```python
#!/usr/bin/env python
# coding: utf-8

# In[6]:


#task 5: chinese segmentation of jieba
#!pip install jieba
import jieba
seg=jieba.cut("把句子中所有的可以成词的词语都扫描出来",cut_all=True)
print(" ".join(seg))


# In[7]:


#TASK 1: BASIC TEXT PROCESSING PIPELINE
import nltk
texts="""When I dare to be powerful – to use my strength in the service of my vision, then it becomes less and less important whether I am afraid"""
for text in texts:
    sentences=nltk.sent_tokenize(texts)
    for sentence in sentences:
        words=nltk.word_tokenize(sentence)
        tagged=nltk.pos_tag(words)
        print(tagged)

for text in texts:
    words=nltk.word_tokenize(texts)
    for word in words:
        word


# In[9]:


#task 2: tweettokenizer
import nltk
from nltk.tokenize import TweetTokenizer
text='The party was soooo fun: D #superfun'
twtkn=TweetTokenizer()
twtkn.tokenize(text)


# In[12]:


#Task 3: Scrapping Data from Web
from urllib import request
url="http://www.gutenberg.org/files/2554/2554-0.txt"
response=request.urlopen(url)
raw=response.read().decode('utf8')
type(raw)
from nltk.tokenize import word_tokenize
tokens=word_tokenize(raw)
type(tokens)

#HOMEWORK

```

```
#HTML => ASCII => TOKEN => TEXT => VOCAB


# In[18]:


from urllib import request
url = "https://timesofindia.indiatimes.com/"
html = request.urlopen(url).read().decode('utf8')
html[:60]

print(html)


# In[51]:


from nltk.tokenize import word_tokenize
from bs4 import BeautifulSoup
raw = BeautifulSoup(html, 'html.parser').get_text()
tokens = word_tokenize(raw)
print(len(tokens))
print(tokens)


# In[53]:


import nltk
tokens = tokens[:10615]
text = nltk.Text(tokens)
text.concordance('News')


# In[ ]:
```