[illegible] preparation and analyzing pipeline

Aim: TO learn and process [illegible] [illegible] [illegible] to a text . the identify the most frequent words.

Algorithm:-

1- load dataset:

Import [illegible] [illegible] from [illegible]. — [illegible] csv.

2- Clean text:-

Convert to lower remove punctuation and special character [illegible] [illegible] spaces.

3- remove stopwords:-

filter and remove stopwords and [illegible] token

4- Apply cleaning:-

[illegible] the [illegible] [illegible] text column to generate clean.

5- Analyze freq:-

plot all token count word freq [illegible] [illegible]

Program:-

```
import pandas as pd
import re
import spacy
nlp = spacy.load ("en_core_web_sm")
df = pd.read_csv ("[illegible] - [illegible] - csv)
Print( df (raw text)) .head() )
def clean text_ process( text ):
    if pd. isnull (text)
return []
```

Output:

we got this for my husband

who is an (OTR)

1.) I am a proffessional OTR truck driver

2.) well, what can I say

3.) not going to write a long review, even though

4.) I've had mine for a year name. review type : Text

```
text = lext . lowe[)
text = rersed (z[n\n\13]. text )
lent = ctral. craled ("adits - leym py'_ allout ('aul'' ))
ulor = n (p(tore) )
of noo is .et and not ... in pord
words . str-
Print last [' ouis tend', 'clearnd . doth ]. head . (500 )
vall - lothos = [ tor of the in dyh danl . ths >
      for ito . Verg in ulee )
 from collution import dount
word - frer = counte call - ctr )
Print (' TP is from words in Agron ravics,, ')
Print ( weed - frequnt , most . common (15 ))
```

result-:

cleaned dotes contains meaningful words withou- net- and the doot frequity wak output the items from the vera.