', 'of', 'value', 'of', 'any', 'product', 'made', 'Must', 'be', 'scalable', 'to', 'g', 'It', '\x92', 's', 'meant', 'to', 'be', 'the', 'same', 'pr

In [49]:
```python
# Tokenization
import nltk
nltk.download('punkt')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to C:\Users\F SAFRIN
[nltk_data]     KAUSER\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to C:\Users\F SAFRIN
[nltk_data]     KAUSER\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[49]: True

In [50]:
```python
# Tokens count
len(text_tokens)
```

Out[50]: 17847

In [51]:
```python
# Remove Stopwords
from nltk.corpus import stopwords
my_stop_words=stopwords.words('english')

sw_list = ['\x92','rt','ye','yeah','haha','Yes','U0001F923','I']
my_stop_words.extend(sw_list)

no_stop_tokens=[word for word in text_tokens if not word in my_stop_words]
print(no_stop_tokens)
```

```
m', 'negative', 'cash', 'flow', 'next', 'year', 'Lowest', 'cost', 'per', 'ton', 'carbon', 'sequestered', 'n
et', 'value', 'product', 'made', 'Must', 'scalable', 'g', 'It', 'meant', 'price', 'countries', 'Only', 'dif
ference', 'taxes', 'shipping', 'This', 'intended', 'Earth', 'may', 'ideas', 'apply', 'Mars', 'XPrize', 'tea
m', 'manage', '100M', 'carbon', 'capture', 'prize', 'Everyone', 'Tesla', 'receives', 'stock', 'My', 'comp',
'stock', 'options', 'take', 'table', 'That', 'missing', 'Back', 'work', 'go', 'Does', 'seem', 'bit', 'hig
h', 'Doge', 'appears', 'inflationary', 'meaningfully', 'fixed', 'coins', 'per', 'unit', 'time', 'whereas',
'Wow', '1', 'Orbital', 'launch', 'tower', 'stack', '2', 'Enough', 'Raptors', 'orbit', 'booster', '3', 'Impr
ove', 'ship', 'booster', 'mass', 'Back', 'work', 'tonight', 'Ð', 'Đogecoin', 'Instructional', 'video', 'Th
e', 'people', 'spoken', 'So', 'cute', 'U0001F495', 'Extremely', 'misleading', 'image', 'reflect', 'true',
'time', 'cost', 'people', 'rain', 'pain', 'U0001F3B6', 'Who', 'let', 'Doge', 'U0001F3B6', 'Hodl', 'rainfore
sts', 'So', 'finally', 'come', 'It', 'fun', 'crypto', 'Its', 'simplicity', 'genius', 'Yup', 'True', 'Not',
'easy', 'This', 'two', 'decades', 'intense', 'work', 'Have', 'look', 'old', 'notes', 'emails', 'texts', 'Le
ssons', 'learned', 'Of', 'Earth', 'Mars', 'Time', 'tell', 'story', 'Tesla', 'SpaceX', 'Have', 'read', 'It',
'great', 'The', 'Second', 'Last', 'Kingdom', 'Haven', 'heard', 'name', 'years', 'Dogecake', 'YOLT', 'That',
'Damian', 'Yeah', 'Sure', 'The', 'great', 'thing', 'restaurants', 'get', 'hang', 'strangers', 'SJM', 'The',
'future', 'currency', 'Earth', 'Just', 'scratch', 'Much', 'wow', 'U0001F5A4', 'Destiny', 'Franz', 'essentia
l', 'That', 'said', 'ship', 'landing', 'burn', 'clear', 'solution', 'My', 'greate', 'Will', 'still', 'use',
'hot', 'gas', 'maneuvering', 'RCS', 'thrusters', 'Higher', 'Isp', 'Intuitively', 'would', 'seem', 'turbopum
pfed', 'Raptors', 'mu', 'RT', 'Falcon', '9', 'launch', 'SPHEREx', 'mission', '\x96', 'collect', 'data', '30
0', 'million', 'galaxies', 'explore', 'un', 'RT', 'Falcon', '9', 'launches', '60', 'Starlink', 'satellite
```

In [52]:
```python
# Normalize the data
lower_words=[Text.lower() for Text in no_stop_tokens]
print(lower_words[100:200])
```

['once', 'predict', 'cash', 'flow', 'reasonably', 'well', 'starlink', 'ipo', 'starlink', 'staggeringly', 'difficult', 'technical', 'economic', 'endeavor', 'however', 'spacex', 'needs', 'pass', 'deep', 'chasm', 'negative', 'cash', 'flow', 'next', 'year', 'lowest', 'cost', 'per', 'ton', 'carbon', 'sequestered', 'net', 'value', 'product', 'made', 'must', 'scalable', 'g', 'it', 'meant', 'price', 'countries', 'only', 'difference', 'taxes', 'shipping', 'this', 'intended', 'earth', 'may', 'ideas', 'apply', 'mars', 'xprize', 'team', 'manage', '100m', 'carbon', 'capture', 'prize', 'everyone', 'tesla', 'receives', 'stock', 'my', 'comp', 'stock', 'options', 'take', 'table', 'that', 'missing', 'back', 'work', 'go', 'does', 'seem', 'bit', 'high', 'doge', 'appears', 'inflationary', 'meaningfully', 'fixed', 'coins', 'per', 'unit', 'time', 'whereas', 'wow', '1', 'orbital', 'launch', 'tower', 'stack', '2', 'enough', 'raptors', 'orbit', 'booster']

In [53]:
```python
# Stemming (Optional)
from nltk.stem import PorterStemmer
ps=PorterStemmer()
stemmed_tokens=[ps.stem(word) for word in lower_words]
print(stemmed_tokens[100:200])
```

['onc', 'predict', 'cash', 'flow', 'reason', 'well', 'starlink', 'ipo', 'starlink', 'staggeringli', 'difficult', 'technic', 'econom', 'endeavor', 'howev', 'spacex', 'need', 'pass', 'deep', 'chasm', 'neg', 'cash', 'flow', 'next', 'year', 'lowest', 'cost', 'per', 'ton', 'carbon', 'sequest', 'net', 'valu', 'product', 'made', 'must', 'scalabl', 'g', 'it', 'meant', 'price', 'countri', 'onli', 'differ', 'tax', 'ship', 'thi', 'intend', 'earth', 'may', 'idea', 'appli', 'mar', 'xprize', 'team', 'manag', '100m', 'carbon', 'captur', 'prize', 'everyon', 'tesla', 'receiv', 'stock', 'my', 'comp', 'stock', 'option', 'take', 'tabl', 'that', 'miss', 'back', 'work', 'go', 'doe', 'seem', 'bit', 'high', 'doge', 'appear', 'inflationari', 'meaning', 'fix', 'coin', 'per', 'unit', 'time', 'wherea', 'wow', '1', 'orbit', 'launch', 'tower', 'stack', '2', 'enough', 'raptor', 'orbit', 'booster']

In [61]:
```
1 !pip install spacy
```

Requirement already satisfied: spacy in c:\users\f safrin kauser\anaconda3\lib\site-packages (3.2.1)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (2.25.1)
Requirement already satisfied: setuptools in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (52.0.0.post20210125)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (1.8.2)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (2.0.6)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (3.3.0)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (3.0.6)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (1.0.1)
Requirement already satisfied: numpy>=1.15.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (1.20.1)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (0.7.5)
Requirement already satisfied: thinc<8.1.0,>=8.0.12 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (8.0.13)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.8 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (3.0.8)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (2.0.6)
Requirement already satisfied: srsly<3.0.0,>=2.4.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (2.4.2)
Requirement already satisfied: pathy>=0.3.5 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (0.6.1)
Requirement already satisfied: jinja2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (2.11.3)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (1.0.6)
Requirement already satisfied: packaging>=20.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (20.9)
Requirement already satisfied: typer<0.5.0,>=0.3.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (0.4.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy) (4.59.0)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (f

```
rom spacy) (0.9.0)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
packaging>=20.0->spacy) (2.4.7)
Requirement already satisfied: smart-open<6.0.0,>=5.0.0 in c:\users\f safrin kauser\anaconda3\lib\site-package
s (from pathy>=0.3.5->spacy) (5.2.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\f safrin kauser\anaconda3\lib\site-packa
ges (from pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4->spacy) (3.7.4.3)
Requirement already satisfied: idna<3,>=2.5 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from requ
ests<3.0.0,>=2.13.0->spacy) (2.10)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from requests<3.0.0,>=2.13.0->spacy) (1.26.4)
Requirement already satisfied: chardet<5,>=3.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy) (4.0.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fro
m requests<3.0.0,>=2.13.0->spacy) (2020.12.5)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om typer<0.5.0,>=0.3.0->spacy) (7.1.2)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
jinja2->spacy) (1.1.1)
```

In [66]:
```
1  !python -m spacy download en_core_web_lg
```

```
Collecting en-core-web-lg==3.2.0

2022-01-13 22:45:46.238277: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dyn
amic library 'cudart64_110.dll'; dlerror: cudart64_110.dll not found
2022-01-13 22:45:46.238420: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror i
f you do not have a GPU set up on your machine.
```

  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-3.2.0/en_core_web_lg-
3.2.0-py3-none-any.whl (https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-3.2.0/en_co
re_web_lg-3.2.0-py3-none-any.whl) (777.4 MB)
Requirement already satisfied: spacy<3.3.0,>=3.2.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om en-core-web-lg==3.2.0) (3.2.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.8 in c:\users\f safrin kauser\anaconda3\lib\site-packa
ges (from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (3.0.8)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\f safrin kauser\anaconda3\lib\site-packag
es (from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (1.0.6)
Requirement already satisfied: typer<0.5.0,>=0.3.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (0.4.0)
Requirement already satisfied: pathy>=0.3.5 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spac
y<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (0.6.1)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (f
rom spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (0.9.0)
Requirement already satisfied: setuptools in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy<
3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (52.0.0.post20210125)
Requirement already satisfied: jinja2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy<3.3.
0,>=3.2.0->en-core-web-lg==3.2.0) (2.11.3)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (3.3.0)
Requirement already satisfied: numpy>=1.15.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spa
cy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (1.20.1)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (2.0.6)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (2.0.6)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fro
m spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (0.7.5)
Requirement already satisfied: srsly<3.0.0,>=2.4.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (2.4.2)
Requirement already satisfied: packaging>=20.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from s
pacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (20.9)
```

```
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4 in c:\users\f safrin kauser\anaconda3\lib
\site-packages (from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (1.8.2)
Requirement already satisfied: thinc<8.1.0,>=8.0.12 in c:\users\f safrin kauser\anaconda3\lib\site-packages (f
rom spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (8.0.13)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\f safrin kauser\anaconda3\lib\site-pack
ages (from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (1.0.1)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages
 (from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (3.0.6)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages
 (from spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (2.25.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (4.59.0)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
 packaging>=20.0->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (2.4.7)
Requirement already satisfied: smart-open<6.0.0,>=5.0.0 in c:\users\f safrin kauser\anaconda3\lib\site-package
s (from pathy>=0.3.5->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (5.2.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\f safrin kauser\anaconda3\lib\site-packa
ges (from pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (3.7.4.3)
Requirement already satisfied: idna<3,>=2.5 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from requ
ests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fro
m requests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (2020.12.5)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages
 (from requests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (1.26.4)
Requirement already satisfied: chardet<5,>=3.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (4.0.0)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om typer<0.5.0,>=0.3.0->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (7.1.2)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
 jinja2->spacy<3.3.0,>=3.2.0->en-core-web-lg==3.2.0) (1.1.1)
Installing collected packages: en-core-web-lg
Successfully installed en-core-web-lg-3.2.0
[+] Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [67]:
```
1  !python -m spacy download en
```

Collecting en-core-web-sm==3.2.0

2022-01-13 22:48:13.004548: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dyn
amic library 'cudart64_110.dll'; dlerror: cudart64_110.dll not found
2022-01-13 22:48:13.004637: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror i
f you do not have a GPU set up on your machine.

  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.2.0/en_core_web_sm-
3.2.0-py3-none-any.whl (https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.2.0/en_co
re_web_sm-3.2.0-py3-none-any.whl) (13.9 MB)
Requirement already satisfied: spacy<3.3.0,>=3.2.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om en-core-web-sm==3.2.0) (3.2.1)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2.0.6)
Requirement already satisfied: pathy>=0.3.5 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spac
y<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (0.6.1)
Requirement already satisfied: setuptools in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy<
3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (52.0.0.post20210125)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (4.59.0)
Requirement already satisfied: jinja2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spacy<3.3.
0,>=3.2.0->en-core-web-sm==3.2.0) (2.11.3)
Requirement already satisfied: packaging>=20.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from s
pacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (20.9)
Requirement already satisfied: thinc<8.1.0,>=8.0.12 in c:\users\f safrin kauser\anaconda3\lib\site-packages (f
rom spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (8.0.13)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (f
rom spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (0.9.0)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2.25.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.8 in c:\users\f safrin kauser\anaconda3\lib\site-packa
ges (from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (3.0.8)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (3.3.0)
Requirement already satisfied: numpy>=1.15.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from spa
cy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.20.1)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\f safrin kauser\anaconda3\lib\site-pack
ages (from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.0.1)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages
 (from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (3.0.6)

```
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\f safrin kauser\anaconda3\lib\site-packages
(from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2.0.6)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fro
m spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (0.7.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\f safrin kauser\anaconda3\lib\site-packag
es (from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.0.6)
Requirement already satisfied: typer<0.5.0,>=0.3.0 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (0.4.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4 in c:\users\f safrin kauser\anaconda3\lib
\site-packages (from spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.8.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2.4.2)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
 packaging>=20.0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2.4.7)
Requirement already satisfied: smart-open<6.0.0,>=5.0.0 in c:\users\f safrin kauser\anaconda3\lib\site-package
s (from pathy>=0.3.5->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (5.2.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\f safrin kauser\anaconda3\lib\site-packa
ges (from pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (3.7.4.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages
 (from requests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.26.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fro
m requests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2020.12.5)
Requirement already satisfied: chardet<5,>=3.0.2 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (4.0.0)
Requirement already satisfied: idna<3,>=2.5 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from requ
ests<3.0.0,>=2.13.0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2.10)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\f safrin kauser\anaconda3\lib\site-packages (fr
om typer<0.5.0,>=0.3.0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (7.1.2)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\f safrin kauser\anaconda3\lib\site-packages (from
 jinja2->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.1.1)
Installing collected packages: en-core-web-sm
Successfully installed en-core-web-sm-3.2.0
[!] As of spaCy v3.0, shortcuts like 'en' are deprecated. Please use the full
pipeline package name 'en_core_web_sm' instead.
[+] Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

In [68]:
```python
# Lemmatization
nlp=spacy.load(name='en_core_web_sm')
doc=nlp(' '.join(lower_words))
print(doc)
```

alien ray tracing cyberpunk hdr nextlevel have tried great interview doge underestimated congratulations te
sla china amazing execution last year now next even happy new year ox frodo underdoge all thought would fai
l himself haha thanks indeed tweets definitely represent realworld time allocation the entertaining outcome
likely just sent just agree clubhouse it getting real bought dogecoin lil x toddler hodler he definitely is
sues sentencing seems bit high thanks fixing please unlock account this true power any crypto wallet give p
rivate keys avoided costs your app sucks rt nasa selected falcon heavy launch first two elements lunar gate
way together one mission once predict cash flow reasonably well starlink ipo starlink staggeringly difficul
t technical economic endeavor however spacex needs pass deep chasm negative cash flow next year lowest cost
per ton carbon sequestered net value product made must scalable g it meant price countries only difference
taxes shipping this intended earth may ideas apply mars xprize team manage 100m carbon capture prize everyo
ne tesla receives stock my comp stock options take table that missing back work go does seem bit high doge
appears inflationary meaningfully fixed coins per unit time whereas wow 1 orbital launch tower stack 2 enou
gh raptors orbit booster 3 improve ship booster mass back work tonight ð ðogecoin instructional video the p
eople spoken so cute u0001f495 extremely misleading image reflect true time cost people rain pain u0001f3b6
who let doge u0001f3b6 hodl rainforests so finally come it fun crypto its simplicity genius yup true not ea
sy this two decades intense work have look old notes emails texts lessons learned of earth mars time tell s
tory tesla spacex have read it great the second last kingdom haven heard name years dogecake yolt that dami
an yeah sure the great thing restaurants get hang strangers sjm the future currency earth just scratch much
wow u0001f5a4 destiny franz essential that said ship landing burn clear solution my greate will still use h

In [69]:
```python
lemmas=[token.lemma_ for token in doc]
print(lemmas)
```

n', 'sandstorm', 'masterpiece', 'we', 'dumb', 'next', 'time', 'try', 'pull', 'method', 'lowkey', 'loki', 'no', 'highs', 'low', 'doge', 'no', 'need', 'gigachad', 'dogecoin', 'people', 'crypto', 'ur', 'welcome', 'the', 'entertaining', 'outcome', 'likely', 'doge', 'high', 'sea', 'wind', 'this', 'tough', 'one', 'rt', 'watch', 'falcon', '9', 'launch', '60', 'starlink', 'satellite', 'u2192', 'off', 'twitter', 'it', 'fill', 'graffiti', 'art', 'u043c', 'u044b', 'u0441', 'u0434', 'u0435', 'u043b', 'u0430', 'u0435', 'u043c', 'giga', 'berl

in', 'progress', 'neuralink', 'work', 'super', 'hard', 'ensure', 'implant', 'safety', 'close', 'communication', 'rt', 'this', 'mission', 'enable', 'access', 'everyday', 'people', 'dream', 'go', 'space', 'rt', 'announce', 'first', 'commercial', 'astronaut', 'mission', 'orbit', 'earth', 'aboard', 'dragon', 'u2192', 'if', 'work', 'advanced', 'wearable', 'phone', 'robot', 'skill', 'need', 'feel', 'weird', 'helping', 'make', 'hopefully', 'good', 'version', 'cyberpunk', 'come', 'true', 'please', 'consider', 'work', 'neuralink', 'shortterm', 'solve', 'brain', 'spine', 'injury', 'longterm', 'human', 'ai', 'symbiosis', 'latte', 'on', 'clubhouse', 'tonight', '10', 'pm', 'la', 'time', 'tom', 'great', 'story', 'experim', 'rt', 'launch', 'alert', 'u0001f680', 'target', 'early', 'april', '20', 'launch', 'second', 'crew', 'rotation', 'mis', 'he', 'become', 'big', 'fan', 'methane', 'that', 'sound', 'correct', 'tom', 'certainly', 'deserve', 'lot', 'cre', 'tom', 'great', 'instrumental', 'develop', 'early', 'ver', 't', 'w', '15', 'accelerate', 'unusually', 'fast', 'high', 't', 'w', 'important', 'reusable', 'v', 'no', 'escaping', 'read', 'whole', 'article', 'warm', 'sunny', 'day', 'snowy', 'mountain', 'what', 'beautiful', 'day', 'la', 'yeah', 'dr', 'frankensteen', 'never', 'use', 'guy', 'he', 'give', 'talk', 'spacex', 'only', 'halo', 'sure', 'hope', 'true', 'tesla', 'spacex', 'cryoproof', 'install', 'engine', 'starship', 'sn9', 'sn10', 'in', 'retrospect', 'inevitable', 'great', 'shot', 'live', 'sword', 'die', 'sword', 'entropy', 'buy', 'hold', 'company', 'make', 'good', 'produce', 'service', 'lo

In [70]:
```python
1  clean_tweets=' '.join(lemmas)
2  clean_tweets
```

Out[70]: 'alien ray trace cyberpunk hdr nextlevel have try great interview doge underestimate congratulation tesla c
hina amazing execution last year now next even happy new year ox frodo underdoge all thought would fail him
self haha thank indeed tweet definitely represent realworld time allocation the entertaining outcome likely
just send just agree clubhouse it get real buy dogecoin lil x toddler hodler he definitely issue sentencing
seem bit high thank fix please unlock account this true power any crypto wallet give private key avoid cost
your app suck rt nasa select falcon heavy launch first two element lunar gateway together one mission once
predict cash flow reasonably well starlink ipo starlink staggeringly difficult technical economic endeavor
however spacex need pass deep chasm negative cash flow next year low cost per ton carbon sequester net valu
e product make must scalable g it mean price country only difference taxis ship this intend earth may idea
apply mars xprize team manage 100 m carbon capture prize everyone tesla receive stock my comp stock option
take table that miss back work go do seem bit high doge appear inflationary meaningfully fix coin per unit
time whereas wow 1 orbital launch tower stack 2 enough raptor orbit booster 3 improve ship booster mass bac
k work tonight ð ðogecoin instructional video the people speak so cute u0001f495 extremely misleading image
reflect true time cost people rain pain u0001f3b6 who let doge u0001f3b6 hodl rainforest so finally come it
fun crypto its simplicity genius yup true not easy this two decade intense work have look old note email te
xt lesson learn of earth mar time tell story tesla spacex have read it great the second last kingdom haven
hear name year dogecake yolt that damian yeah sure the great thing restaurant get hang stranger sjm the fut
ure currency earth just scratch much wow u0001f5a4 destiny franz essential that say ship landing burn clear
solution my greate will still use hot gas maneuvering rcs thruster high isp intuitively would seem turbopum

Feature Extaction

1. Using CountVectorizer

In [71]:
```python
1  from sklearn.feature_extraction.text import CountVectorizer
2  cv=CountVectorizer()
3  tweetscv=cv.fit_transform(lemmas)
```

In [72]:
```python
1   print(cv.vocabulary_)
```

```
{'alien': 194, 'ray': 2355, 'trace': 3020, 'cyberpunk': 766, 'hdr': 1362, 'nextlevel': 1965, 'have': 1357,
'try': 3051, 'great': 1308, 'interview': 1561, 'doge': 890, 'underestimate': 3184, 'congratulation': 672,
'tesla': 2929, 'china': 581, 'amazing': 216, 'execution': 1058, 'last': 1645, 'year': 3401, 'now': 1993, 'n
ext': 1964, 'even': 1031, 'happy': 1344, 'new': 1962, 'ox': 2094, 'frodo': 1208, 'underdoge': 3183, 'all':
196, 'thought': 2964, 'would': 3387, 'fail': 1092, 'himself': 1405, 'haha': 1332, 'thank': 2938, 'indeed':
1502, 'tweet': 3066, 'definitely': 803, 'represent': 2422, 'realworld': 2370, 'time': 2980, 'allocation': 1
98, 'the': 2943, 'entertaining': 1001, 'outcome': 2073, 'likely': 1699, 'just': 1603, 'send': 2565, 'agre
e': 179, 'clubhouse': 610, 'it': 1580, 'get': 1267, 'real': 2365, 'buy': 501, 'dogecoin': 892, 'lil': 1700,
'toddler': 2991, 'hodler': 1411, 'he': 1363, 'issue': 1579, 'sentencing': 2571, 'seem': 2557, 'bit': 418,
'high': 1397, 'fix': 1154, 'please': 2188, 'unlock': 3208, 'account': 137, 'this': 2960, 'true': 3046, 'pow
er': 2224, 'any': 239, 'crypto': 749, 'wallet': 3292, 'give': 1279, 'private': 2262, 'key': 1615, 'avoid':
337, 'cost': 713, 'your': 3412, 'app': 250, 'suck': 2830, 'rt': 2492, 'nasa': 1941, 'select': 2558, 'falco
n': 1096, 'heavy': 1373, 'launch': 1654, 'first': 1150, 'two': 3071, 'element': 970, 'lunar': 1757, 'gatewa
y': 1251, 'together': 2992, 'one': 2040, 'mission': 1876, 'once': 2039, 'predict': 2236, 'cash': 539, 'flo
w': 1165, 'reasonably': 2374, 'well': 3324, 'starlink': 2768, 'ipo': 1571, 'staggeringly': 2760, 'difficul
t': 856, 'technical': 2909, 'economic': 946, 'endeavor': 989, 'however': 1436, 'spacex': 2724, 'need': 195
1, 'pass': 2126, 'deep': 800, 'chasm': 573, 'negative': 1953, 'low': 1749, 'per': 2148, 'ton': 2997, 'carbo
n': 529, 'sequester': 2578, 'net': 1956, 'value': 3240, 'product': 2276, 'make': 1775, 'must': 1930, 'scala
ble': 2526, 'mean': 1819, 'price': 2251, 'country': 723, 'only': 2043, 'difference': 853, 'taxis': 2903, 's
```

In [73]:
```python
1   print(cv.get_feature_names()[100:200])
```

```
['74', '78', '7th', '90', '9007', '922', '948', '95', '99', 'aber', 'able', 'abo', 'aboard', 'abort', 'about',
'above', 'absence', 'absolute', 'absolutely', 'absorb', 'absorption', 'absurd', 'absurdly', 'ac', 'academia',
'accel', 'accelera', 'accelerate', 'acceleration', 'accelerator', 'accept', 'acceptable', 'access', 'accessibl
e', 'accident', 'accidental', 'accommodate', 'account', 'accura', 'accuracy', 'accurate', 'ace', 'achieve', 'a
chievement', 'achy', 'acquisition', 'across', 'action', 'active', 'activity', 'actual', 'actuall', 'actually',
'actuary', 'adagio', 'add', 'additive', 'address', 'administer', 'adult', 'advanc', 'advance', 'advanced', 'ad
vantage', 'adventure', 'advertise', 'advice', 'advise', 'aero', 'afb', 'affair', 'affect', 'affordable', 'afri
ca', 'after', 'afternoon', 'age', 'ago', 'agony', 'agree', 'ah', 'ahead', 'ahem', 'ai', 'aim', 'air', 'aircraf
t', 'airplane', 'ak', 'aka', 'alert', 'alexander', 'algo', 'algorithm', 'alien', 'align', 'all', 'allocati',
'allocation', 'allow']
```

In [74]:
```python
print(tweetscv.toarray()[100:200])
```

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

In [75]:
```python
print(tweetscv.toarray().shape)
```

```
(11486, 3421)
```

2. CountVectorizer with N-grams (Bigrams & Trigrams)

In [76]:
```python
cv_ngram_range=CountVectorizer(analyzer='word',ngram_range=(1,3),max_features=100)
bow_matrix_ngram=cv_ngram_range.fit_transform(lemmas)
```

In [77]:
```python
print(cv_ngram_range.get_feature_names())
print(bow_matrix_ngram.toarray())
```

```
['actually', 'ai', 'also', 'back', 'big', 'booster', 'car', 'come', 'cool', 'could', 'crew', 'day', 'design',
 'do', 'dragon', 'earth', 'engine', 'even', 'ever', 'exactly', 'falcon', 'first', 'flight', 'fsd', 'future', 'g
et', 'go', 'good', 'great', 'haha', 'hard', 'high', 'if', 'it', 'just', 'land', 'launch', 'life', 'like', 'loo
k', 'lot', 'love', 'make', 'many', 'mar', 'maybe', 'mission', 'model', 'much', 'need', 'new', 'next', 'no', 'n
ot', 'ok', 'one', 'part', 'people', 'point', 'pretty', 'probably', 'production', 'right', 'rocket', 'rt', 'sa
y', 'seem', 'soon', 'space', 'spacex', 'starlink', 'starship', 'still', 'super', 'sure', 'take', 'tesla', 'tes
t', 'thank', 'that', 'the', 'there', 'they', 'think', 'this', 'time', 'true', 'try', 'ufe0f', 'use', 'way', 'w
e', 'week', 'well', 'will', 'work', 'would', 'yeah', 'year', 'you']
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

3. TF-IDF Vectorizer

In [78]:
```python
from sklearn.feature_extraction.text import TfidfVectorizer
tfidfv_ngram_max_features=TfidfVectorizer(norm='l2',analyzer='word',ngram_range=(1,3),max_features=500)
tfidf_matix_ngram=tfidfv_ngram_max_features.fit_transform(lemmas)
```

In [79]:
```python
1  print(tfidfv_ngram_max_features.get_feature_names())
2  print(tfidf_matix_ngram.toarray())
```

```
['10', '100', '13', '20', '2020', '30', '300', '40', '4d', '60', 'able', 'absolutely', 'access', 'accurat
e', 'actually', 'add', 'ago', 'agree', 'ai', 'aim', 'air', 'all', 'almost', 'already', 'also', 'although',
'always', 'amazing', 'and', 'another', 'appear', 'apply', 'appreciate', 'approval', 'arm', 'around', 'artic
le', 'as', 'astronaut', 'at', 'autopilot', 'awesome', 'back', 'bad', 'bar', 'base', 'battery', 'be', 'becom
e', 'berlin', 'beta', 'big', 'bit', 'boca', 'body', 'booster', 'brain', 'bring', 'bs', 'build', 'burn', 'bu
t', 'buy', 'cake', 'call', 'camera', 'can', 'car', 'cargo', 'cast', 'catch', 'cell', 'ch4', 'chance', 'chan
ge', 'china', 'civilization', 'close', 'cold', 'color', 'come', 'company', 'complete', 'computer', 'confir
m', 'congratulation', 'control', 'cool', 'cost', 'could', 'course', 'cov', 'crazy', 'create', 'crew', 'crit
ical', 'curevac', 'cyberpunk', 'cybertruck', 'datum', 'day', 'deep', 'definitely', 'delivery', 'design', 'd
etail', 'difference', 'different', 'do', 'dock', 'doge', 'dojo', 'dragon', 'drive', 'droneship', 'early',
'earth', 'easy', 'electric', 'enable', 'end', 'energy', 'engine', 'engineer', 'engineering', 'enough', 'ent
ire', 'especially', 'essential', 'esthetic', 'et', 'even', 'ever', 'every', 'everyone', 'everything', 'exac
tly', 'expect', 'extreme', 'extremely', 'fair', 'falcon', 'far', 'fast', 'favorite', 'feel', 'fine', 'fir
e', 'first', 'fix', 'flap', 'flight', 'flow', 'fly', 'for', 'fsd', 'ft', 'full', 'fully', 'fun', 'fundament
al', 'future', 'game', 'gas', 'gene', 'get', 'giant', 'giga', 'give', 'glad', 'go', 'goal', 'good', 'govern
ment', 'great', 'haha', 'half', 'happen', 'hard', 'have', 'he', 'head', 'hear', 'heavy', 'help', 'high', 'h
old', 'home', 'hop', 'hope', 'hopefully', 'hot', 'hour', 'human', 'idea', 'if', 'imo', 'important', 'impres
sive', 'improve', 'improvement', 'in', 'increase', 'indeed', 'inference', 'interesting', 'irony', 'it', 'ju
st', 'km', 'know', 'la', 'lab', 'land', 'landing', 'large', 'last', 'later', 'launch', 'least', 'leave', 'l
eg', 'less', 'let', 'level', 'life', 'liftoff', 'like', 'likely', 'limit', 'liquid', 'literally', 'little',
'live', 'long', 'longterm', 'look', 'lose', 'lot', 'love', 'low', 'machine', 'main', 'major', 'make', 'man
y', 'mar', 'mass', 'matter', 'may', 'maybe', 'mean', 'might', 'mile', 'million', 'min', 'mind', 'minute',
'mission', 'model', 'money', 'month', 'moon', 'most', 'mostly', 'motor', 'mount', 'move', 'movie', 'much',
'must', 'my', 'name', 'nasa', 'need', 'net', 'neural', 'neuralink', 'never', 'new', 'news', 'next', 'nigh
t', 'no', 'nosecone', 'not', 'note', 'nothing', 'now', 'number', 'of', 'ok', 'old', 'one', 'only', 'open',
'operational', 'option', 'orbit', 'orbital', 'original', 'pad', 'part', 'pass', 'pcr', 'people', 'per', 'pe
rson', 'physics', 'place', 'play', 'please', 'pm', 'point', 'positive', 'possible', 'potential', 'power',
'pressure', 'pretty', 'price', 'primarily', 'prior', 'probability', 'probably', 'problem', 'product', 'prod
uction', 'progress', 'propellant', 'prototype', 'provide', 'public', 'push', 'put', 'quite', 'range', 'rapt
or', 'reach', 'read', 'real', 'really', 'release', 'require', 'return', 'review', 'right', 'rise', 'roadste
r', 'rocket', 'rt', 'safe', 'safety', 'satellite', 'say', 'scale', 'second', 'see', 'seem', 'sell', 'send',
'sense', 'separation', 'service', 'set', 'several', 'ship', 'short', 'shot', 'should', 'show', 'single', 's
ite', 'slightly', 'slow', 'small', 'sn5', 'sn8', 'sn9', 'so', 'software', 'solar', 'solution', 'solve', 'so
me', 'something', 'song', 'soon', 'sorry', 'sound', 'south', 'space', 'spacecraft', 'spacex', 'speed', 'spl
ashdown', 'stage', 'starlink', 'starship', 'start', 'static', 'station', 'steel', 'step', 'still', 'stock',
'story', 'super', 'support', 'sure', 'symptom', 'system', 'take', 'talk', 'tank', 'target', 'tbh', 'team',
'technically', 'technology', 'tesla', 'test', 'testing', 'texas', 'thank', 'that', 'the', 'there', 'they',
'thing', 'think', 'this', 'three', 'thrust', 'time', 'to', 'today', 'tomorrow', 'ton', 'tonight', 'top', 't
otal', 'tough', 'traffic', 'training', 'travel', 'true', 'try', 'tunnel', 'turn', 'twitter', 'two', 'u0001f
```

```
3b6', 'u0001f440', 'u0001f495', 'u0001f525', 'u0001f5a4', 'u043b', 'u043e', 'u2192', 'ufe0f', 'understand',
'update', 'upgrade', 'ur', 'use', 'useful', 'vaccine', 'value', 'vehicle', 'version', 'very', 'video', 'vie
w', 'volume', 'vs', 'wait', 'want', 'watch', 'way', 'we', 'week', 'welcome', 'well', 'what', 'will', 'witho
ut', 'word', 'work', 'world', 'worth', 'would', 'wow', 'yeah', 'year', 'yes', 'yet', 'you', 'yup', 'zero']
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

Generate Word Cloud

In [80]:

```python
# Define a function to plot word cloud
def plot_cloud(wordcloud):
    plt.figure(figsize=(40,30))
    plt.imshow(wordcloud)
    plt.axis('off')

# Generate Word Cloud
STOPWORDS.add('pron')
STOPWORDS.add('rt')
STOPWORDS.add('yeah')
wordcloud=WordCloud(width=3000,height=2000,background_color='black',max_words=50,
                    colormap='Set1',stopwords=STOPWORDS).generate(clean_tweets)
plot_cloud(wordcloud)
```

Named Entity Recognition (NER)

In [82]:

```python
# Parts Of Speech (POS) Tagging
nlp=spacy.load('en_core_web_sm')

one_block=clean_tweets
doc_block=nlp(one_block)
spacy.displacy.render(doc_block,style='ent',jupyter=True)
```

indeed tweet definitely represent realworld time allocation the entertaining outcome likely just send just agree clubhouse it get real

buy dogecoin   lil x toddler **PRODUCT**   hodler he definitely issue sentencing seem bit high thank fix please unlock account this true

power any crypto wallet give private key avoid cost your app suck rt   nasa **ORG**   select falcon heavy launch   first **ORDINAL**   two

CARDINAL   element lunar gateway together   one **CARDINAL**   mission once predict cash flow reasonably well starlink ipo starlink

staggeringly difficult technical economic endeavor however spacex need pass deep chasm negative cash flow   next year **DATE**

low cost per ton carbon sequester net value product make must scalable g it mean price country only difference taxis ship this intend

earth may idea apply mars xprize team manage   100 m **QUANTITY**   carbon capture prize everyone tesla receive stock my comp

stock option take table that miss back work go do seem bit high doge appear inflationary meaningfully fix coin per unit time whereas

wow   1 **CARDINAL**   orbital launch tower stack   2 **CARDINAL**   enough raptor orbit booster   3 **CARDINAL**   improve ship booster

In [83]:
```python
for token in doc_block[100:200]:
    print(token,token.pos_)
```

```
once ADV
predict VERB
cash NOUN
flow NOUN
reasonably ADV
well ADV
starlink VERB
ipo NOUN
starlink NOUN
staggeringly ADV
difficult ADJ
technical ADJ
economic ADJ
endeavor NOUN
however ADV
spacex VERB
need NOUN
pass VERB
deep ADJ
```

In [84]:
```python
# Filtering the nouns and verbs only
nouns_verbs=[token.text for token in doc_block if token.pos_ in ('NOUN','VERB')]
print(nouns_verbs[100:200])
```

```
['time', 'launch', 'tower', 'stack', 'raptor', 'orbit', 'booster', 'improve', 'ship', 'booster', 'mass', 'wor
k', 'tonight', 'ðogecoin', 'video', 'people', 'speak', 'cute', 'image', 'reflect', 'time', 'cost', 'people',
'rain', 'pain', 'u0001f3b6', 'let', 'come', 'fun', 'crypto', 'simplicity', 'genius', 'yup', 'decade', 'work',
'look', 'note', 'email', 'text', 'lesson', 'learn', 'earth', 'time', 'tell', 'story', 'tesla', 'spacex', 'rea
d', 'hear', 'name', 'year', 'dogecake', 'yolt', 'thing', 'restaurant', 'get', 'hang', 'stranger', 'sjm', 'curr
ency', 'earth', 'scratch', 'u0001f5a4', 'destiny', 'franz', 'say', 'ship', 'landing', 'burn', 'solution', 'gre
ate', 'use', 'gas', 'maneuvering', 'rcs', 'thruster', 'seem', 'turbopumpfed', 'raptor', 'falcon', 'launch', 'c
ollect', 'galaxy', 'explore', 'launch', 'starlink', 'satellite', 'orbit', 'mission', 'pad', '39a', 'deck', 'de
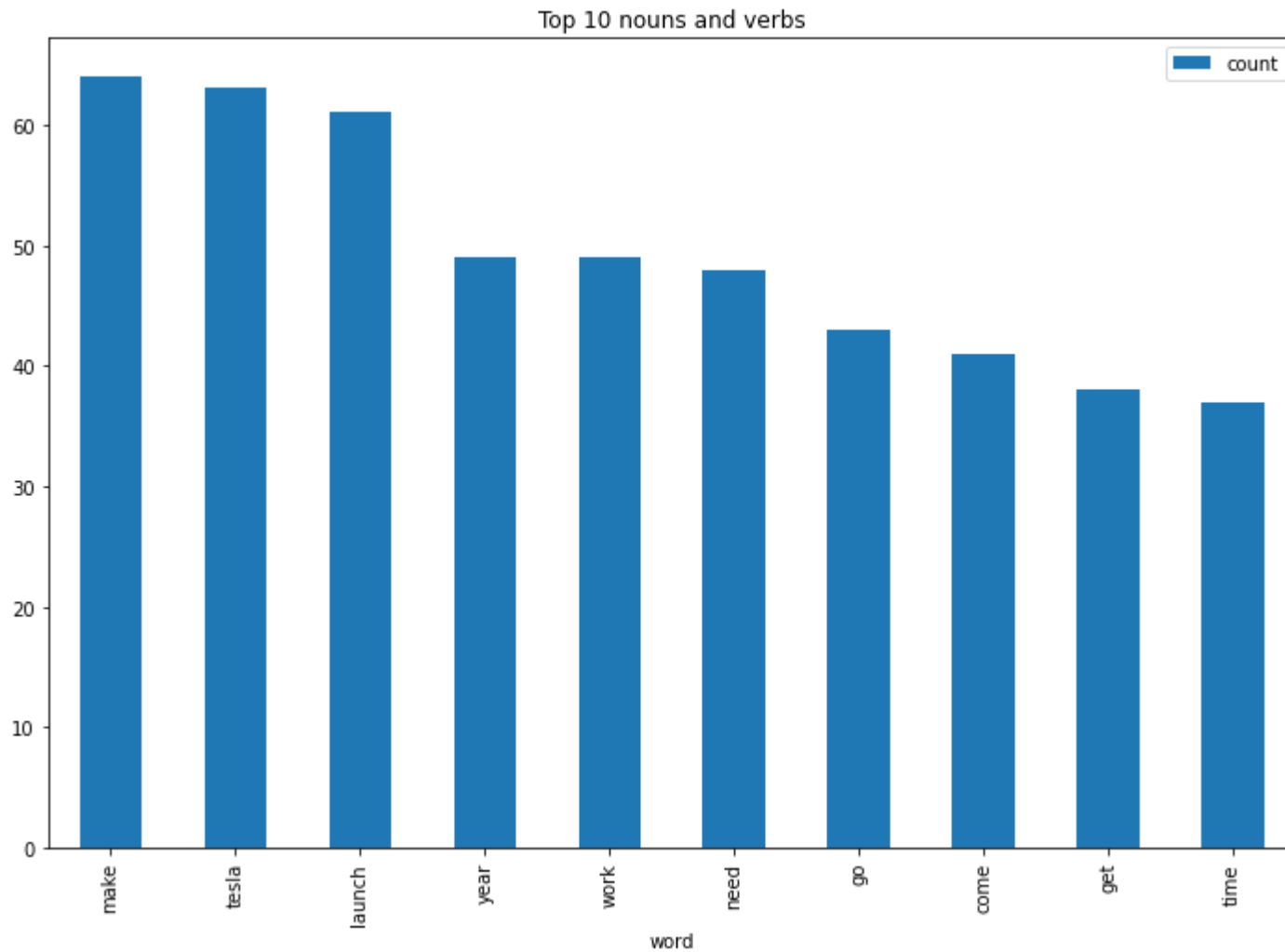fault', 'engine', 'lever', 'arm', 'shut', 'engine', 'min', 'throttle']
```

In [85]:

```python
# Counting the noun & verb tokens
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer()

X=cv.fit_transform(nouns_verbs)
sum_words=X.sum(axis=0)

words_freq=[(word,sum_words[0,idx]) for word,idx in cv.vocabulary_.items()]
words_freq=sorted(words_freq, key=lambda x: x[1], reverse=True)

wd_df=pd.DataFrame(words_freq)
wd_df.columns=['word','count']
wd_df[0:10] # viewing top ten results
```

Out[85]:

	word	count
0	make	64
1	tesla	63
2	launch	61
3	year	49
4	work	49
5	need	48
6	go	43
7	come	41
8	get	38
9	time	37

```
In [86]:    1  # Visualizing results (Barchart for top 10 nouns + verbs)
            2  wd_df[0:10].plot.bar(x='word',figsize=(12,8),title='Top 10 nouns and verbs');
```


Top 10 nouns and verbs

Emotion Mining - Sentiment Analysis

In [87]:
```python
from nltk import tokenize
sentences=tokenize.sent_tokenize(' '.join(tweets))
sentences
```

ing real Bought some Dogecoin for lil X, so he can be a toddler hodler @JoshHannode he definitely has issue
s, but the sentencing seems a bit high @freewalletorg Thanks for fixing @freewalletorg Please unlock my acc
ount @AstroJordy <U+0001F923><U+0001F923> This is true power haha https://t.co/Fc9uhQSd7O (https://t.co/Fc9
uhQSd7O) @freewalletorg Any crypto wallet that won\x92t give you your private keys should be avoided at all
costs @freewalletorg Your app sucks RT @SpaceX: NASA has selected Falcon Heavy to launch the first two elem
ents of the lunar Gateway together on one mission!',
 'https://t.co/3pWt @ajtourville Yes @BLKMDL3 @RationalEtienne @Adamklotz_ Once we can predict cash flow re
asonably well, Starlink will IPO @RationalEtienne @Adamklotz_ Starlink is a staggeringly difficult technica
l & economic endeavor.',
 'However, if we don\x85 https://t.co/9Z8Ac6skqx (https://t.co/9Z8Ac6skqx) @RationalEtienne @Adamklotz_ Spa
ceX needs to pass through a deep chasm of negative cash flow over the next year or\x85 https://t.co/7J1c92h
djl (https://t.co/7J1c92hdjl) @ID_AA_Carmack Lowest cost per ton of carbon sequestered, net of value of any
product made.',
 'Must be scalable to g\x85 https://t.co/XMyI7qWSgw (https://t.co/XMyI7qWSgw) @Adamklotz_ It\x92s meant to
 be the same price in all countries.',
 'Only difference should be taxes & shipping.',
 '@tobyliiiiiiiiii This is intended for Earth, but there may be some ideas that apply to Mars too @CryptoSh
rikar @CoinDesk @Tesla @Dan_Z_Palmer <U+0001F923><U+0001F923> XPrize team will manage the $100M carbon capt
ure prize https://t.co/fSw5IanL0r (https://t.co/fSw5IanL0r) @SRuhle Everyone at Tesla receives stock.',
 'My comp is all stock/options, which I do not take off the table.',

In [88]:
```python
sent_df=pd.DataFrame(sentences,columns=['sentence'])
sent_df
```

Out[88]:

	sentence
0	@kunalb11 I m an alien @ID_AA_Carmack Ray trac...
1	Have you tried it?
2	@joerogan @Spotify Great interview!
3	@gtera27 Doge is underestimated @teslacn Congr...
4	Now on to the next for even more!!
...	...
919	@kenyanwalstreet Not actually a payout, just a...
920	It may never pay out, as the stock can t b ht...
921	Details Aug 28.
922	AI symbiosis while u wait @vistacruiser7 @flcn...
923	@TeslaGong @PPathole Samwise Gamgee @PPathole ...

924 rows × 1 columns

In [96]:
```python
# Emotion Lexicon - Affin
afin=pd.read_csv('Afinn.csv',sep=',',encoding='Latin-1')
afin
```

Out[96]:

	word	value
0	abandon	-2
1	abandoned	-2
2	abandons	-2
3	abducted	-2
4	abduction	-2
...	...	...
2472	yucky	-2
2473	yummy	3
2474	zealot	-2
2475	zealots	-2
2476	zealous	2

2477 rows × 2 columns

In [97]:
```python
affinity_scores=afin.set_index('word')['value'].to_dict()
affinity_scores
```

```
'absolved': 2,
'absolves': 2,
'absolving': 2,
'absorbed': 1,
'abuse': -3,
'abused': -3,
'abuses': -3,
'abusive': -3,
'accept': 1,
'accepted': 1,
'accepting': 1,
'accepts': 1,
'accident': -2,
'accidental': -2,
'accidentally': -2,
'accidents': -2,
'accomplish': 2,
'accomplished': 2,
'accomplishes': 2,
```

In [98]:
```python
# Custom function: score each word in a sentence in lemmatised form, but calculate the score for the whole (
nlp=spacy.load('en_core_web_sm')
sentiment_lexicon=affinity_scores

def calculate_sentiment(text:str=None):
    sent_score=0
    if text:
        sentence=nlp(text)
        for word in sentence:
            sent_score+=sentiment_lexicon.get(word.lemma_,0)
    return sent_score
```

In [99]:
```python
# manual testing
calculate_sentiment(text='great')
```

Out[99]: 3

In [100]:
```python
# Calculating sentiment value for each sentence
sent_df['sentiment_value']=sent_df['sentence'].apply(calculate_sentiment)
sent_df['sentiment_value']
```

Out[100]:
```
0       0
1       0
2       3
3       3
4       0
       ..
919     0
920    -4
921     0
922    -2
923     0
Name: sentiment_value, Length: 924, dtype: int64
```

In [101]:
```python
# how many words are there in a sentence?
sent_df['word_count']=sent_df['sentence'].str.split().apply(len)
sent_df['word_count']
```

Out[101]:
```
0      13
1       4
2       4
3      13
4       8
       ..
919    11
920    31
921     3
922    47
923    15
Name: word_count, Length: 924, dtype: int64
```

In [102]:
```python
sent_df.sort_values(by='sentiment_value')
```

Out[102]:

	sentence	sentiment_value	word_count
647	Very ba https://t.co/tJsh1Exz1Q @justpaulinel...	-8	60
64	Also, the road to hell is mostly paved with ba...	-7	11
837	Cool Model 3 review by @iamjamiefoxx https://t...	-7	61
611	Then static fire, checkouts, static fire, fly ...	-4	12
920	It may never pay out, as the stock can t b ht...	-4	31
...	...	...	...
81	@teslaownersSV This is a good one @MrBeastYT l...	13	38
585	The open areas https://t.co/rabjKrtQlw @Sav...	14	138
719	We just haven t observed the https://t.co/mez...	15	72
36	@ajtourville @Erdayastronaut @SpaceX Yes, but ...	16	231
105	@Erdayastronaut @SpaceX Was also thinking that...	16	94

924 rows × 3 columns

In [103]:
```python
# Sentiment score of the whole review
sent_df['sentiment_value'].describe()
```

Out[103]:
```
count    924.000000
mean       1.345238
std        2.684749
min       -8.000000
25%        0.000000
50%        0.000000
75%        3.000000
max       16.000000
Name: sentiment_value, dtype: float64
```

In [104]:
```python
# negative sentiment score of the whole review
sent_df[sent_df['sentiment_value']<=0]
```

Out[104]:

	sentence	sentiment_value	word_count
0	@kunalb11 I m an alien @ID_AA_Carmack Ray trac...	0	13
1	Have you tried it?	0	4
4	Now on to the next for even more!!	0	8
5	Happy New Year of the Ox!	0	6
6	https://t.co/9WFKMYu2oj Frodo was the underdog...	-2	14
...	...	...	...
919	@kenyanwalstreet Not actually a payout, just a...	0	11
920	It may never pay out, as the stock can t b ht...	-4	31
921	Details Aug 28.	0	3
922	AI symbiosis while u wait @vistacruiser7 @flcn...	-2	47
923	@TeslaGong @PPathole Samwise Gamgee @PPathole ...	0	15

496 rows × 3 columns

In [105]:

```python
# positive sentiment score of the whole review
sent_df[sent_df['sentiment_value']>0]
```

Out[105]:

	sentence	sentiment_value	word_count
2	@joerogan @Spotify Great interview!	3	4
3	@gtera27 Doge is underestimated @teslacn Congr...	3	13
7	https://t.co/zGxJFDzzrM @OwenSparks_ @flcnhvy ...	2	10
9	The most entertaining outcome is the most like...	3	109
17	Back to work I go @CapybaraSurfer @MattWallace...	4	38
...	...	...	...
911	He was one of the very best.	3	7
913	@Ali_Afshari In general, we need to improve ho...	4	87
915	@burakaydik True Wow, IHOP & GitHub are cl...	3	15
917	This is both great & terrifying.	3	6
918	Everything we ve ever sensed or thought has be...	3	17

428 rows × 3 columns

In [106]:
```python
# Adding index cloumn
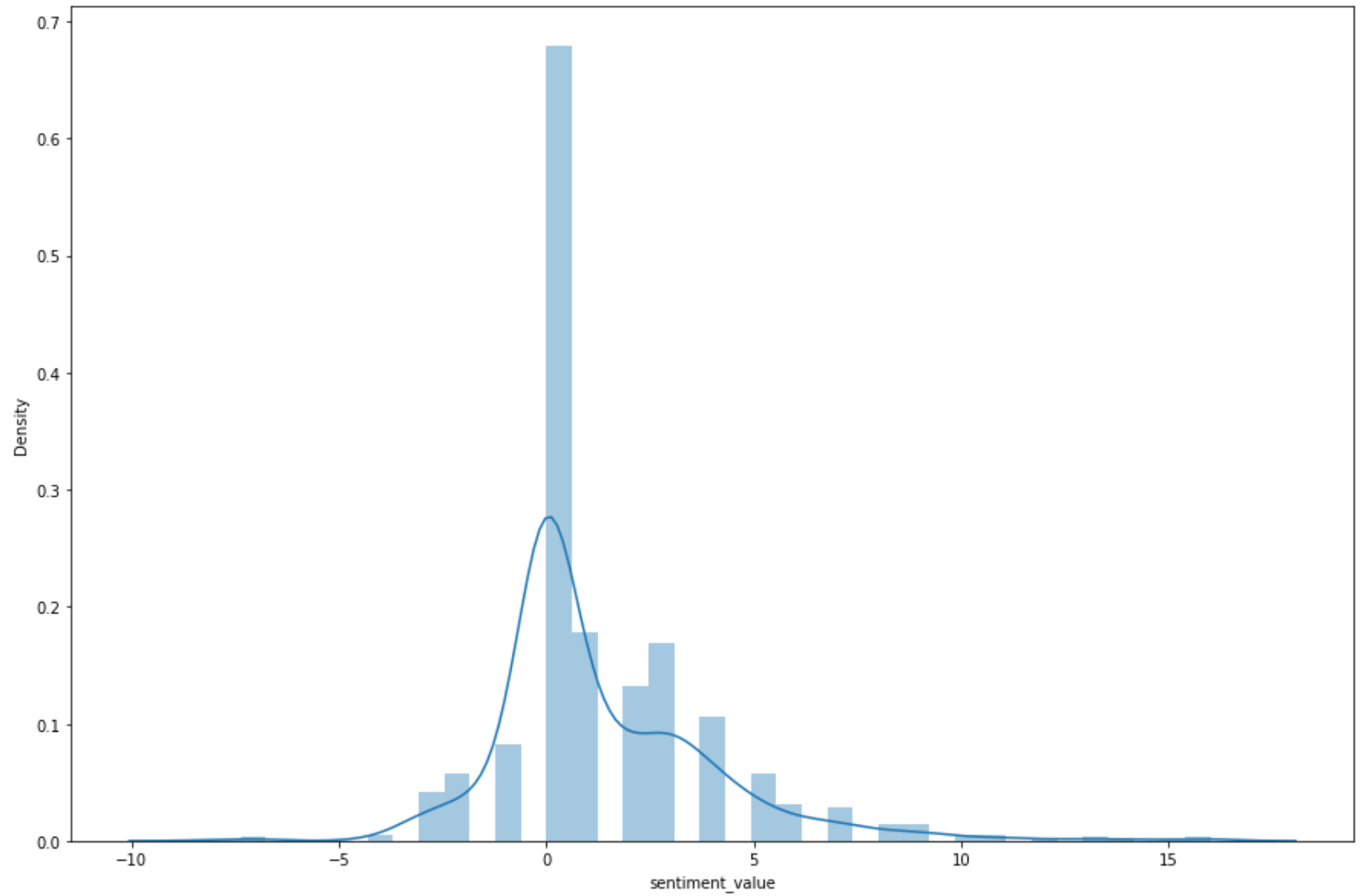sent_df['index']=range(0,len(sent_df))
sent_df
```

Out[106]:

	sentence	sentiment_value	word_count	index
0	@kunalb11 I m an alien @ID_AA_Carmack Ray trac...	0	13	0
1	Have you tried it?	0	4	1
2	@joerogan @Spotify Great interview!	3	4	2
3	@gtera27 Doge is underestimated @teslacn Congr...	3	13	3
4	Now on to the next for even more!!	0	8	4
...	...	...	...	...
919	@kenyanwalstreet Not actually a payout, just a...	0	11	919
920	It may never pay out, as the stock can t b ht...	-4	31	920
921	Details Aug 28.	0	3	921
922	AI symbiosis while u wait @vistacruiser7 @flcn...	-2	47	922
923	@TeslaGong @PPathole Samwise Gamgee @PPathole ...	0	15	923

924 rows × 4 columns

In [107]:
```python
1  # Plotting the sentiment value for whole review
2  import seaborn as sns
3  plt.figure(figsize=(15,10))
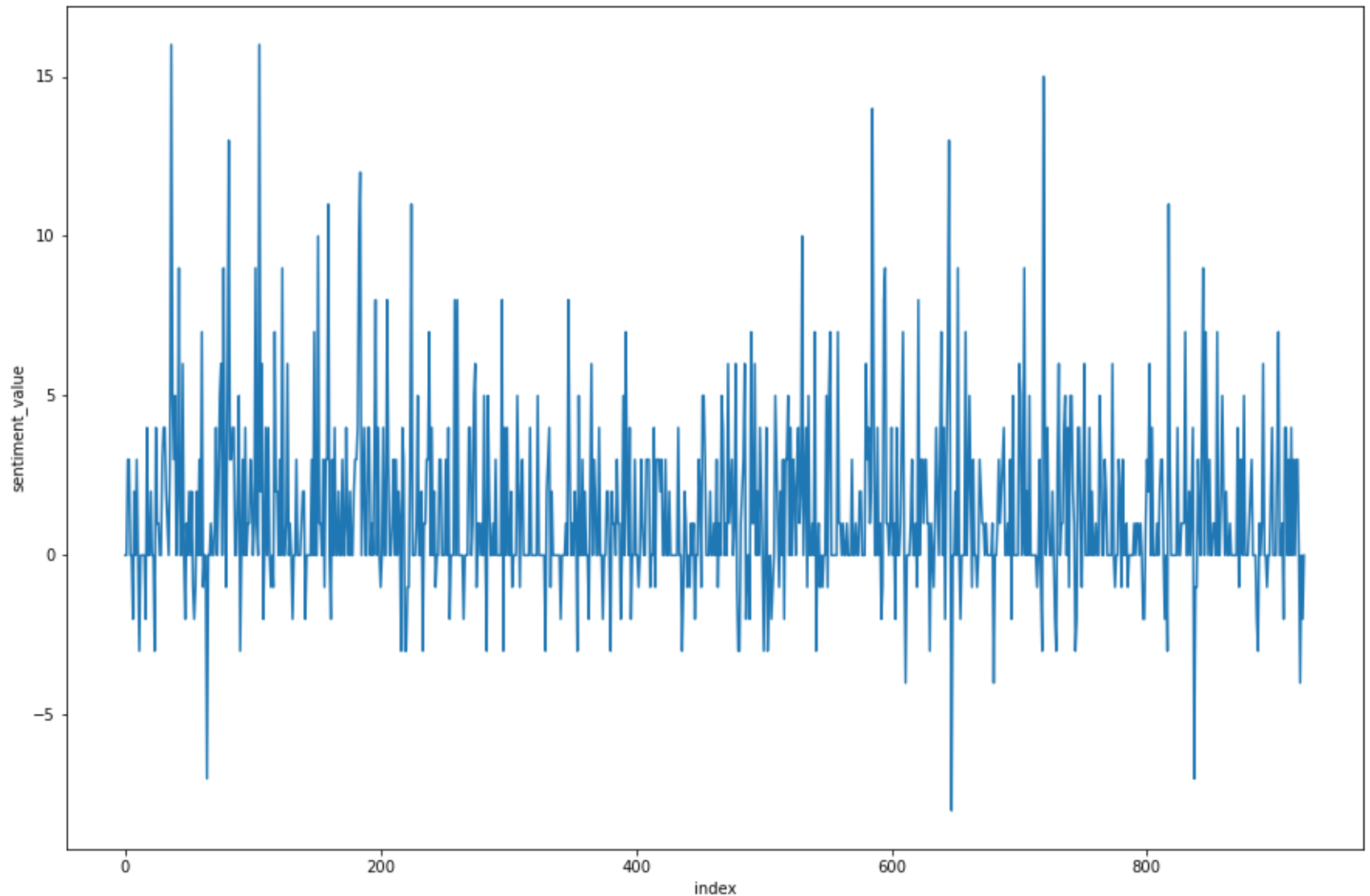4  sns.distplot(sent_df['sentiment_value'])
```

C:\Users\F SAFRIN KAUSER\anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot`
is a deprecated function and will be removed in a future version. Please adapt your code to use either `displo
t` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)

Out[107]: <AxesSubplot:xlabel='sentiment_value', ylabel='Density'>

In [108]:
```python
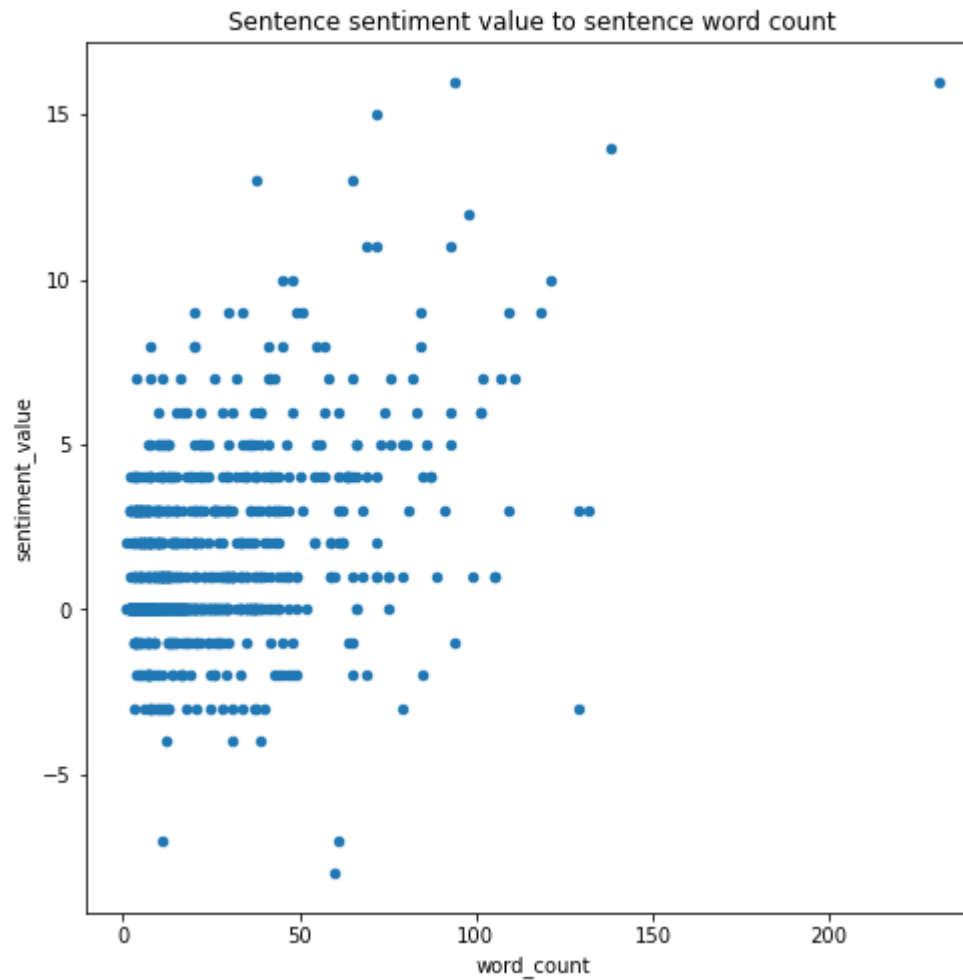# Plotting the line plot for sentiment value of whole review
plt.figure(figsize=(15,10))
sns.lineplot(y='sentiment_value',x='index',data=sent_df)
```

Out[108]: <AxesSubplot:xlabel='index', ylabel='sentiment_value'>

In [109]:

```python
# Correlation analysis
sent_df.plot.scatter(x='word_count',y='sentiment_value',figsize=(8,8),title='Sentence sentiment value to sente
```

Out[109]: <AxesSubplot:title={'center':'Sentence sentiment value to sentence word count'}, xlabel='word_count', ylabel='sentiment_value'>

In [ ]:

```python

```

localhost:8888/notebooks/Desktop/Ruksana/DataScience/ExcelR_Assignements/18.Text Mining/Elon_musk_Text-Mining_Solution.ipynb

49/49