In [1]:

```
import numpy as np
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from textblob import Word, TextBlob
from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

In [2]:

```
def wiki_preprocess(text, Barplot=False, Wordcloud=False, Tokenize=False, Lemmatize=False):
    """
    Metinler üzerinde ön işleme  işlemlerini yapar.

    :param text: text Frame'deki metinlerin olduğu değişken
    :param Barplot: Barplot görselleştirme
    :param WordCloud: Word cloud görselleştirme
    :param Tokenize: Cümleleri kelimelere ayırma
    :param Lemmatize: Kelimeri köklerine ayırma
    :return: text

    Example:
        wiki_preprocess(textframe(["text"]))
    """
    # Normalizing Case Folding
    text=text.str.lower()

    # Punctuations
    text=text.apply(lambda x: re.sub("[^\w\s]","",str(x)))
    text=text.apply(lambda x: re.sub("\n"," ",str(x)))
    text=text.apply(lambda x: re.sub("â"," ",str(x)))

    # Numbers
    text=text.fillna('').apply(lambda x: ''.join([i for i in x if not i.isdigit()]))

    # Stop Words
    sw=stopwords.words("english")
    text=text.apply(lambda x: " ".join(x for x in str(x).split() if x not in sw))

    # Rare Words / Custom Words
    clear_text=pd.Series(" ".join(text).split()).value_counts()[-1000:]
    text=text.apply(lambda x: " ".join(x for x in str(x).split() if x not in clear_text))

    if Barplot:
        tf=pd.Series(" ".join(text).split()).value_counts()
        tf=pd.DataFrame(tf)
        tf.reset_index(inplace=True)
        tf.columns=["words","counts"]
        tf[tf["counts"]>10000].plot.bar(x="words",y="counts")
        plt.show()

    if Wordcloud:
        wordcloud_text=" ".join(i for i in text)
        wordcloud=WordCloud(max_font_size=10000,
                            max_words=1000,
                            background_color="black").generate(wordcloud_text)
        plt.figure()
        plt.imshow(wordcloud,interpolation="bilinear")
        plt.axis("off")
        plt.show()

    if Tokenize:
        text=text.apply(lambda x: TextBlob(x).words)
    if Lemmatize:
        text=text.apply(lambda x: " ".join(Word(x).lemmatize() for x in str(x).split()))

    return text
```

In [3]:

```
df=pd.read_csv("C:\\Users\\Dell\\Desktop\\NLP\\Case Study 2\\wiki_data.csv",index_col=0)
df.head()
```

Out[3]:

	text
1	Anovo\n\nAnovo (formerly A Novo) is a computer...
2	Battery indicator\n\nA battery indicator (also...
3	Bob Pease\n\nRobert Allen Pease (August 22, 19...
4	CAVNET\n\nCAVNET was a secure military forum w...
5	CLidar\n\nThe CLidar is a scientific instrumen...

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 10859 entries, 1 to 10859
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    10859 non-null  object
dtypes: object(1)
memory usage: 169.7+ KB
```

In [5]:

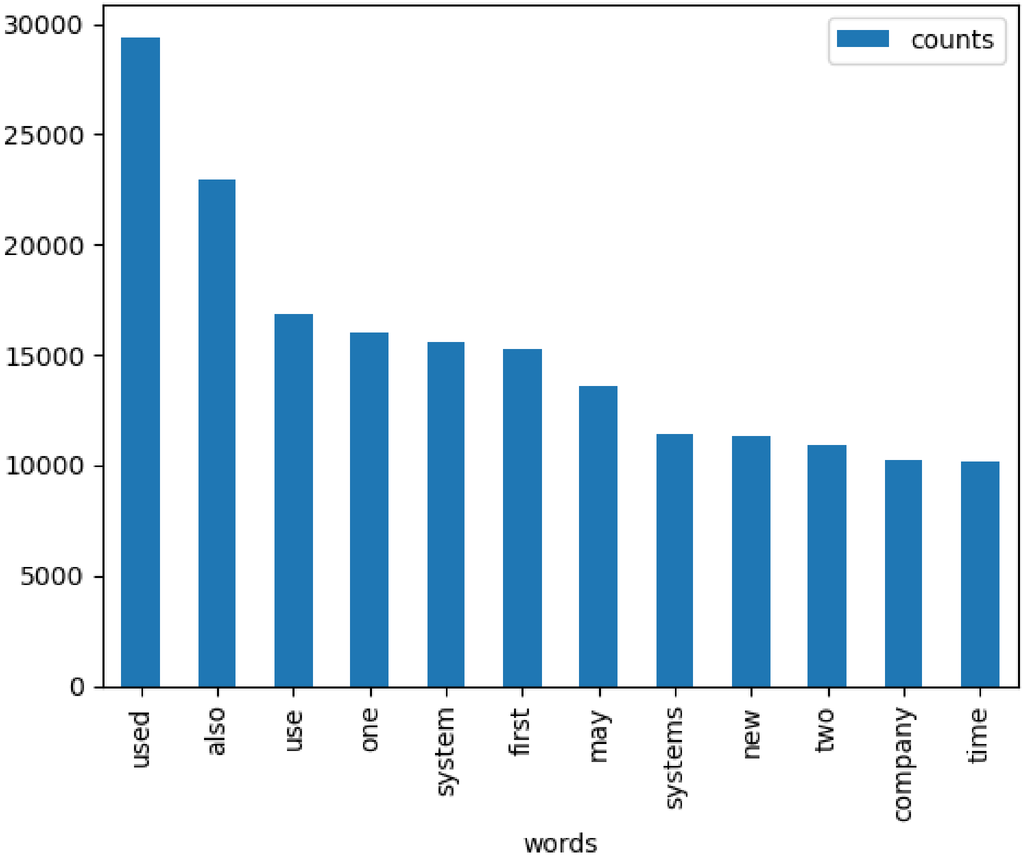
```
df.iloc[2:3].values
```

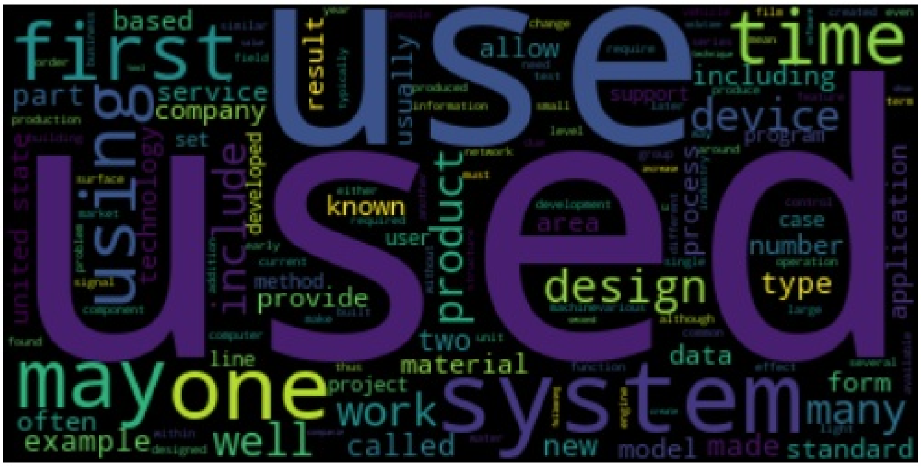
Out[5]:

```
array([['Bob Pease\n\nRobert Allen Pease (August 22, 1940Â\xa0â€" June 18, 2011) was an a
nalog integrated circuit design expert and technical author. He designed several very suc
cessful "best-seller" integrated circuits, many of them in continuous production for mult
iple decades. These include the LM331 voltage-to-frequency converter, and the LM337 adjus
table negative voltage regulator (complement to the LM317).\n\nPease was born on August 2
2, 1940 in Rockville, Connecticut. He attended Northfield Mount Hermon School in Massachu
setts, and subsequently obtained a Bachelor of Science in Electrical Engineering (BSEE) d
egree from Massachusetts Institute of Technology in 1961.\n\nHe started work in the early
1960s at George A. Philbrick Researches (GAP-R). GAP-R pioneered the first reasonable-cos
t, mass-produced operational amplifier (op-amp), the K2-W. At GAP-R, Pease developed many
high-performance op-amps, built with discrete solid-state components.\n\nIn 1976, Pease m
oved to National Semiconductor Corporation (NSC) as a designer and applications engineer,
where he began designing analog monolithic integrated circuits, as well as design referen
ce circuits using these devices. He had advanced to staff scientist by the time of his de
parture in 2009. During his tenure at NSC, he began writing a popular continuing monthly
column called "Pease Porridge" in "Electronic Design" about his experiences in the world
of electronic design and application.\n\nTHOR-LVX (photo-nuclear) microtron Advanced Expl
osives contraband Detection System: "A Dual-Purpose Ion-Accelerator for Nuclear-Reaction-
Based Explosives-and SNM-Detection in Massive Cargo" was the last project he was designin
g for.\n\nPease was the author of eight books, including "Troubleshooting Analog Circuits
", and held 21 patents.\n\nHis other interests included hiking and biking in remote place
s, and working on his old Volkswagen Beetle, which he often mentioned in his columns. Pea
se\'s writing was "strongly opinionated, but he could communicate with a wry sense of hum
or that endeared him to readers whether they agreed with him or not".\n\nPease was killed
in the crash of his 1969 Volkswagen Beetle, on June 18, 2011. He was leaving a gathering
in memory of Jim Williams, who was another well-known analog circuit designer, a technica
l author, and a renowned staff engineer working at Linear Technology. Pease was 70 years
old, and was survived by his wife, two sons, and three grandchildren. The sudden death of
Pease triggered a small flood of remembrances and tributes from fellow technical writers,
practicing engineers, and electronics hardware hacking enthusiasts.\n\n\n\n']]
practicing engineers, and electronics hardware hacking enthusiasts.\n\n\n\n ']],
      dtype=object)
```

In [6]:

```
wiki_preprocess(df["text"],True,True)
```





Out[6]:

```
1        anovo anovo formerly novo computer services co...
2        battery indicator battery indicator also known...
3        bob pease robert allen pease august june analo...
4        cavnet cavnet secure military forum became ope...
5        clidar clidar scientific instrument used measu...
                               ...
10855    soundcast soundcast llc privately funded compa...
10856    spectrum analyzer spectrum analyzer measures m...
10857    telepresence technology telepresence technolog...
10858    transpacific profiler network transpacific pro...
10859    transfer case transfer case part drivetrain fo...
Name: text, Length: 10859, dtype: object
```

In [7]:

```
data=wiki_preprocess(df["text"],Lemmatize=True)
data.iloc[2:3].values
```

Out[7]:

```
array(['bob pea robert allen pea august june analog integrated circuit design expert tech
nical author designed several successful bestseller integrated circuit many continuous pr
oduction multiple decade include lm voltagetofrequency converter lm adjustable negative v
oltage regulator complement lm pea born august rockville connecticut attended northfield 
mount hermon school massachusetts subsequently obtained bachelor science electrical engin
eering bsee degree massachusetts institute technology started work early george philbrick
research gapr gapr pioneered first reasonablecost massproduced operational amplifier opam
p kw gapr pea developed many highperformance opamps built discrete solidstate component p
ea moved national semiconductor corporation nsc designer application engineer began desig
ning analog monolithic integrated circuit well design reference circuit using device adva
nced staff scientist time departure tenure nsc began writing popular continuing monthly c
olumn called pea porridge electronic design experience world electronic design applicatio
n thorlvx photonuclear microtron advanced explosive contraband detection system dualpurpo
se ionaccelerator nuclearreactionbased explosivesand snmdetection massive cargo last proj
ect designing pea author eight book including troubleshooting analog circuit held patent 
interest included hiking biking remote place working old volkswagen beetle often mentione
d column pea writing strongly opinionated could communicate wry sense humor endeared read
er whether agreed pea killed crash volkswagen beetle june leaving gathering memory jim wi
lliams another wellknown analog circuit designer technical author renowned staff engineer
working linear technology pea year old survived wife two son three grandchild sudden deat
h pea triggered small flood remembrance tribute fellow technical writer practicing engine
er electronics hardware hacking enthusiast'],
      dtype=object)
```

In [ ]: