

- Yêu cầu:

Sử dụng các phần mềm: anaconda 3 version mới nhất, visual studio code

- Các bước thực hiện:

- Cài đặt anaconda
- Cài đặt visual studio code
- Thiết lập môi trường python của visual studio code là python 3
- Chạy dòng lệnh sau:

```
import nltk  
nltk.download()
```

Sẽ xuất hiện một window như vậy:

The screenshot shows the Visual Studio Code interface. The editor window displays a Python file named `test.py` with the following code:

```
1 import nltk
2 nltk.download()
```

The NLTK Downloader window is open, showing a table of available packages. The 'Collections' tab is selected, and the 'all' collection is highlighted. The table lists the following collections:

Identifier	Name	Size	Status
all	All packages	n/a	installed
all-corpora	All the corpora	n/a	installed
all-nltk	All packages available on nltk_data gh-pages branch	n/a	installed
book	Everything used in the NLTK Book	n/a	installed
popular	Popular packages	n/a	installed
tests	Packages for running tests	n/a	installed
third-party	Third-party data packages	n/a	installed

The 'Download' button is visible at the bottom of the window. The terminal window shows the following output:

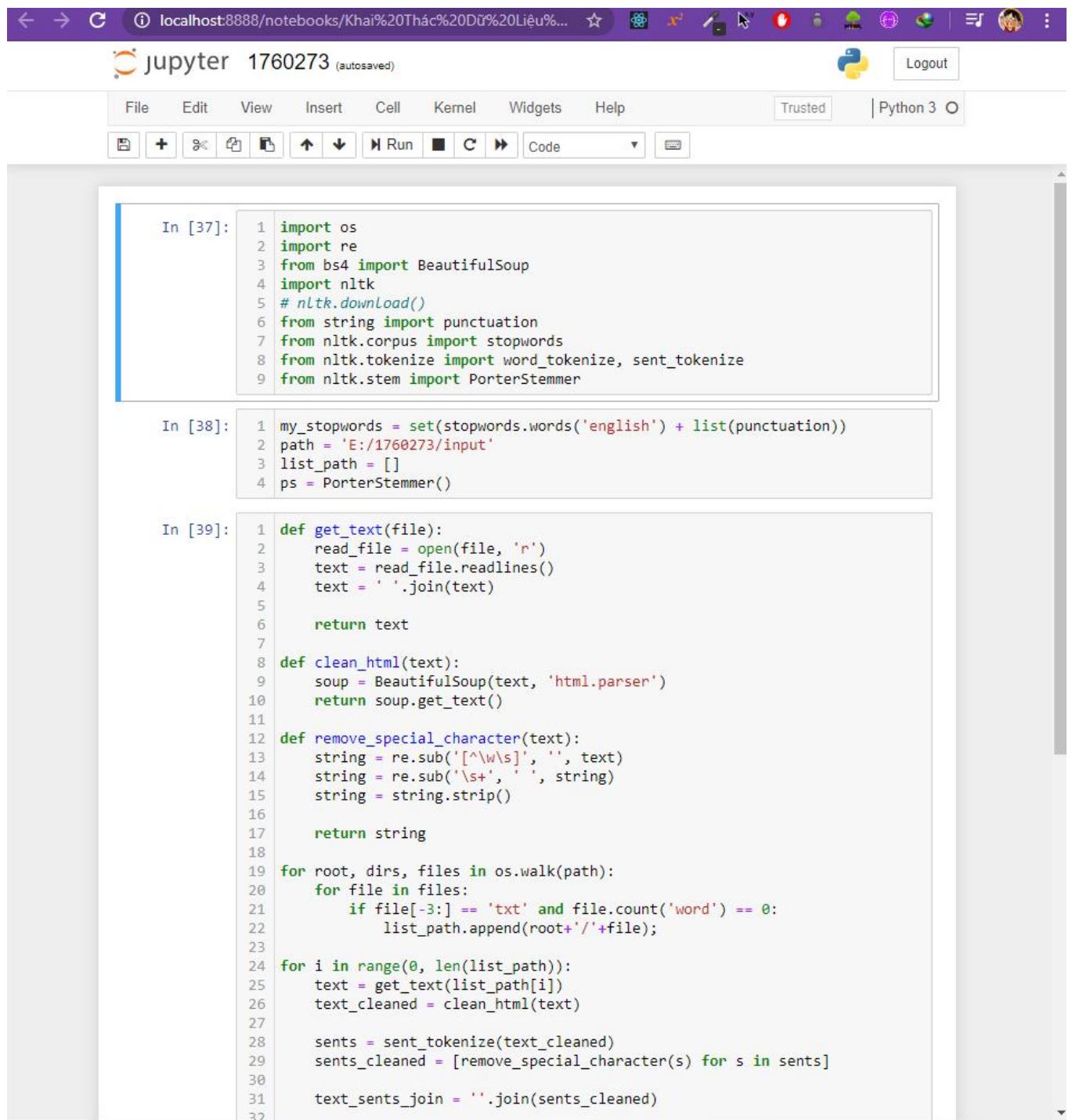
```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS D:\Data-Science\Complete Data Science Training> conda activate base
PS D:\Data-Science\Complete Data Science Training> & C:/Users/cuong/anaconda3/python.exe d:/test.py
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

nhấn button download đến khi tất cả các gói đều xanh như hình trên là thành công.

- Mở file 1760273.ipynb thông qua anaconda

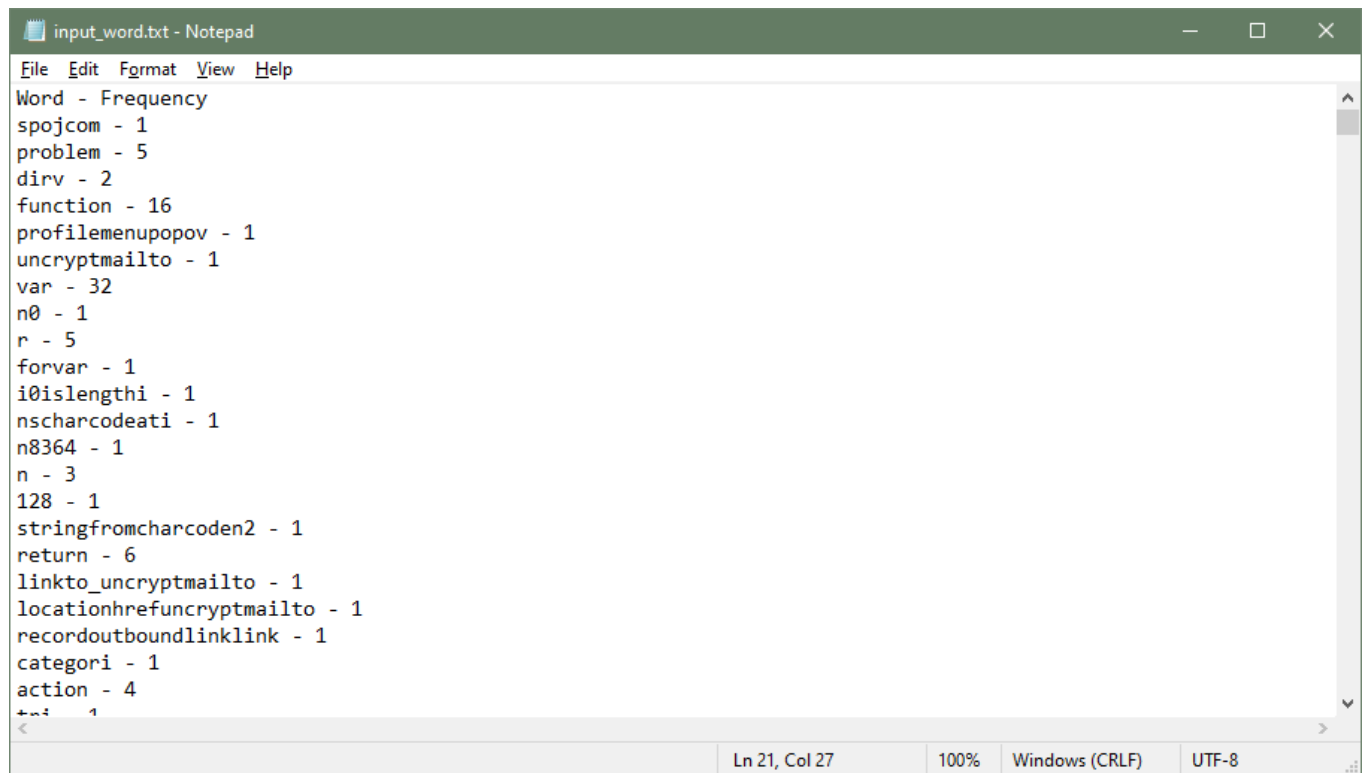


```
In [37]: 1 import os
2 import re
3 from bs4 import BeautifulSoup
4 import nltk
5 # nltk.download()
6 from string import punctuation
7 from nltk.corpus import stopwords
8 from nltk.tokenize import word_tokenize, sent_tokenize
9 from nltk.stem import PorterStemmer

In [38]: 1 my_stopwords = set(stopwords.words('english') + list(punctuation))
2 path = 'E:/1760273/input'
3 list_path = []
4 ps = PorterStemmer()

In [39]: 1 def get_text(file):
2     read_file = open(file, 'r')
3     text = read_file.readlines()
4     text = ''.join(text)
5
6     return text
7
8 def clean_html(text):
9     soup = BeautifulSoup(text, 'html.parser')
10    return soup.get_text()
11
12 def remove_special_character(text):
13    string = re.sub('[^\w\s]', '', text)
14    string = re.sub('\s+', ' ', string)
15    string = string.strip()
16
17    return string
18
19 for root, dirs, files in os.walk(path):
20     for file in files:
21         if file[-3:] == '.txt' and file.count('word') == 0:
22             list_path.append(root+'/'+file);
23
24 for i in range(0, len(list_path)):
25     text = get_text(list_path[i])
26     text_cleaned = clean_html(text)
27
28     sents = sent_tokenize(text_cleaned)
29     sents_cleaned = [remove_special_character(s) for s in sents]
30
31     text_sents_join = ''.join(sents_cleaned)
32
```

- Chọn thẻ cell -> run all
- Do tôi đã chuẩn sẵn 2 trang html được lưu dạng file txt nằm trong thư mục **input**, cái ta cần làm bây giờ là truy cập vào thư mục **output** để xem kết quả.
- Đây là cách cấu trúc cho các file kết quả trong thư mục **output**



The image shows a Notepad window titled "input_word.txt - Notepad". The window contains a list of words and their frequencies, formatted as "Word - Frequency". The words are listed in descending order of frequency. The text is as follows:

```
File Edit Format View Help
Word - Frequency
spojcom - 1
problem - 5
dirv - 2
function - 16
profilemenupopov - 1
uncryptmailto - 1
var - 32
n0 - 1
r - 5
forvar - 1
i0islengthi - 1
nscharcodeati - 1
n8364 - 1
n - 3
128 - 1
stringfromcharcoden2 - 1
return - 6
linkto_uncryptmailto - 1
locationhrefuncryptmailto - 1
recordoutboundlinklink - 1
categori - 1
action - 4
+ni 1
```

The status bar at the bottom indicates the current position is "Ln 21, Col 27", the zoom is "100%", the line ending is "Windows (CRLF)", and the encoding is "UTF-8".

bao gồm từ và tần suất xuất hiện của từ đó.