```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
!pip install contractions
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
Collecting contractions
  Downloading contractions-0.1.72-py2.py3-none-any.whl (8.3 kB)
Collecting textsearch>=0.0.21
  Downloading textsearch-0.0.24-py2.py3-none-any.whl (7.6 kB)
Collecting pyahocorasick
  Downloading pyahocorasick-1.4.4-cp37-cp37m-manylinux_2_17_x86_64.manylinux2
     |████████████████████████████████| 106 kB 5.2 MB/s
Collecting anyascii
  Downloading anyascii-0.3.1-py3-none-any.whl (287 kB)
     |████████████████████████████████| 287 kB 37.1 MB/s
Installing collected packages: pyahocorasick, anyascii, textsearch, contracti
Successfully installed anyascii-0.3.1 contractions-0.1.72 pyahocorasick-1.4.4
```

```
import pandas as pd
import contractions
import re
import string
# Importing wordcloud for plotting word clouds and textwrap for wrapping longe
from wordcloud import WordCloud, STOPWORDS
from textwrap import wrap
# For visualizations
import matplotlib.pyplot as plt
```

0. Loading and preparing data

```
# import pandas as pd
# jsonObj = pd.read_json(path_or_buf= "/content/drive/MyDrive/DATA SCIENCE/CAS
```

```
# jsonObj
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

	id	conversation_id	text	speaker	meta	reply-to	timesta
0	L1045	L1044	They do not!	u0	{'movie_id': 'm0', 'parsed': [{'rt': 1, 'toks'...	L1044	N
1	L1044	L1044	They do to!	u2	{'movie_id': 'm0', 'parsed':	None	N

```
# text_data = jsonObj["text"]
# text_data
```

```
0                                          They do not!
1                                           They do to!
2                                            I hope so.
3                                             She okay?
4                                             Let's go.
                                ...
304708    Lord Chelmsford seems to want me to stay back ...
304709    I'm to take the Sikali with the main column to...
304710                              Your orders, Mr Vereker?
304711    Good ones, yes, Mr Vereker. Gentlemen who can ...
304712    Colonel Durnford... William Vereker. I hear yo...
Name: text, Length: 304713, dtype: object
```

```
# text_data.to_csv("/content/drive/MyDrive/DATA SCIENCE/CASE STUDY 2/Movie_quo
```

```
movie_quotes = pd.read_csv("/content/drive/MyDrive/DATA SCIENCE/CASE STUDY 2/M
movie_quotes
```

1. EDA

1 They do to!

```
df = movie_quotes
df.head()
```

	text
0	They do not!
1	They do to!
2	I hope so.
3	She okay?
4	Let's go.
304712	Colonel Durnford... William Vereker, I hear yo

1.1 Checking for null values and removing them

```
df.isnull().sum()
```

```
text    267
dtype: int64
```

```
df.dropna(subset=['text'],inplace = True)
df.isnull().sum()
```

```
text    0
dtype: int64
```

There were 267 dialogues, that were null, so we removed them from our dataset.

1.2 Removing contractions

```
' '.join(df['text'].tolist())
```

```
'They do not! They do to! I hope so. She okay? Let\'s go. Wow Okay -- you\'re
ometimes you
e my fear of
d get to the
good stuff eventually. Thank God!  If I had to hear one more story about your
coiffure... Me.  This endless...blonde babble. I\'m like, boring myself. Wha
```



```
df['text_new'] = df.apply(lambda row : contractions.fix(row['text']), axis = 1
```

```
' '.join(df['text_new'].tolist())
```

```
'They do not! They do to! I hope so. She okay? Let us go. Wow Okay -- you are
going to need to learn how to lie. No I am kidding.  You know how sometimes y
ou just become this "persona"?  And you do not know how to quit? Like my fear
of wearing pastels? The "real you". What good stuff? I figured you would get
to the good stuff eventually. Thank God!  If I had to hear one more story abo
ut your coiffure... Me.  This endless ...blonde babble. I am like, boring mys
```

```
df.head()
```

	text	text_new
0	They do not!	They do not!
1	They do to!	They do to!
2	I hope so.	I hope so.
3	She okay?	She okay?
4	Let's go.	Let us go.

```
df = df.drop(columns=['text'])
df.head()
```

	text_new
0	They do not!
1	They do to!
2	I hope so.
3	She okay?
4	Let us go.

```
1
```

1.3 Turning to lower case

```
df['cleaned']=df['text_new'].apply(lambda x: x.lower())
' '.join(df['text_new'].tolist())
```

```
'They do not! They do to! I hope so. She okay? Let us go. Wow Okay -- you are
going to need to learn how to lie. No I am kidding.  You know how sometimes y
ou just become this "persona"?  And you do not know how to quit? Like my fear
of wearing pastels? The "real you". What good stuff? I figured you would get
to the good stuff eventually. Thank God!  If I had to hear one more story abo
ut your coiffure... Me.  This endless ...blonde babble. I am like, boring mys
```


1.4 Removing punctuations

```
df['cleaned']=df['cleaned'].apply(lambda x: re.sub('[%s]' % re.escape(string.p
' '.join(df['text_new'].tolist())
```

```
'They do not! They do to! I hope so. She okay? Let us go. Wow Okay -- you are
going to need to learn how to lie. No I am kidding.  You know how sometimes y
ou just become this "persona"?  And you do not know how to quit? Like my fear
of wearing pastels? The "real you". What good stuff? I figured you would get
to the good stuff eventually. Thank God!  If I had to hear one more story abo
ut your coiffure... Me.  This endless ...blonde babble. I am like, boring mys
```

1.5 Removing extra spaces

```
df['cleaned']=df['cleaned'].apply(lambda x: re.sub(' +',' ',x))
' '.join(df['cleaned'].tolist())
```

```
'they do not they do to i hope so she okay let us go wow okay you are going t
o need to learn how to lie no i am kidding you know how sometimes you just be
come this persona and you do not know how to quit like my fear of wearing pas
tels the real you what good stuff i figured you would get to the good stuff e
ventually thank god if i had to hear one more story about your coiffure me th
is endless blonde babble i am like boring myself what crap do you listen to t
```

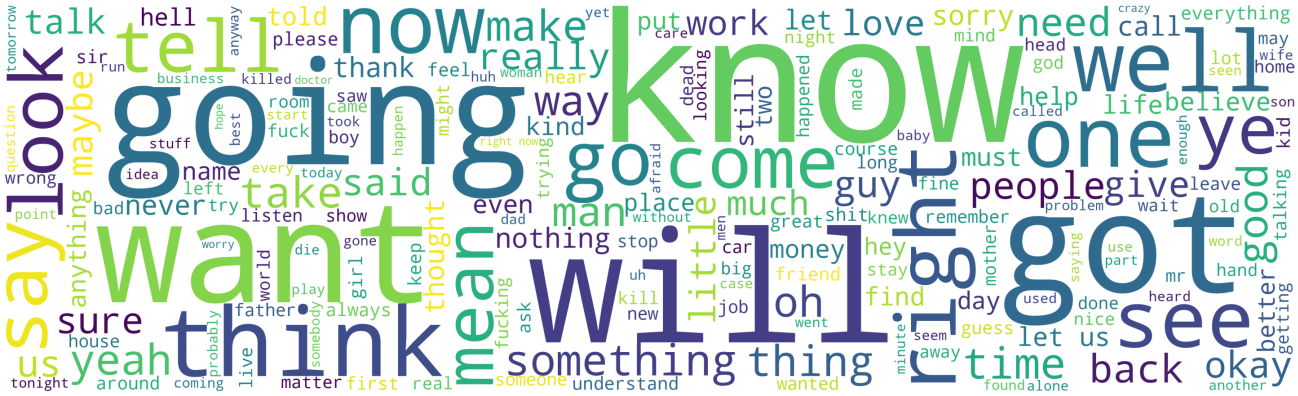
1.5 Stopwords

```
all_words = ""
for quote  in df.cleaned:
  # split the value
  tokens = quote.split()
  all_words += " ".join(tokens)+" "
```

```
stopwords = set(STOPWORDS)


wordcloud = WordCloud(width = 4000, height = 1200,
                background_color ='white',
                stopwords = stopwords,
                min_font_size = 10).generate(all_words)

# plot the WordCloud image
plt.figure(figsize = (40, 12), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
```



1

Automatic saving failed. This file was updated remotely or in another tab. Show diff

✓ 0s completed at 20:05