In [1]:
```
import sys
!{sys.executable} -m spacy download en_core_web_sm
import re, numpy as np, pandas as pd
from pprint import pprint

from IPython.display import Image
from IPython.display import display
import seaborn as sns

from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer,T
fidfTransformer
from sklearn.decomposition import NMF
from sklearn.preprocessing import normalize;

import matplotlib.pyplot as plt

# NLTK Stop words
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk import FreqDist
stop_words = stopwords.words('english')
#stop_words.extend(['from', 'subject', 're', 'edu', 'use', 'not', 'would', 'sa
y', 'could', '_', 'be', 'know', 'good', 'go', 'get', 'do', 'done', 'try', 'man
y', 'some', 'nice', 'thank', 'think', 'see', 'rather', 'easy', 'easily', 'lo
t', 'lack', 'make', 'want', 'seem', 'run', 'need', 'even', 'right', 'line', 'e
ven', 'also', 'may', 'take', 'come'])

#suppress all warnings with this
import warnings
warnings.filterwarnings("ignore")
```

```
2023-12-07 16:57:59.427220: E tensorflow/compiler/xla/stream_executor/cuda/cu
da_dnn.cc:9342] Unable to register cuDNN factory: Attempting to register fact
ory for plugin cuDNN when one has already been registered
2023-12-07 16:57:59.427293: E tensorflow/compiler/xla/stream_executor/cuda/cu
da_fft.cc:609] Unable to register cuFFT factory: Attempting to register facto
ry for plugin cuFFT when one has already been registered
2023-12-07 16:57:59.427328: E tensorflow/compiler/xla/stream_executor/cuda/cu
da_blas.cc:1518] Unable to register cuBLAS factory: Attempting to register fa
ctory for plugin cuBLAS when one has already been registered
2023-12-07 16:57:59.436498: I tensorflow/core/platform/cpu_feature_guard.cc:1
82] This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild
TensorFlow with the appropriate compiler flags.
2023-12-07 16:58:01.328123: W tensorflow/compiler/tf2tensorrt/utils/py_utils.
cc:38] TF-TRT Warning: Could not find TensorRT
Collecting en-core-web-sm==3.6.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_
core_web_sm-3.6.0/en_core_web_sm-3.6.0-py3-none-any.whl (12.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 53.1 MB/s eta 0:0
0:00
Requirement already satisfied: spacy<3.7.0,>=3.6.0 in /usr/local/lib/python3.
10/dist-packages (from en-core-web-sm==3.6.0) (3.6.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/
python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0)
(3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/
python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0)
(1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/py
thon3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.
0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python
3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.0.9)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (8.1.12)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python
3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.1.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/pyth
on3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.1
0)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in /usr/local/lib/python
3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.9.0)
Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.10/dis
t-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.10.3)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/pyt
hon3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (6.4.
0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (4.66.1)
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.10/dis
t-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.23.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/pyth
```

```
on3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.31.
0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/l
ocal/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==
3.6.0) (1.10.13)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packa
ges (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.1.2)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-p
ackages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (67.7.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/d
ist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (23.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/pyth
on3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.3.
0)
Requirement already satisfied: typing-extensions>=4.2.0 in /usr/local/lib/pyt
hon3.10/dist-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.7.
0,>=3.6.0->en-core-web-sm==3.6.0) (4.5.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/pyt
hon3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-
core-web-sm==3.6.0) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist
-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm=
=3.6.0) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.1
0/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-w
eb-sm==3.6.0) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.1
0/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-w
eb-sm==3.6.0) (2023.11.17)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.1
0/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core-web-s
m==3.6.0) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/pyt
hon3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core
-web-sm==3.6.0) (0.1.4)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.
10/dist-packages (from typer<0.10.0,>=0.3.0->spacy<3.7.0,>=3.6.0->en-core-web
-sm==3.6.0) (8.1.7)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/d
ist-packages (from jinja2->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.1.
3)
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')

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

In [2]:
```
import re, nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
from sklearn.feature_extraction.text import CountVectorizer
from nltk.corpus import stopwords
from wordcloud import WordCloud
nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer
from tqdm import tqdm
from gensim.models import Word2Vec
from sklearn.manifold import TSNE
import gensim
from gensim import corpora
from gensim.models.coherencemodel import CoherenceModel
import matplotlib.colors as mcolors
from collections import Counter
from matplotlib.ticker import FuncFormatter
from bokeh.plotting import figure, output_file, show
from bokeh.models import Label
from bokeh.io import output_notebook
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

Topic Modeling

In [3]:
```
import pandas as pd
df = pd.read_csv('Suicide_Detection.csv',on_bad_lines='skip',engine="python")
```

In [4]:
```
df.columns
```

Out[4]:
```
Index(['Unnamed: 0', 'text', 'class'], dtype='object')
```

In [5]:
```
import warnings
warnings.filterwarnings("ignore")
```

In [6]:
```
df['text'].head(3)
```

Out[6]:
```
0    Ex Wife Threatening SuicideRecently I left my ...
1    Am I weird I don't get affected by compliments...
2    Finally 2020 is almost over... So I can never ...
Name: text, dtype: object
```

In [7]:
```
df['text'].str.split()
# Lower all srings
df['Text_clean'] = df['text'].str.lower()
```

In [8]:
```
#Clean the content by removing all the  punctuation,
df['Text_clean'] = df['Text_clean'].str.replace('[^\w\s]','')
```

In [9]:
```
import re
import inflect
p = inflect.engine()
```

In [10]:
```
def f(row):
    return num2words(row['Text_clean'])
```

In [11]:
```
#Clean the content by removing all the  numbers
df['Text_nonumber'] = df['Text_clean'].str.replace('\d+', '')
```

In [12]:
```
#Remove white space
df['Text_clean'] = df['Text_clean'].str.strip()
```

In [13]:
```
df['words'] = df.Text_clean.str.strip().str.split('[\W_]+')
```

In [85]:
```
#word count
words_list = df['Text_clean'].tolist()
raw_text = ''.join(words_list)
```

In [15]:
```
all_words = raw_text.split()
```

In [16]:
```
word_dict = {}

## For each word in the text
for word in all_words:
    # if the word wasn't already in the dictionary
    if word not in word_dict.keys():
        # add it
        word_dict[word] = 1
    # otherwise
    else:
        # add 1 to the existing count
        word_dict[word] = word_dict[word] + 1
```

In [17]:
```
word_dict
```

Out[17]:

```
{'ex': 1653,
 'wife': 1184,
 'threatening': 199,
 'suiciderecently': 1,
 'i': 508134,
 'left': 5897,
 'my': 182398,
 'for': 82872,
 'good': 13652,
 'because': 33058,
 'she': 33022,
 'has': 16801,
 'cheated': 534,
 'on': 49808,
 'me': 122587,
 'twice': 799,
 'and': 270447,
 'lied': 514,
 'to': 312182,
 'so': 67218,
 'much': 17902,
 'that': 102780,
 'have': 79223,
 'decided': 2611,
 'refuse': 338,
 'go': 19898,
 'back': 12962,
 'her': 30205,
 'as': 30400,
 'of': 130334,
 'a': 187040,
 'few': 7968,
 'days': 6213,
 'ago': 6340,
 'began': 763,
 'suicide': 10389,
 'tirelessly': 8,
 'spent': 1722,
 'these': 6145,
 'paat': 1,
 'talking': 4575,
 'out': 33729,
 'it': 114597,
 'keeps': 1244,
 'hesitating': 15,
 'wants': 2956,
 'believe': 3295,
 'ill': 10665,
 'come': 5210,
 'know': 39331,
 'lot': 7234,
 'people': 27770,
 'will': 23878,
 'threaten': 110,
 'this': 70107,
 'in': 95573,
 'order': 760,
```

```
'get': 32777,
'their': 8266,
'way': 13211,
'but': 86609,
'what': 32893,
'happens': 1541,
'if': 39345,
'really': 25169,
'does': 4501,
'do': 46953,
'how': 25763,
'am': 30435,
'supposed': 1988,
'handle': 1584,
'death': 3989,
'hands': 876,
'still': 12882,
'love': 11537,
'cannot': 2041,
'deal': 2730,
'with': 61127,
'getting': 8370,
'again': 9535,
'constantly': 2637,
'feeling': 7656,
'insecure': 260,
'im': 98007,
'worried': 1120,
'today': 5666,
'may': 2904,
'be': 57338,
'the': 206392,
'day': 17274,
'hope': 5373,
'doesnt': 8483,
'happenam': 1,
'weird': 2064,
'dont': 64149,
'affected': 247,
'by': 13876,
'compliments': 97,
'its': 35029,
'coming': 2199,
'from': 23074,
'someone': 13217,
'irl': 345,
'feel': 36791,
'when': 27002,
'internet': 1051,
'strangers': 380,
'itfinally': 3,
'2020': 396,
'is': 82764,
'almost': 4584,
'over': 12258,
'can': 25503,
'never': 20143,
```

```
'hear': 2104,
'been': 28889,
'bad': 9294,
'year': 10625,
'ever': 10532,
'swear': 372,
'fucking': 14059,
'god': 2675,
'annoyingi': 12,
'need': 12063,
'helpjust': 13,
'help': 15696,
'crying': 2368,
'hardim': 12,
'losthello': 1,
'name': 1719,
'adam': 31,
'16': 1305,
'ive': 32227,
'struggling': 1268,
'years': 15505,
'afraid': 2581,
'through': 9169,
'past': 5115,
'thoughts': 6152,
'fear': 1848,
'anxiety': 3958,
'close': 3843,
'limit': 185,
'quiet': 491,
'long': 7923,
'too': 13837,
'scared': 4586,
'family': 12070,
'about': 40267,
'feelings': 2529,
'3': 4404,
'losing': 1421,
'aunt': 310,
'triggered': 185,
'all': 41846,
'everyday': 2406,
'hopeless': 848,
'lost': 5021,
'guilty': 892,
'remorseful': 9,
'things': 14625,
'done': 6547,
'lifebut': 4,
'like': 54448,
'little': 5574,
'experienced': 482,
'life': 35229,
'only': 19807,
'time': 25491,
'revealed': 72,
'broke': 2284,
```

```
'down': 7999,
'where': 9446,
'they': 30353,
'saw': 1983,
'cuts': 286,
'watching': 1235,
'them': 20067,
'something': 12215,
'portrayed': 9,
'an': 18353,
'average': 540,
'made': 8651,
'absolutely': 1550,
'dreadful': 38,
'later': 2612,
'found': 4004,
'was': 64682,
'attempt': 1688,
'survivor': 50,
'odoverdose': 1,
'pills': 2014,
'hanging': 1055,
'happened': 3450,
'blackout': 43,
'went': 5877,
'noose': 279,
'during': 2009,
'first': 7718,
'therapy': 2331,
'diagnosed': 1055,
'severe': 888,
'depression': 6662,
'social': 2908,
'eating': 1281,
'disorder': 1064,
'transferred': 81,
'fucken': 20,
'group': 1548,
'some': 17257,
'reason': 6243,
'which': 8546,
'more': 18072,
'anxious': 865,
'eventually': 1766,
'before': 9166,
'last': 9931,
'session': 181,
'1': 2360,
'showed': 448,
'results': 364,
'daily': 960,
'check': 1093,
'up': 32459,
'feelingswhich': 1,
'2': 5823,
'step': 916,
'survey': 97,
```

```
'momdad': 2,
'find': 7565,
'putting': 964,
'horrible': 1945,
'afraidanxious': 1,
'mom': 6635,
'doing': 8214,
'amazing': 1184,
'described': 86,
'happiest': 181,
'shes': 4390,
'seen': 1863,
'helped': 1473,
'him': 14408,
'put': 5144,
'sertaline': 1,
'anti': 241,
'or': 38781,
'sorry': 4793,
'forgot': 596,
'finished': 560,
'prescription': 177,
'nor': 685,
'right': 10610,
'type': 1190,
'depressant': 19,
'thought': 8572,
'wanted': 7141,
'drugs': 1452,
'took': 3399,
'off': 9784,
'recommended': 108,
'pill': 248,
'schedule': 259,
'after': 11516,
'week': 4313,
'stopped': 2159,
'taking': 2982,
'worse': 5853,
'damage': 471,
'worry': 1290,
'caused': 730,
'even': 27425,
'now': 27262,
'here': 12692,
'everything': 12010,
'going': 20229,
'relapsed': 109,
'cutting': 1019,
'developed': 361,
'insomnia': 125,
'worthless': 1579,
'questioning': 113,
'why': 14312,
'whats': 2984,
'motivation': 1192,
'move': 2572,
```

```
'bed': 2921,
'keep': 8140,
'ask': 3449,
'myself': 33090,
'nearly': 841,
'every': 11659,
'night': 4957,
'having': 6723,
'break': 2026,
'everytime': 503,
'please': 5408,
'anyone': 10759,
'might': 4637,
'drastic': 66,
'shaped': 47,
'idk': 2552,
'anymorehonetly': 1,
'idki': 36,
'just': 76478,
'there': 17056,
'nothing': 11635,
'nowhere': 737,
'either': 3229,
'unbearably': 15,
'sad': 4320,
'ignoring': 284,
'friends': 18231,
'opitunity': 1,
'loosing': 93,
'girlfriend': 3355,
'hurt': 4768,
'everyone': 9407,
'talk': 11067,
'cause': 2927,
'anything': 13456,
'behind': 1446,
'education': 579,
'alone': 6596,
'not': 49134,
'enjoyed': 332,
'no': 32565,
'hopes': 413,
'dreams': 1005,
'care': 8807,
'complicated': 179,
'words': 1507,
'describe': 398,
'would': 25552,
'end': 11292,
'strong': 1418,
'brave': 205,
'enough': 7039,
'knowing': 1724,
'weak': 894,
'makes': 5440,
'sadder': 49,
'thing': 9668,
```

```
'push': 794,
'away': 7708,
'emotion': 309,
'empty': 1311,
'used': 4184,
'being': 16683,
'normal': 2635,
'understand': 3645,
'mentioned': 417,
'got': 14471,
'die': 10949,
'havnt': 31,
'brought': 725,
'realised': 304,
'cant': 30497,
'comprehend': 111,
'meaning': 644,
'rambling': 216,
'probably': 5365,
'regret': 870,
'posting': 1587,
'think': 21595,
'place': 4341,
'gun': 1397,
'head': 4073,
'encoage': 1,
'who': 18359,
'see': 12776,
'instead': 2010,
'suvive': 1,
'plus': 570,
'meaningless': 483,
'future': 2995,
'bleak': 122,
'while': 7057,
'could': 13595,
'cure': 201,
'cancer': 718,
'useful': 230,
'wasting': 498,
'your': 12437,
'timetrigger': 1,
'warning': 201,
'excuse': 490,
'self': 3747,
'inflicted': 20,
'burnsi': 2,
'crisis': 422,
'line': 674,
'panic': 1013,
'attack': 693,
'healthy': 632,
'did': 9292,
'stupid': 3512,
'impulse': 110,
'burned': 174,
'father': 2295,
```

```
'daughter': 650,
'knows': 1961,
'history': 833,
'we': 20207,
'were': 11134,
'together': 2917,
'12': 1164,
'hes': 3824,
'at': 37610,
'worst': 2201,
'had': 25811,
'always': 11940,
'cut': 2330,
'ankles': 24,
'wrists': 457,
'thinking': 5884,
'easier': 864,
'than': 10860,
'one': 26525,
'work': 9999,
'car': 2544,
'hadnt': 437,
'harmed': 154,
'without': 5744,
'usual': 443,
'moment': 2099,
'should': 8536,
'say': 10302,
'touched': 187,
'under': 1364,
'hood': 44,
'hot': 667,
'curved': 17,
'pattern': 90,
'forearm': 23,
'then': 15774,
'side': 1557,
'wrist': 283,
'are': 26621,
'inch': 124,
'kind': 3458,
'wide': 93,
'deep': 1286,
'working': 2682,
'explain': 1189,
'burns': 75,
'maybe': 6533,
'wire': 28,
'smooshed': 1,
'engine': 43,
'fix': 1295,
'want': 44437,
'harm': 1010,
'able': 4984,
'thisit': 13,
'ends': 649,
'tonighti': 103,
```

```
'anymore': 11919,
'quiteveryone': 1,
'edgy': 103,
'making': 4021,
'conscious': 201,
'stand': 1286,
'draw': 312,
'yes': 1754,
'play': 2473,
'guitar': 196,
'honestly': 2985,
'stuck': 1685,
'taste': 244,
'music': 1631,
'rock': 459,
'alt': 105,
'metal': 174,
'2000s': 29,
'90s': 55,
'make': 14388,
'unique': 121,
'style': 170,
'seeing': 1885,
'classmates': 282,
'into': 10335,
'rap': 112,
'edm': 5,
'hard': 6753,
'fit': 716,
'others': 3387,
'copying': 18,
'id': 5705,
'another': 5408,
'quirky': 45,
'kid': 1959,
'whos': 799,
'cringey': 28,
'phase': 201,
'many': 6320,
'look': 4963,
'grunge': 6,
'kinda': 2161,
'agree': 362,
'continue': 1499,
'wore': 73,
'crosses': 22,
'wallet': 60,
'chains': 33,
'tiktoks': 21,
'feels': 4190,
'categories': 14,
'confuse': 27,
'clout': 22,
'chaser': 2,
'tiktok': 202,
'e': 433,
'boy': 994,
```

```
'goddamn': 373,
'hate': 9981,
'lifemy': 67,
'20': 1640,
'oldhello': 1,
'old': 5716,
'balding': 40,
'male': 1004,
'hairline': 28,
'trash': 481,
'matters': 577,
'huge': 1110,
'bipolar': 646,
'crippling': 217,
'cherry': 38,
'top': 1374,
'wear': 553,
'hat': 70,
'247': 310,
'room': 2989,
'stop': 6519,
'pop': 801,
'xanax': 248,
'try': 8037,
'numb': 784,
'pain': 7147,
'works': 993,
'bit': 3076,
'comes': 1810,
'crashing': 181,
'once': 4109,
'communicate': 166,
'relationship': 3837,
'popular': 378,
'kids': 2241,
'dad': 4601,
'passed': 950,
'dark': 1178,
'hole': 581,
'arrested': 148,
'numerous': 154,
'times': 5473,
'rehab': 111,
'mental': 3940,
'hospitals': 122,
'you': 54534,
'havent': 4639,
'killed': 1550,
'yet': 3528,
'brothers': 525,
'didnt': 12115,
'dead': 3288,
'point': 8310,
'support': 2137,
'isnt': 4813,
'alive': 3104,
'guy': 3480,
```

```
'himself': 1124,
'bald': 44,
'looks': 1114,
'child': 1631,
'molestor': 1,
'choosei': 2,
'rest': 1908,
'sleeping': 1127,
'painkillersi': 2,
'wait': 1643,
'struggled': 477,
'6': 2060,
'finally': 4051,
'ending': 1725,
'itcan': 15,
'imagine': 1192,
'neitherwrinkles': 1,
'weight': 1196,
'gain': 283,
'hair': 946,
'loss': 468,
'messed': 477,
'teeth': 375,
'bones': 121,
'health': 2537,
'issues': 2144,
'menopause': 3,
'hormones': 127,
'hating': 368,
'new': 5189,
'generations': 49,
'amp': 1085,
'world': 6571,
'progress': 395,
'useless': 1396,
'angry': 1746,
'piece': 1341,
'shit': 10048,
'take': 9724,
'itself': 533,
'totally': 612,
'depended': 15,
'secretly': 135,
'already': 4674,
'yourself': 1968,
'happy': 7860,
'avoid': 587,
'thisdo': 10,
'hit': 1909,
'train': 614,
'painfulguns': 1,
'country': 1525,
'trains': 45,
'suffer': 1157,
'though': 5493,
'painless': 396,
'method': 556,
```

```
 'suicidedeath': 2,
 'continuedi': 1,
 'posted': 1073,
 'interesting': 577,
 'asked': 3035,
 'information': 435,
 'bunch': 804,
 'same': 6059,
 'spit': 101,
 'personal': 770,
 'obviously': 1062,
 'least': 3658,
 'trolls': 31,
 'laughs': 91,
 'desire': 572,
 'selfterminate': 2,
 'grows': 67,
 'stronger': 373,
 'bitterness': 28,
 'main': 674,
 'goal': 425,
 'throughout': 589,
 'process': 546,
 'minimize': 28,
 'subsequent': 10,
 'fallout': 55,
 'certainly': 342,
 'nice': 2676,
 'patrons': 3,
 'forum': 98,
 'respectful': 64,
 'privacy': 132,
 'ridiculous': 281,
 'expectation': 36,
 'considering': 855,
 'sourcebeen': 1,
 'suicidaleditfuck': 1,
 'verizon': 6,
 'smart': 773,
 'app': 315,
 'watch': 1621,
 'porn': 503,
 'privately': 37,
 'wtf': 414,
 'featureim': 1,
 'seems': 3703,
 'young': 1464,
 'transgender': 209,
 'sure': 6149,
 'tell': 8643,
 'lying': 771,
 'actually': 6237,
 'trans': 671,
 'overwhelmed': 315,
 'emotions': 1096,
 'wish': 6210,
 'religious': 400,
```

```
 'accepting': 182,
 'alleviate': 31,
 'yesterday': 1333,
 'barely': 1873,
 'drew': 87,
 'blood': 773,
 'correctly': 128,
 'pursue': 187,
 'theres': 5614,
 'money': 4550,
 'field': 327,
 'unless': 797,
 'become': 2484,
 'famous': 112,
 'thats': 8897,
 'happening': 1004,
 'currently': 1461,
 'seriously': 1641,
 'debating': 108,
 'longer': 2566,
 'born': 1438,
 'girl': 5906,
 'crywell': 1,
 'screwed': 272,
 'locked': 387,
 'school': 13341,
 'toilet': 221,
 'edit': 757,
 'lived': 1075,
 'storyim': 7,
 'fucked': 2681,
 'assignment': 205,
 'due': 2288,
 'tomorrow': 1857,
 'started': 6875,
 'yetyeaputting': 1,
 'knife': 686,
 'give': 6641,
 'any': 14571,
 'hesitation': 37,
 'free': 1940,
 'fun': 2305,
 'depressing': 316,
 'sister': 2264,
 'goes': 1917,
 'friend': 8954,
 'hurts': 1997,
 'realize': 1435,
 'hahai': 12,
 'goodbye': 892,
 'everyonei': 57,
 '36': 116,
 '37': 66,
 'disability': 285,
 'ptsd': 487,
 'rheumatoid': 9,
 'arthritis': 34,
```

```
'400': 87,
'lbs': 110,
'sick': 3146,
'living': 5991,
'tired': 6405,
'single': 2505,
'rejected': 481,
'monster': 329,
'connect': 273,
'companionship': 36,
'loneliness': 583,
'taken': 1553,
'swallowed': 96,
'inside': 1927,
'consumed': 127,
'darkness': 326,
'everywhere': 458,
'towards': 1044,
'reminds': 191,
'reads': 220,
'deadme': 1,
'toxic': 629,
'house': 3886,
'best': 6701,
'cope': 763,
'etc': 2037,
'rona': 21,
'hahahaha': 14,
'stay': 3188,
'home': 5872,
'forcefully': 21,
'brrrrrrrrrtrapped': 1,
'voiddear': 1,
'whoever': 312,
'cares': 1674,
'read': 2941,
'doubt': 724,
'fall': 1474,
'criteria': 34,
'trapped': 653,
'void': 288,
'monotonous': 40,
'tasks': 151,
'forward': 1373,
'aspirations': 102,
'whatsoever': 206,
'saving': 259,
'grace': 54,
'felt': 6113,
'accepted': 545,
'maren': 7,
'idiotic': 63,
'sounds': 1054,
'assured': 44,
'looked': 1168,
'kept': 1759,
'briefly': 60,
```

```
'sleep': 4656,
'worth': 2704,
'okay': 2400,
'most': 7341,
'loved': 2761,
'open': 1597,
'completely': 2803,
'ease': 156,
'told': 9145,
'else': 6889,
'cared': 890,
'shared': 241,
'intimate': 87,
'moments': 628,
'dare': 200,
'talked': 1896,
'each': 2921,
'other': 11839,
'devising': 4,
'crazy': 1261,
'scenarios': 75,
'us': 4740,
'apart': 1099,
'except': 1310,
'ty': 28,
'problems': 3434,
'facing': 191,
'related': 328,
'breakup': 311,
'starr': 9,
'whenever': 1252,
'needed': 1927,
'better': 11803,
'shining': 38,
'star': 216,
'pitchblack': 2,
'sky': 138,
'none': 1287,
'slowly': 975,
'drifted': 70,
'realise': 330,
'exactly': 975,
'regrets': 116,
'involved': 388,
'blame': 982,
'intentional': 25,
'nonetheless': 66,
'clingy': 88,
'tendencies': 158,
'said': 8603,
'suppose': 445,
'overexaggerating': 6,
'wasnt': 4385,
'interaction': 253,
'fair': 523,
'around': 8193,
'own': 5587,
```

```
'extremely': 1565,
'selfdestructive': 49,
'spend': 1637,
'four': 694,
'months': 6248,
'legitimately': 122,
'thanks': 1831,
'stress': 1321,
'looking': 3421,
'migraines': 55,
'becoming': 819,
'far': 2685,
'between': 1412,
'replay': 17,
'those': 4953,
'ridiculing': 1,
'various': 285,
'reasons': 1681,
'beyond': 681,
'control': 1510,
'hold': 1631,
'share': 1156,
'incredibly': 692,
'receptive': 7,
'lie': 1093,
'leave': 4406,
'details': 385,
'person': 9239,
'seemed': 851,
'genuinely': 904,
'concerned': 333,
'tried': 8213,
'dispel': 2,
'hateful': 105,
'attractive': 546,
'ten': 359,
'prior': 192,
'willingly': 44,
'picture': 577,
'send': 1060,
'snapchat': 213,
'pictures': 388,
'hideous': 63,
'hindsight': 57,
'miss': 1881,
'return': 513,
'happen': 2565,
'extreme': 453,
'fine': 1920,
'start': 4028,
'highmaintenance': 1,
'causes': 291,
'drift': 104,
'until': 5420,
'altogether': 49,
'sometimes': 3598,
'abruptly': 25,
```

```
'lives': 1916,
'case': 1110,
'morning': 1806,
'weeks': 3280,
'ended': 2086,
'whim': 35,
'lose': 1730,
'significant': 246,
'also': 8799,
'driving': 753,
'two': 5438,
'eat': 1949,
'laid': 317,
'staring': 366,
'ceiling': 139,
'tears': 674,
'trying': 7291,
'gone': 3618,
'wrong': 4243,
'cried': 713,
'hid': 109,
'outside': 1597,
'view': 377,
'contain': 37,
'complete': 1020,
'luck': 597,
'reconnect': 28,
'idea': 2780,
'online': 2197,
'facebook': 421,
'absolute': 508,
'lowest': 243,
'sent': 1149,
'message': 1000,
'expectations': 275,
'prepared': 224,
'ignored': 482,
'wonderful': 528,
'conversation': 966,
'reminiscing': 10,
'our': 5422,
'childhood': 863,
'remember': 3387,
'subject': 332,
'changed': 1144,
'stayed': 618,
'partly': 79,
'upset': 1146,
'couldnt': 3956,
'especially': 1391,
'decade': 305,
'joked': 34,
'reconnecting': 8,
'red': 564,
'string': 77,
'fate': 197,
'dumb': 977,
```

```
 'problem': 2435,
 'lies': 402,
 'fact': 2845,
 'replicate': 6,
 'socially': 441,
 'awkward': 709,
 'freeze': 76,
 'boring': 604,
 'subreddits': 100,
 'rr4r': 1,
 'rkikpals': 1,
 'actual': 738,
 'friendship': 521,
 'fighting': 1027,
 'badgering': 2,
 'hang': 1765,
 'game': 1698,
 'thrones': 23,
 'boyfriend': 2541,
 'selfworth': 45,
 'bottom': 524,
 'fault': 1476,
 'relationships': 1132,
 'dissolve': 13,
 'joined': 272,
 'blaming': 154,
 'dissolution': 3,
 'unlovable': 79,
 'true': 1421,
 'stems': 37,
 'growing': 704,
 ...}
```

In [18]:
```
##NLTK
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

Out[18]:
```
True
```

In [19]:
```
## word_tokenize
df['tokenized_text'] = df['Text_clean'].apply(nltk.word_tokenize)
```

In [20]:
```
import numpy as np

from nltk.tokenize import TweetTokenizer
from nltk import ngrams
```

In [21]:
```
df['class'].value_counts()
```

Out[21]:
```
non-suicide    35164
suicide        35033
Name: class, dtype: int64
```

In [22]:
```
df.loc[df['class'] == "suicide",'coding'] = 1
df.loc[df['class'] == "non-suicide",'coding'] = 0

df_train = df.sample(frac=.8,random_state = 123).copy()
df_test = df.drop(df_train.index).copy()
```

In [23]:
```
df
```

Out[23]:

	Unnamed: 0	text	class	Text_clean	Text_nonumber	
0	2	Ex Wife Threatening SuicideRecently I left my ...	suicide	ex wife threatening suiciderecently i left my ...	ex wife threatening suiciderecently i left my ...	thr suicider
1	3	Am I weird I don't get affected by compliments...	non-suicide	am i weird i dont get affected by compliments ...	am i weird i dont get affected by compliments ...	[am, aff
2	4	Finally 2020 is almost over... So I can never ...	non-suicide	finally 2020 is almost over so i can never hea...	finally is almost over so i can never hear h...	[finally, almost,
3	8	i need helpjust help me im crying so hard	suicide	i need helpjust help me im crying so hard	i need helpjust help me im crying so hard	[i, need help cryi
4	9	I'm so lostHello, my name is Adam (16) and I'v...	suicide	im so losthello my name is adam 16 and ive bee...	im so losthello my name is adam and ive been ...	[im, so, my, adam
...	...	...	...	...	...	
70192	105341	mint choc chip ice cream ? https://forms.gle/9...	non-suicide	mint choc chip ice cream httpsformsgle9ugzj16...	mint choc chip ice cream httpsformsgleugzjvqh...	[mint, cl ic httpsforn
70193	105344	Probably a stupid question. So for context, I'...	non-suicide	probably a stupid question so for context im 1...	probably a stupid question so for context im ...	[pr stupid, so, fo
70194	105345	i'm too embarrassed to tell people i know of m...	suicide	im too embarrassed to tell people i know of my...	im too embarrassed to tell people i know of my...	embarra tell,
70195	105346	Suicidal thoughts, partner doesn't support me ...	suicide	suicidal thoughts partner doesnt support me or...	suicidal thoughts partner doesnt support me or...	thoughts s
70196	105347	I just made this meal from an idea and it's re...	non-suicide	i just made this meal from an idea and its rea...	i just made this meal from an idea and its rea...	[i, ju this, m an, i

70197 rows × 8 columns

In [24]:
```
from nltk.corpus import stopwords
nltk.download('stopwords')
## remove stopwords
stop=stopwords.words('english')
df["Text_stop_removed"]=df['Text_clean'].apply(lambda x: ' '.join([word for wo
rd in x.split() if word not in (stop)]))
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

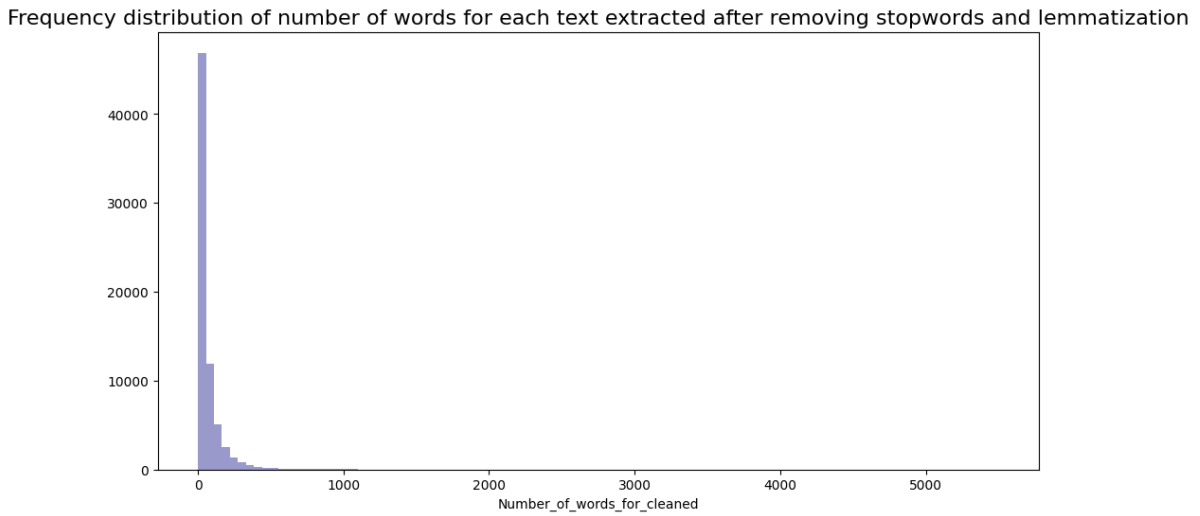
In [25]:
```
df["tokenized"]=df["Text_stop_removed"].apply(lambda x: nltk.word_tokenize(x))
```

In [26]:
```
import nltk
nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer
def word_lemmatizer(text):
    lem_text = [WordNetLemmatizer().lemmatize(i,pos='v') for i in text]
    return lem_text
df["lemmatized"]=df["tokenized"].apply(lambda x: word_lemmatizer(x))
df["lemmatize_joined"]=df["lemmatized"].apply(lambda x: ' '.join(x))
```

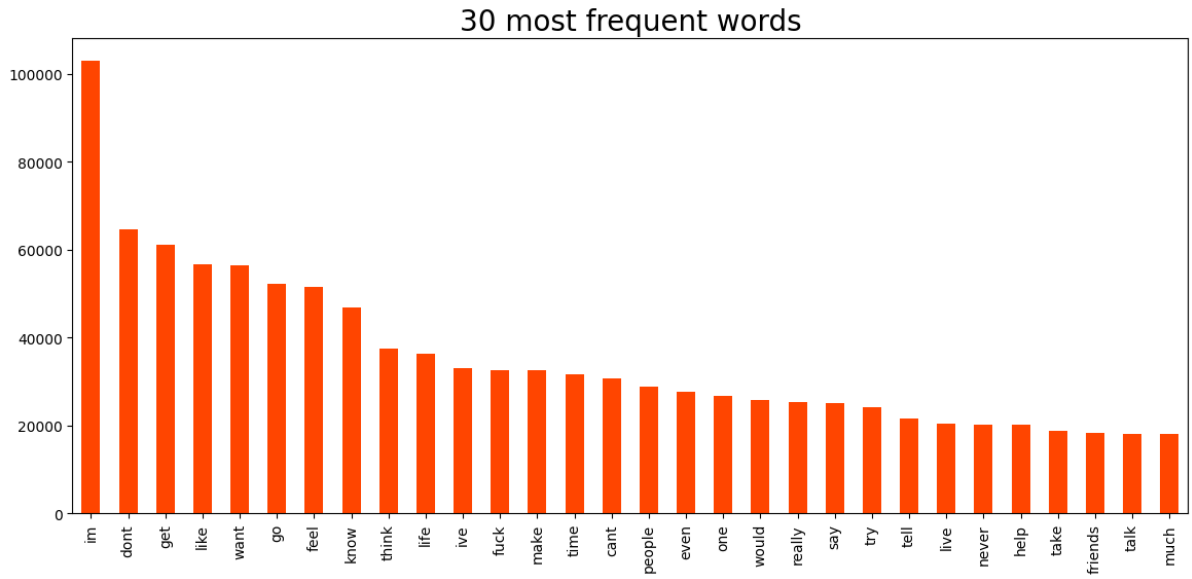
```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

In [27]:
```
df['Number_of_words_for_cleaned'] = df['lemmatize_joined'].apply(lambda x:len
(str(x).split()))
```

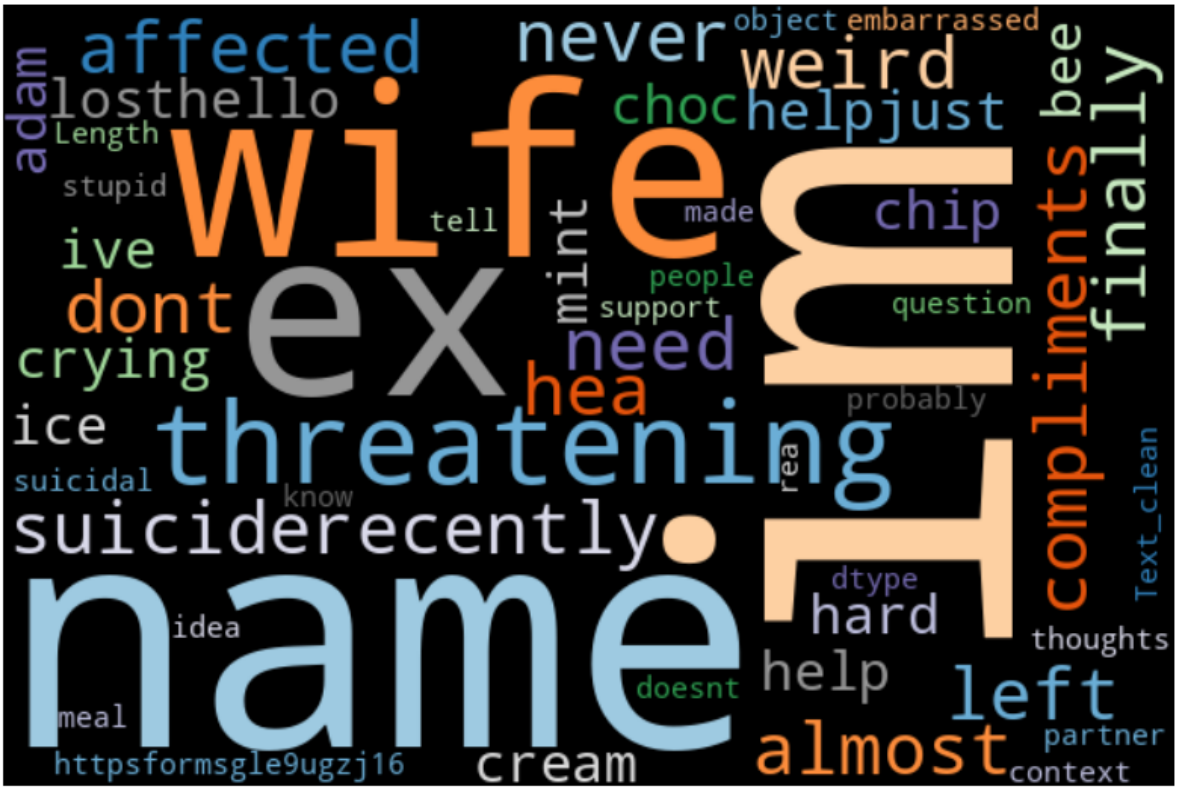
In [28]:
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12,6))
sns.distplot(df['Number_of_words_for_cleaned'],kde = False, color= "navy", bin
s = 100)
plt.title("Frequency distribution of number of words for each text extracted a
fter removing stopwords and lemmatization", size=16)
plt.show()
```



In [29]:
```
plt.figure(figsize=(14,6))
freq=pd.Series(" ".join(df["lemmatize_joined"]).split()).value_counts()[:30]
freq.plot(kind="bar", color = "orangered")
plt.title("30 most frequent words",size=20)
plt.show()
```



In [30]:
```
from wordcloud import WordCloud
cloud=WordCloud(colormap='tab20c',width=600,height=400).generate(str(df["Text_
clean"]))
fig=plt.figure(figsize=(10,15))
plt.axis("off")
plt.imshow(cloud,interpolation='bilinear')
plt.show()
```



Ngrams

In [31]:
```
import plotly.graph_objects as go
```

In [32]:
```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer,T
fidfTransformer
from sklearn.decomposition import NMF
from sklearn.preprocessing import normalize;
def get_top_n_words(corpus, n=None):
    vec = CountVectorizer().fit(corpus)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.i
tems()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return words_freq[:n]
common_words = get_top_n_words(df['text'], 20)

df1 = pd.DataFrame(common_words, columns = ['text' , 'count'])

fig = go.Figure([go.Bar(x=df1['text'], y=df1['count'])])
fig.update_layout(title=go.layout.Title(text="Top 20 words in the dataset befo
re removing stop words"))
fig.show()
```

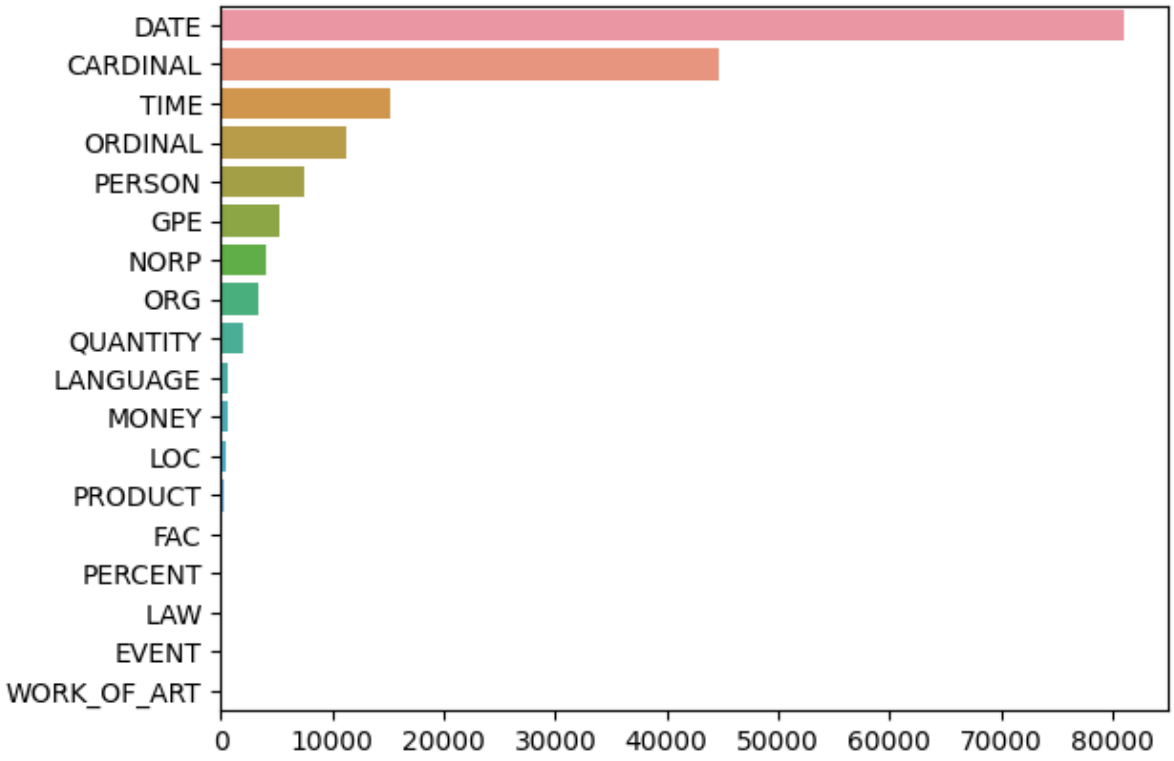
In [33]:
```
def get_top_n_bigram(corpus, n=None):
    vec = CountVectorizer(ngram_range=(2, 2), stop_words='english').fit(corpu
s)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.i
tems()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return words_freq[:n]
common_words = get_top_n_bigram(df["Text_clean"], 20)
df3 = pd.DataFrame(common_words, columns = ['bigram' , 'count'])

fig = go.Figure([go.Bar(x=df3['bigram'], y=df3['count'])])
fig.update_layout(title=go.layout.Title(text="Top 20 bigrams in the text after
removing stop words and lemmatization"))
fig.show()
```

In [34]:
```
def get_top_n_trigram(corpus, n=None):
    vec = CountVectorizer(ngram_range=(3, 3), stop_words='english').fit(corpu
s)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.i
tems()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return words_freq[:n]
common_words = get_top_n_trigram(df["text"], 20)
df4 = pd.DataFrame(common_words, columns = ['trigram' , 'count'])

fig = go.Figure([go.Bar(x=df4['trigram'], y=df4['count'])])
fig.update_layout(title=go.layout.Title(text="Top 20 trigrams in text"))
fig.show()
```

In [35]:
```
from collections import Counter
import sys
!{sys.executable} -m spacy download en_core_web_sm
# Spacy for preprocessing
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
def plot_named_entity_barchart(text):
    nlp = spacy.load("en_core_web_sm")

    def _get_ner(text):
        doc=nlp(text)
        return [X.label_ for X in doc.ents]

    ent=text.apply(lambda x : _get_ner(x))
    ent=[x for sub in ent for x in sub]
    counter=Counter(ent)
    count=counter.most_common()

    x,y=map(list,zip(*count))
    sns.barplot(x=y,y=x)

plot_named_entity_barchart(df['Text_clean'])
```

```
2023-12-07 17:03:28.156296: E tensorflow/compiler/xla/stream_executor/cuda/cu
da_dnn.cc:9342] Unable to register cuDNN factory: Attempting to register fact
ory for plugin cuDNN when one has already been registered
2023-12-07 17:03:28.156375: E tensorflow/compiler/xla/stream_executor/cuda/cu
da_fft.cc:609] Unable to register cuFFT factory: Attempting to register facto
ry for plugin cuFFT when one has already been registered
2023-12-07 17:03:28.156426: E tensorflow/compiler/xla/stream_executor/cuda/cu
da_blas.cc:1518] Unable to register cuBLAS factory: Attempting to register fa
ctory for plugin cuBLAS when one has already been registered
2023-12-07 17:03:28.169475: I tensorflow/core/platform/cpu_feature_guard.cc:1
82] This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild
TensorFlow with the appropriate compiler flags.
2023-12-07 17:03:30.040810: W tensorflow/compiler/tf2tensorrt/utils/py_utils.
cc:38] TF-TRT Warning: Could not find TensorRT
Collecting en-core-web-sm==3.6.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_
core_web_sm-3.6.0/en_core_web_sm-3.6.0-py3-none-any.whl (12.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 54.4 MB/s eta 0:0
0:00
Requirement already satisfied: spacy<3.7.0,>=3.6.0 in /usr/local/lib/python3.
10/dist-packages (from en-core-web-sm==3.6.0) (3.6.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/
python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0)
(3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/
python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0)
(1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/py
thon3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.
0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python
3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.0.9)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (8.1.12)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python
3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.1.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/pyth
on3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.1
0)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in /usr/local/lib/python
3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.9.0)
Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.10/dis
t-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.10.3)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/pyt
hon3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (6.4.
0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.
10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (4.66.1)
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.10/dis
t-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.23.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/pyth
```

```
on3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.31.
0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/l
ocal/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==
3.6.0) (1.10.13)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packa
ges (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.1.2)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-p
ackages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (67.7.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/d
ist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (23.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/pyth
on3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.3.
0)
Requirement already satisfied: typing-extensions>=4.2.0 in /usr/local/lib/pyt
hon3.10/dist-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.7.
0,>=3.6.0->en-core-web-sm==3.6.0) (4.5.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/pyt
hon3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-
core-web-sm==3.6.0) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist
-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm=
=3.6.0) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.1
0/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-w
eb-sm==3.6.0) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.1
0/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-w
eb-sm==3.6.0) (2023.11.17)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.1
0/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core-web-s
m==3.6.0) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/pyt
hon3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core
-web-sm==3.6.0) (0.1.4)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.
10/dist-packages (from typer<0.10.0,>=0.3.0->spacy<3.7.0,>=3.6.0->en-core-web
-sm==3.6.0) (8.1.7)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/d
ist-packages (from jinja2->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.1.
3)
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```



Word2Vec

In [36]:
```
all_words = df["text"].apply(lambda x: nltk.word_tokenize(x))
```

In [37]:
```
from gensim.models import Word2Vec
w2v_model = Word2Vec(all_words,
                     min_count=600,
                     window=10,
                     #size=250,
                     alpha=0.03,
                     min_alpha=0.0007,
                     workers = 4,
                     seed = 42)
```

In [38]:
```
w2v_model.wv.index_to_key
```

Out[38]:
```
['I',
 '.',
 ',',
 'to',
 'and',
 'the',
 'a',
 'my',
 'of',
 'it',
 'me',
 'that',
 '’',
 "n't",
 'in',
 'do',
 'is',
 'have',
 'for',
 'i',
 'just',
 'but',
 'was',
 "'m",
 'this',
 'with',
 'so',
 'be',
 'you',
 'like',
 'on',
 '?',
 'not',
 "'s",
 'want',
 'about',
 'all',
 'know',
 'or',
 't',
 'feel',
 'at',
 'life',
 'out',
 'can',
 'myself',
 'up',
 'if',
 'get',
 'because',
 'what',
 'her',
 'they',
 'am',
 'she',
 'been',
 'as',
```

```
'no',
'm',
'would',
')',
'are',
'people',
"'ve",
'(',
'had',
'time',
'even',
'one',
'really',
'now',
'when',
'he',
'how',
'will',
'from',
'think',
'them',
'going',
's',
'It',
'go',
'!',
'My',
'only',
'never',
'ca',
'who',
'there',
'an',
'much',
'we',
'more',
'friends',
'has',
'...',
'did',
':',
'could',
';',
'some',
'being',
'day',
'help',
'*',
'don',
'years',
"''",
'got',
'him',
'The',
'make',
'things',
'``',
```

```
'too',
'any',
'But',
'good',
'by',
'anything',
'then',
'way',
'here',
'school',
'someone',
'fucking',
'anymore',
'back',
'see',
'still',
'something',
'over',
'family',
'\u200d',
'better',
'your',
'And',
'need',
'other',
'always',
'end',
'die',
'live',
'love',
'why',
'talk',
'kill',
'everything',
'fuck',
'than',
'anyone',
'year',
'nothing',
'say',
'into',
'right',
'ever',
'suicide',
'shit',
'work',
'So',
'na',
've',
"'ll",
'were',
'off',
'hate',
'again',
'every',
'thing',
'after',
```

```
                'take',
                'does',
                'should',
                'person',
                'bad',
                'last',
                'friend',
                'through',
                'where',
                'told',
                '-',
                'very',
                'care',
                'before',
                'parents',
                'everyone',
                'made',
                'since',
                'point',
                'said',
                'thought',
                'his',
                'tell',
                'im',
                'getting',
                'doing',
                'She',
                'filler',
                'around',
                'job',
                'tried',
                'their',
                'down',
                'which',
                '&',
                'try',
                'keep',
                'few',
                'away',
                'happy',
                'long',
                'Filler',
                'feeling',
                'find',
                'dont',
                'also',
                'If',
                'first',
                '..',
                'else',
                'lot',
                'pain',
                'trying',
                'enough',
                'wanted',
                'This',
                'He',
```

```
'started',
'hard',
'most',
'alone',
'done',
'world',
'mom',
'while',
'best',
'having',
'depression',
'give',
'tired',
'ago',
'stop',
'suicidal',
'reason',
'thoughts',
'What',
'days',
'months',
'many',
'actually',
'felt',
'wish',
"'re",
'living',
'same',
'worse',
'girl',
'well',
'left',
'sure',
'home',
'2',
"'d",
'went',
'thinking',
'old',
'”',
'own',
'without',
'these',
'--',
'post',
'little',
'times',
'They',
'though',
'its',
'“',
'You',
'makes',
'let',
'We',
'hope',
'until',
```

```
'two',
'gon',
'probably',
'another',
'our',
'come',
'past',
'A',
'put',
'night',
'today',
'depressed',
'able',
'lost',
'new',
'look',
'hurt',
'pretty',
'already',
'amp',
'sleep',
'those',
'Why',
'scared',
'3',
'wan',
'dad',
'mind',
'might',
'money',
'talking',
'No',
'maybe',
'That',
'almost',
'place',
'leave',
'There',
'week',
'sad',
'us',
'wrong',
'#',
'guess',
'used',
'head',
'feels',
'sorry',
'next',
'start',
'Ni',
'college',
'cant',
'found',
'wo',
'making',
'guys',
```

```
'death',
'house',
'Just',
'anxiety',
'll',
'relationship',
'close',
'mental',
'mother',
'finally',
'\u200e\u200f\u200f\u200e',
'once',
'matter',
'gone',
'How',
'stuff',
'self',
'understand',
'least',
'high',
'please',
'seems',
'When',
'Im',
'hours',
'Now',
'guy',
'whole',
'Fuck',
'such',
'real',
'stupid',
'happened',
'call',
'saying',
'problems',
'kind',
'ask',
'dead',
'girlfriend',
'mean',
'took',
'others',
'looking',
'weeks',
'remember',
'believe',
'alive',
'5',
'All',
'part',
'yet',
'came',
'great',
'FUCK',
'bit',
'sick',
```

```
                'either',
                'stay',
                'future',
                'change',
                'different',
                'In',
                'killing',
                'asked',
                'room',
                'https',
                'literally',
                'bed',
                'month',
                'together',
                'taking',
                'wants',
                ']',
                'body',
                'soon',
                '[',
                'read',
                'didn',
                'Not',
                'social',
                're',
                'fact',
                'idea',
                'gets',
                'completely',
                'each',
                'called',
                'loved',
                'deal',
                '4',
                'sometimes',
                'class',
                "'",
                'far',
                'use',
                'worth',
                'nice',
                'cause',
                'telling',
                'Do',
                'seem',
                'fucked',
                'normal',
                'longer',
                'working',
                'brother',
                'man',
                'happen',
                'may',
                'boyfriend',
                'decided',
                'both',
                '  ',
```

```
'afraid',
'Like',
'honestly',
'knew',
'constantly',
'car',
'1',
'move',
'later',
'face',
'....',
'feelings',
'says',
'couple',
'health',
'hell',
'become',
'Please',
'Then',
'single',
'play',
'hospital',
'recently',
'problem',
'cry',
'Is',
'less',
'nobody',
'deserve',
'father',
'crying',
'wake',
'fun',
'd',
'sister',
'For',
'big',
'advice',
'okay',
'cut',
'broke',
'girls',
'phone',
'therapy',
'story',
'kids',
'Maybe',
'situation',
'coming',
'6',
'due',
'online',
'At',
'issues',
'stopped',
'support',
'doesn',
```

```
'moment',
'worst',
'x200B',
'kinda',
'whatever',
'10',
'heart',
'hear',
'text',
'weird',
'plan',
'everyday',
'died',
'ended',
'To',
'basically',
'yourself',
'lonely',
'break',
'etc',
'pills',
'age',
'Paul',
'Every',
'Even',
'kid',
'Jake',
'knows',
'supposed',
'half',
'saw',
'horrible',
'After',
'second',
'gt',
'lives',
'fine',
'failed',
'needed',
'Day',
'inside',
'gave',
'\u200f\u200f\u200e',
'goes',
'free',
'turn',
'ready',
'therapist',
'hurts',
'cum',
'talked',
'eat',
'hit',
'rest',
'miss',
'gotten',
'pay',
```

```
                'tomorrow',
                'entire',
                'shitty',
                'morning',
                'reading',
                'met',
                'brain',
                'seen',
                'barely',
                'chance',
                'anyway',
                'fear',
                'truly',
                'games',
                'comes',
                'seeing',
                'As',
                'instead',
                'u',
                'full',
                'idk',
                'often',
                'moved',
                'kept',
                'hang',
                'angry',
                'rather',
                'writing',
                'attempt',
                'reddit',
                '20',
                'sense',
                'during',
                '%',
                'write',
                'lose',
                'won',
                'human',
                'failure',
                'name',
                'cares',
                'ending',
                'spent',
                'terrible',
                'ones',
                'means',
                'experience',
                'stuck',
                'child',
                'game',
                'ex',
                'small',
                'burden',
                'reasons',
                'birthday',
                'suffering',
                'fight',
```

```
                'tonight',
                'starting',
                'dying',
                'sex',
                'wait',
                'wanting',
                'selfish',
                '15',
                'short',
                'spend',
                'knowing',
                'minutes',
                'show',
                'state',
                'enjoy',
                'handle',
                'three',
                'hold',
                'music',
                'open',
                'People',
                'worthless',
                'listen',
                'side',
                'outside',
                'attention',
                'watch',
                'quite',
                'happens',
                'country',
                'course',
                'posting',
                'Well',
                'turned',
                'asking',
                '.-',
                'video',
                'group',
                'taken',
                'possible',
                'classes',
                'killed',
                'extremely',
                'happiness',
                'run',
                'continue',
                'words',
                'control',
                'One',
                'fault',
                'Because',
                '/',
                'bring',
                'absolutely',
                'god',
                '18',
                'eventually',
```

```
                'eyes',
                'bored',
                'seriously',
                'Nothing',
                'helped',
                'hour',
                'born',
                'grade',
                'behind',
                'question',
                'fall',
                'given',
                'young',
                'food',
                'commit',
                'broken',
                'strong',
                'easy',
                'drugs',
                'realize',
                'dream',
                'true',
                'yeah',
                'giving',
                'waiting',
                'sort',
                'usually',
                'ugly',
                'gun',
                'useless',
                'playing',
                'ok',
                'ass',
                'against',
                'exist',
                '8',
                'trust',
                'losing',
                'miserable',
                'forward',
                'lt',
                'Also',
                'forever',
                'late',
                'meet',
                'between',
                'grades',
                'top',
                'realized',
                'medication',
                'under',
                'currently',
                'sit',
                'meds',
                'Everything',
                'crush',
                'mine',
```

```
'date',
'16',
'7',
'Everyone',
'worked',
'empty',
'Or',
'Life',
'piece',
'stress',
'pop',
'sub',
'cool',
'Can',
'dog',
'worry',
'ways',
'fix',
'leaving',
'heard',
'teacher',
'hand',
'crazy',
'account',
'option',
'constant',
'stand',
'God',
'\u200e',
'funny',
'answer',
'struggling',
'note',
'walk',
'mad',
'fail',
'drive',
'Its',
'themselves',
'ive',
'damn',
'eating',
'Thanks',
'deep',
'became',
'keeps',
'14',
'Thank',
'front',
'buy',
'sucks',
'wife',
'older',
'serious',
'mentally',
'motivation',
'explain',
```

```
                'weight',
                'haven',
                'especially',
                'set',
                'posts',
                'super',
                'straight',
                '$',
                'afford',
                'middle',
                'watching',
                'Anyone',
                'share',
                'Today',
                'except',
                'suffer',
                'isn',
                'act',
                'sitting',
                'amazing',
                'amount',
                'doctor',
                'didnt',
                'type',
                'awful',
                'abuse',
                'looked',
                'woman',
                'peace',
                'low',
                'dark',
                'thinks',
                'waste',
                'upset',
                'needs',
                'lol',
                'important',
                'changed',
                'pathetic',
                'emotional',
                'abusive',
                'hurting',
                'yes',
                'physically',
                'himself',
                'lazy',
                'sent',
                'joke',
                'case',
                'drink',
                'liked',
                'worried',
                'along',
                'drunk',
                'relationships',
                'yesterday',
                'Sorry',
```

```
'somehow',
'chat',
'Reddit',
'water',
'emotions',
'perfect',
'speak',
'apart',
'energy',
'bullshit',
'sleeping',
'gay',
'simply',
'lie',
'beautiful',
'herself',
'somewhere',
'lately',
'reality',
'thank',
'forget',
'several',
'imagine',
'looks',
'lived',
'physical',
'17',
'using',
'mess',
'wasn',
'huge',
'posted',
'12',
'meant',
'keeping',
'reach',
'Does',
'women',
'must',
'Idk',
'ill',
'check',
'diagnosed',
'society',
'Sometimes',
'30',
'hanging',
'towards',
'positive',
'attempted',
'Any',
'fighting',
'happening',
'escape',
'harm',
'drinking',
'pass',
```

```
                'dating',
                'works',
                'dreams',
                'jobs',
                'existence',
                'sounds',
                'Hey',
                'Some',
                'planning',
                'send',
                'contact',
                'cutting',
                'jump',
                'panic',
                'choice',
                'struggle',
                'none',
                'On',
                'complete',
                'TO',
                'purpose',
                'conversation',
                'random',
                '13',
                'harder',
                'wont',
                'tells',
                'message',
                'internet',
                'Cake',
                'blame',
                'subreddit',
                'helping',
                'quit',
                'dumb',
                'painful',
                'likely',
                'male',
                'Have',
                'loves',
                'student',
                'thats',
                'THE',
                'chest',
                'disorder',
                'anybody',
                'despite',
                'emotionally',
                'save',
                'university',
                'failing',
                'thanks',
                'sound',
                'boy',
                'fell',
                'takes',
                'exactly',
```

```
 'number',
 'whenever',
 'known',
 'passed',
 'hair',
 'early',
 'putting',
 'interested',
 'anyways',
 'difficult',
 'daily',
 'mostly',
 'slowly',
 'wonder',
 'asleep',
 'multiple',
 '100',
 'honest',
 'moving',
 'test',
 'badly',
 'song',
 'ruined',
 'anywhere',
 'bother',
 'AND',
 'younger',
 'gives',
 'hoping',
 'comment',
 'word',
 ...]
```

In [39]:
```
v1 = w2v_model.wv['die']
print(v1)
```

```
[-0.01656012 -1.7131358  -0.09624957 -3.054288    1.8481854  -4.3829455
  1.5676743   0.29733747 -1.5906522   5.099951    1.3984878  -2.5427542
  2.3415697  -0.5322488   1.1859593   2.8046238  -2.319534   -1.5975673
 -1.9541273   1.12719     0.52480924 -0.7334962   3.2115426  -4.930841
 -1.0315009  -1.4635807   0.4215247   0.63209593  0.10335546 -2.249634
  0.10248864 -3.4661367  -1.0411446   2.9805512  -1.2724049  -1.8033476
  0.6938161  -4.3397474  -4.2274704   2.1040442  -3.0417728  -0.695654
  2.0104563  -1.9942942   0.5602244  -2.2543266  -0.5196703  -1.7350476
 -1.5876626   1.8630496   1.2265964   0.5577995   1.5125577  -1.027955
 -2.09221     1.7761525   6.4169593   2.66189     1.2226387   1.8583586
  0.00933917  0.05589342 -1.0777882   4.1776066  -2.4164255  -2.3233685
 -0.11373189 -0.77969456  2.2942984   1.2836883  -0.6158414   2.5823781
  0.84133494 -0.7874491  -0.4480739  -0.22722903 -3.464209   -1.4618119
 -1.3226708   0.8291202  -2.1058211   0.34111485 -0.3040787   1.8007078
  0.7327085   5.141138   -0.4072581  -0.9810349  -0.09684868  0.13757314
  3.213012   -0.38885373 -2.8611238  -0.5259004   0.7222548  -1.5868611
 -0.34265354 -3.6504548  -0.36142778 -0.7879854 ]
```

In [40]:
```
sim_words = w2v_model.wv.most_similar('help')
print(sim_words)
```

```
[('advice', 0.5728095769882202), ('support', 0.5703648924827576), ('vent', 0.
5393311977386475), ('Help', 0.5368391871452332), ('explain', 0.53323906660079
96), ('tell', 0.49944809079170227), ('fix', 0.47687309980392456), ('stop', 0.
4637908637523651), ('talk', 0.4610207974910736), ('accept', 0.441024869680404
66)]
```

t-SNE

In [41]:
```
import matplotlib.cm as cm
```

In [42]:
```
from sklearn.manifold import TSNE
def tsne_plot():
    labels = []
    tokens = []

    # Extracting words and their vectors from our trained model
    for word in model.wv.index_to_key:
        tokens.append(model.wv[word])
        labels.append(word)

    # Train t-SNE
    tsne_model = TSNE(perplexity=45, n_components=2, init='pca', n_iter=2500,
random_state=23)
    new_values = tsne_model.fit_transform(tokens)
    x = []
    y = []

    for value in new_values:
        x.append(value[0])
        y.append(value[1])

    plt.figure(figsize=(16, 16))
    for i in range(len(x)):
        plt.scatter(x[i],y[i])
        plt.annotate(labels[i],
                     xy=(x[i], y[i]),
                     xytext=(5, 2),
                     textcoords='offset points',
                     ha='right',
                     va='bottom')
        plt.xlabel("dimension 1")
        plt.ylabel("dimension 2")
    plt.show()
```

In [43]:
```
type(all_words)
```

Out[43]:
```
pandas.core.series.Series
```

In [44]:
```
# Words that occur atleast 50 times
model = Word2Vec(all_words, window=20, min_count=50, workers=4)
```

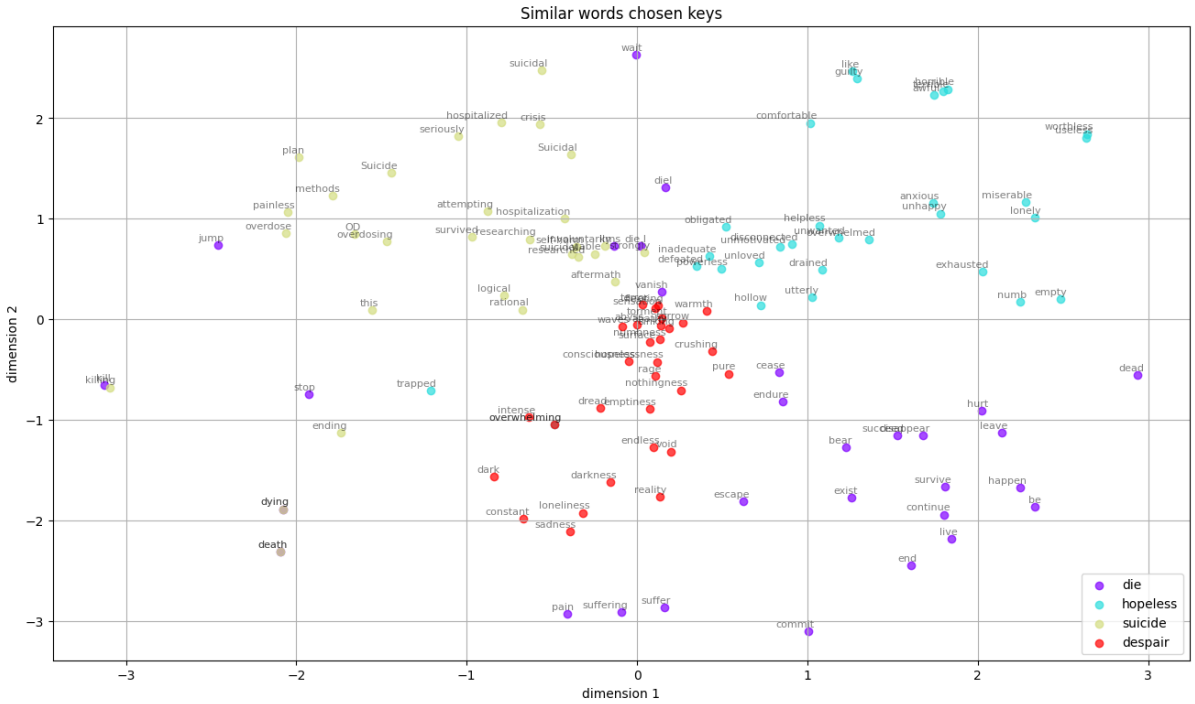
In [45]:
```
model.wv.most_similar('die')
```

Out[45]:
```
[('disappear', 0.6950932145118713),
 ('live', 0.6424233913421631),
 ('death', 0.6397140026092529),
 ('end', 0.610454797744751),
 ('survive', 0.591528058052063),
 ('suffer', 0.5883954167366028),
 ('dieI', 0.5737757682800293),
 ('kill', 0.555905818939209),
 ('be', 0.5555461645126343),
 ('continue', 0.5551039576530457)]
```

In [46]:
```
keys = ['die', 'hopeless', 'suicide', 'despair']

# this array will contain the vectors(dimension 100) and the labels
embedding_clusters = []
word_clusters = []
for word in keys:
    embeddings = []
    words = []
    for similar_word, _ in model.wv.most_similar(word, topn=30):
        words.append(similar_word)
        embeddings.append(model.wv[similar_word])
    embedding_clusters.append(embeddings)
    word_clusters.append(words)
```

In [47]:
```
embedding_clusters = np.array(embedding_clusters)
n, m, k = embedding_clusters.shape
tsne_model_en_2d = TSNE(perplexity=50, n_components=2, init='pca', n_iter=350
0, random_state=32)
embeddings_en_2d = np.array(tsne_model_en_2d.fit_transform(embedding_clusters.
reshape(n * m, k))).reshape(n, m, 2)
```

In [48]:
```
def tsne_plot_similar_words(title, labels, embedding_clusters, word_clusters,
a, filename=None):
    plt.figure(figsize=(16, 9))
    colors = cm.rainbow(np.linspace(0, 1, len(labels)))
    for label, embeddings, words, color in zip(labels, embedding_clusters, wor
d_clusters, colors):
        x = embeddings[:, 0]
        y = embeddings[:, 1]
        plt.scatter(x, y, c=color, alpha=a, label=label)
        for i, word in enumerate(words):
            plt.annotate(word, alpha=0.5, xy=(x[i], y[i]), xytext=(5, 2),
                         textcoords='offset points', ha='right', va='bottom',
size=8)
    plt.legend(loc=4)
    plt.title(title)
    plt.grid(True)
    plt.xlabel("dimension 1")
    plt.ylabel("dimension 2")
    if filename:
        plt.savefig(filename, format='png', dpi=150, bbox_inches='tight')
    plt.show()


tsne_plot_similar_words('Similar words chosen keys', keys, embeddings_en_2d, w
ord_clusters, 0.7,
                        'similar_words.png')
```



LDA

In [49]:
```
import re, nltk
from nltk.corpus import stopwords
nltk.download('stopwords')

def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']):
    output = []
    for sent in texts:
        doc = nlp(" ".join(sent))
        output.append([token.lemma_ for token in doc if
        token.pos_ in allowed_postags])
    return output
# function to remove stopwords
def remove_stopwords(rev):
    rev_new = " ".join([i for i in rev if i not in stop_words])
    return rev_new
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [50]:
```
# remove short words (length < 3)
df['Text_clean'] = df['Text_clean'].apply(lambda x: ' '.join([w for
                   w in x.split() if len(w)>2]))
# remove stopwords from the text
text = [remove_stopwords(r.split()) for r in df['Text_clean']]
# make entire text lowercase
text = [r.lower() for r in text]
```

In [51]:
```
tokenized_text = pd.Series(text).apply(lambda x: x.split())
print(tokenized_text[1])
```

```
['weird', 'dont', 'get', 'affected', 'compliments', 'coming', 'someone', 'kno
w', 'irl', 'feel', 'really', 'good', 'internet', 'strangers']
```

In [52]:
```
nlp = spacy.load("en_core_web_sm")
text_2 = lemmatization(tokenized_text)
print(text_2[1]) # print lemmatized headline
```

```
['weird', 'get', 'affect', 'compliment', 'come', 'know', 'irl', 'feel', 'real
ly', 'good', 'internet', 'stranger']
```

In [54]:
```
ntopics = 4
dictionary = corpora.Dictionary(text_2)
doc_term_matrix = [dictionary.doc2bow(text) for lyric in text_2]
```

In [55]:
```
import time
t0 = time.time()
# Creating the object for LDA model using gensim library
LDA = gensim.models.ldamodel.LdaModel
# Build LDA model
lda_model = LDA(corpus=doc_term_matrix, id2word=dictionary,
                num_topics=ntopics, random_state=100, chunksize=1000,
                passes=50)
print('\nThe LDA_MODEL dataset done in '+'%s seconds'% (time.time() - t0))
```

```
The LDA_MODEL dataset done in 677.3147797584534 seconds
```

In [56]:
```
## word lists
for i in range(0,ntopics):
    temp = lda_model.show_topic(i, 10)
    terms = []
    for term in temp:
        terms.append(term)
    print("\nTop 10 terms for topic #" + str(i) + ": "+ ", ".join([i[0] for i
in terms]))
```

```
Top 10 terms for topic #0: tried, heartdelete, tell, life, gorl, going, hey,
helphelp, isk, bored

Top 10 terms for topic #1: tried, heartdelete, tell, life, gorl, going, hey,
helphelp, isk, bored

Top 10 terms for topic #2: tried, heartdelete, tell, life, gorl, going, hey,
helphelp, isk, bored

Top 10 terms for topic #3: hello, woke, heartdelete, tell, life, gorl, going,
tried, bored, sexy
```

In [57]:
```
def get_lda_topics(model, num_topics):
    word_dict = {};
    for i in range(ntopics):
        words = model.show_topic(i, topn = 20);
        word_dict['Topic # ' + '{:02d}'.format(i+1)] = [i[0] for i in words];
    return pd.DataFrame(word_dict);
```

In [58]:
```
get_lda_topics(lda_model, ntopics)
```

Out[58]:

	Topic # 01	Topic # 02
0	going	going
1	therelol	therelol
2	goodbye	goodbye
3	read	read
4	deleted	deleted
5	curious	curious
6	anything	anything
7	inadequateplease	inadequateplease
8	cthgisnialpnidenodnaba60117102mocefilretfalaru...	cthgisnialpnidenodnaba60117102mocefilretfalaru... ct
9	tell	tell
10	life	life
11	gorl	gorl
12	tried	tried
13	heartdelete	heartdelete
14	bored	bored
15	helphelp	helphelp
16	hey	hey
17	isk	isk
18	sexy	sexy
19	pog	pog

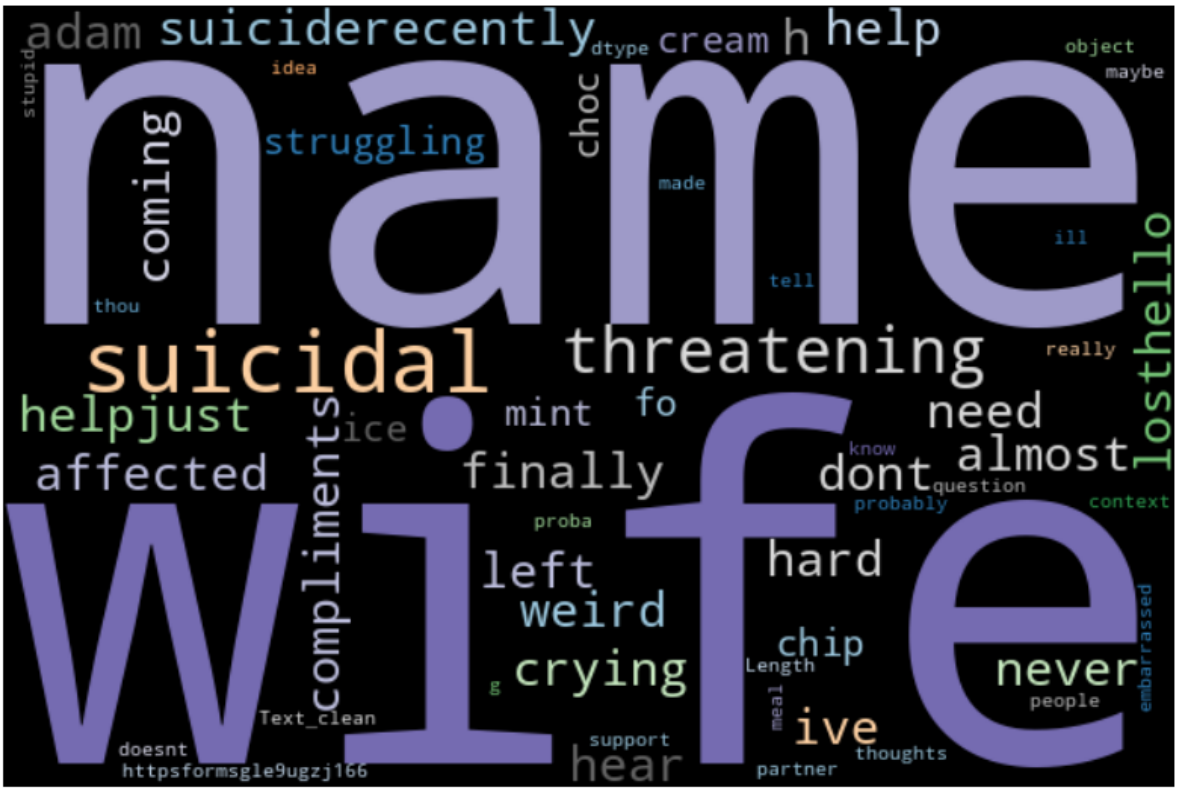
In [59]:
```
# Assign each document to most prevalent topic
lda_topic_assignment = [max(p,key=lambda item: item[1]) for p in lda_model[doc
_term_matrix]]
```

In [60]:
```
Topic_list_cln =[p[0] for p in lda_topic_assignment ]
```

In [61]:
```
df["Topic_LDA"] = Topic_list_cln
```

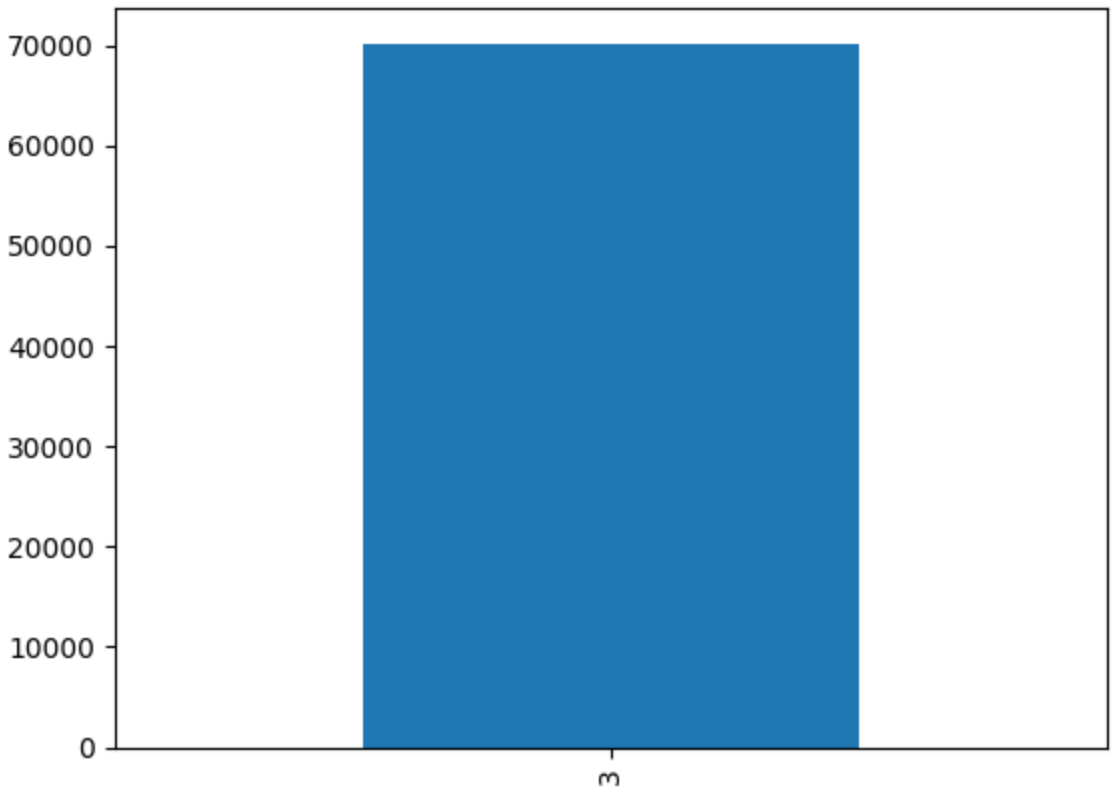
Vizualisation using Wordcloud

In [69]:
```
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
stopwords = set(STOPWORDS)

from wordcloud import WordCloud
cloud=WordCloud(colormap='tab20c',width=600,height=400).generate(str(df["Text_
clean"]))
fig=plt.figure(figsize=(10,15))
plt.axis("off")
plt.imshow(cloud,interpolation='bilinear')
plt.show()
```

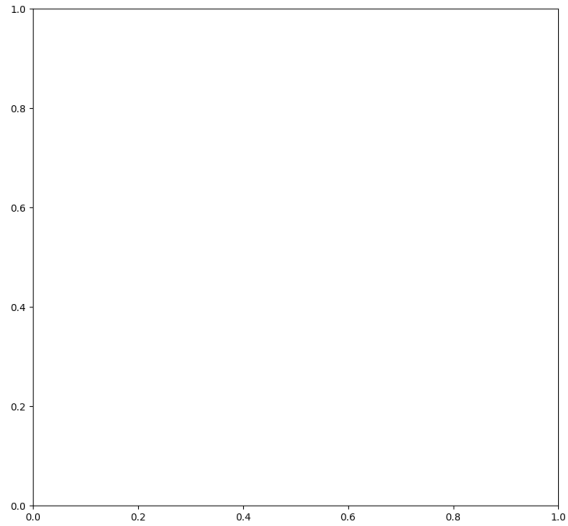


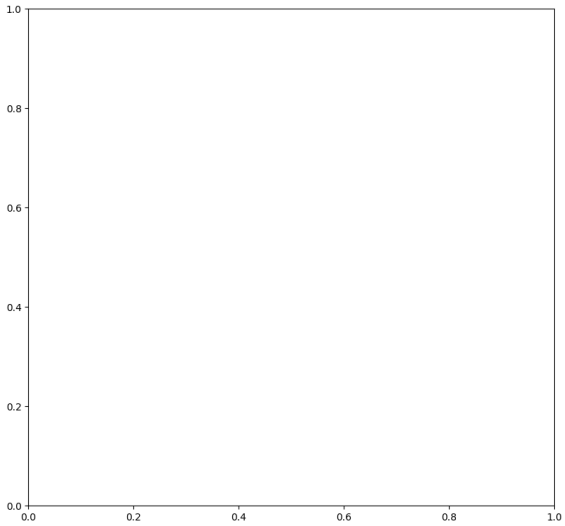
LDA Topics

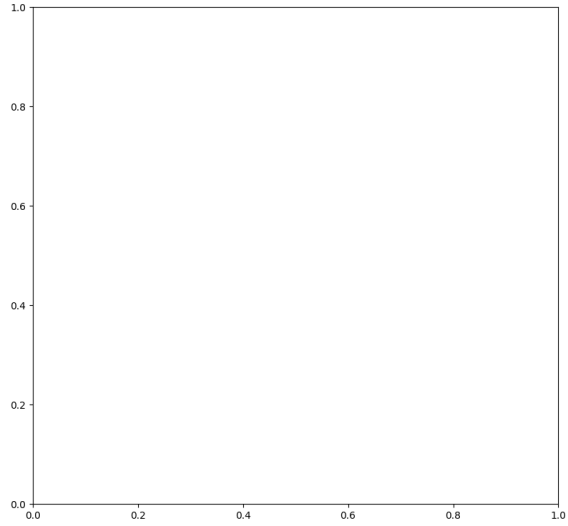
In [70]:
```
df['Topic_LDA'].value_counts().plot(kind = 'bar')
plt.show()
```

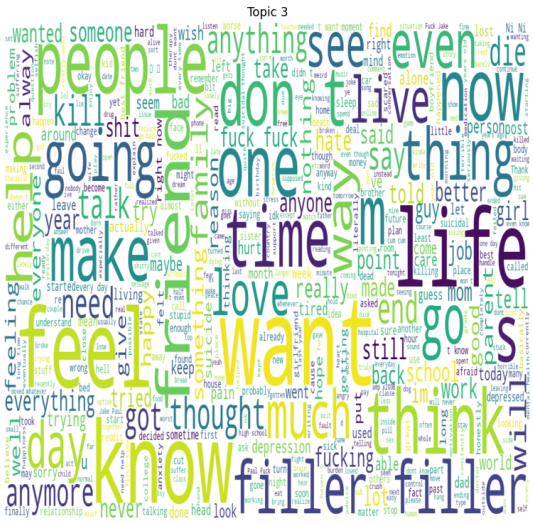


In [71]:
```
fig, axs = plt.subplots(2,2, figsize=(21,20))

for item  in enumerate(list(df['Topic_LDA'].unique())):
    wc = WordCloud(background_color="White",stopwords = stopwords,
                   max_words=1000, max_font_size= 200,  width=1600, height=800,min
_font_size = 10)
    wc.generate(" ".join(df[df['Topic_LDA']== item[1]]['text']))

    axs[item[1]//2, item[1]%2].set_title("Topic %d" % item[1])
    axs[item[1]//2, item[1]%2].imshow(wc, aspect='auto', interpolation='biline
ar')
    axs[item[1]//2, item[1]%2].axis("off")
```









NMF Topics

In [72]:
```
vectorizer = CountVectorizer(analyzer='word', max_features=5000);
cln_counts = vectorizer.fit_transform(df['Text_clean'])
```

In [73]:
```
transformer_cln = TfidfTransformer(smooth_idf=False);
x_tfidf_cln = transformer_cln.fit_transform(cln_counts);
```

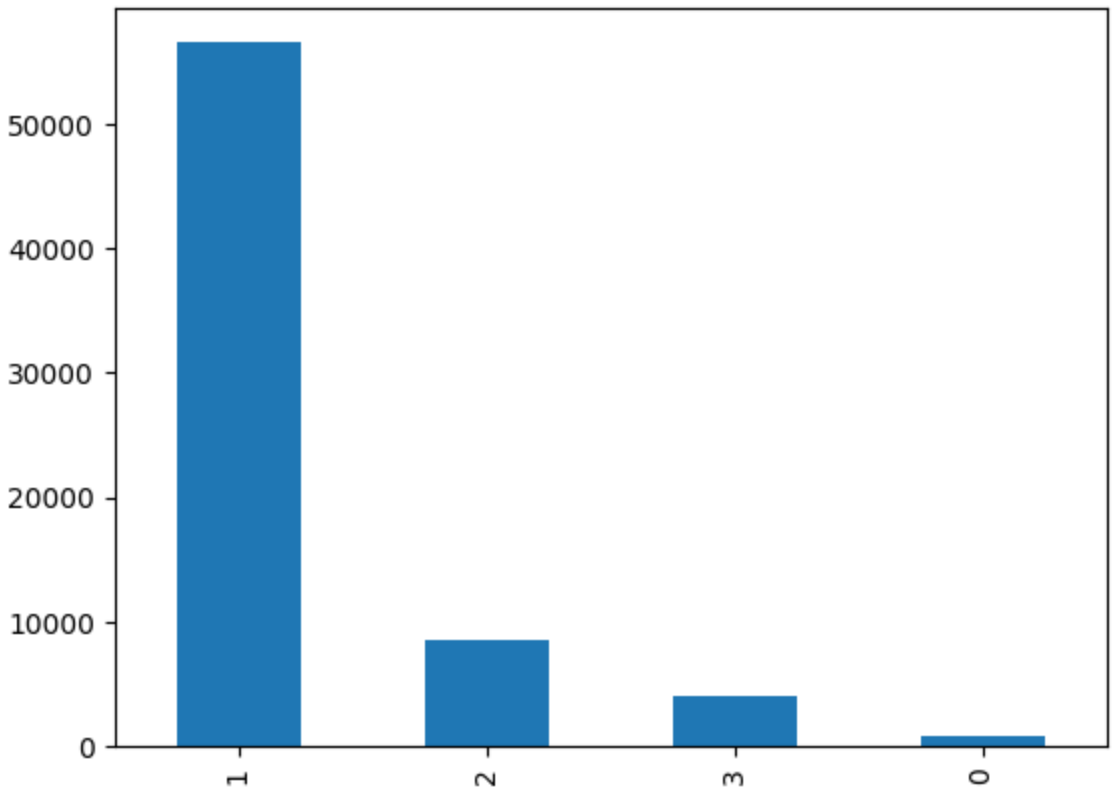
In [74]:
```
xtfidf_norm_cln = normalize(x_tfidf_cln, norm='l1', axis=1)
```

In [75]:
```
#obtain a NMF model.
model_cln = NMF(n_components=ntopics, random_state = 50,init='nndsvd');
#fit the model
W_mat_cln = model_cln.fit_transform(xtfidf_norm_cln)
H_mat_cln = model_cln.components_
```

In [76]:
```
df["Topic_NMF"] = np.argmax(W_mat_cln, axis =1)
```

In [77]:
```
def get_nmf_topics(model, n_top_words):

    #the word ids obtained need to be reverse-mapped to the words so we can pr
int the topic names.
    feat_names = vectorizer.get_feature_names()

    word_dict = {};
    for i in range(ntopics):

        #for each topic, obtain the largest values, and add the words they map
to into the dictionary.
        words_ids = model.components_[i].argsort()[:-20 - 1:-1]
        words = [feat_names[key] for key in words_ids]
        word_dict['Topic # ' + '{:02d}'.format(i+1)] = words;

    return pd.DataFrame(word_dict);
```

In [78]:
```
df['Topic_NMF'].value_counts().plot(kind = 'bar')
plt.show()
```



In [79]:
```
fig, axs = plt.subplots(2,2, figsize=(21,20))

for item  in enumerate(list(df['Topic_NMF'].unique())):
    wc = WordCloud(background_color="White",stopwords = stopwords,
                   max_words=1000, max_font_size= 200,  width=1600, height=800,min
_font_size = 10)
    wc.generate(" ".join(df[df['Topic_NMF']== item[1]]['text']))

    axs[item[1]//2, item[1]%2].set_title("Topic %d" % item[1])
    axs[item[1]//2, item[1]%2].imshow(wc, aspect='auto', interpolation='biline
ar')
    axs[item[1]//2, item[1]%2].axis("off")
```

In [79]: