# **Web Scraping with Requests and Beautiful Soup**

**Author:** Ian Patricio  
**Date:** August 25, 2025  
**Language:** Python  
**Purpose:** This program scrapes the Python.org website, extracts visible text using BeautifulSoup, removes stopwords, and generates a word cloud. It demonstrates practical use of web scraping, text preprocessing, and visualization in NLP. The program follows best practices including modular design and clear documentation.

### **Python Pseudocode:**

# Module: Web Scraping + Word Cloud

# Purpose: Scrape text from python.org, clean it, remove stopwords, and visualize in a word cloud.

# Import libraries:

# requests -> download web content

# BeautifulSoup -> parse HTML

# re -> text cleaning

# wordcloud -> generate word cloud

# matplotlib -> display results

# Step 1: Define URL (https://www.python.org)

# Step 2: Send GET request to fetch HTML

# Step 3: Parse HTML with BeautifulSoup

# - Remove <script> and <style> elements

# - Extract visible text

# Step 4: Clean text

# - Convert to lowercase

# - Remove extra spaces and punctuation

# Step 5: Remove stopwords using WordCloud’s built-in stopword list

# Step 6: Recombine words into cleaned text

# Step 7: Generate WordCloud object

# Step 8: Display with matplotlib

# Step 9: Save word cloud image to file

### **Python Source Code:**

"""

Module: Web Scraping + Word Cloud

Description: This module scrapes python.org, cleans extracted text, removes

stopwords, and generates a word cloud for visualization.

Author: Fernando Ian Patricio

Date: August 25, 2025

"""

import requests

from bs4 import BeautifulSoup

from wordcloud import WordCloud, STOPWORDS

import matplotlib.pyplot as plt

import re

# Download webpage

url = "https://www.python.org"

response = requests.get(url)

html\_content = response.text

# Parse HTML with BeautifulSoup

soup = BeautifulSoup(html\_content, "html.parser")

# Remove scripts and styles

for script in soup(["script", "style"]):

script.extract()

# Extract visible text

text = soup.get\_text(separator=" ")

# Clean text

text = re.sub(r'\s+', ' ', text) # Remove extra spaces

text = text.lower()

# Remove stopwords

stopwords = set(STOPWORDS)

words = [word for word in text.split() if word not in stopwords]

cleaned\_text = " ".join(words)

# Generate WordCloud

wordcloud = WordCloud(

width=1000,

height=600,

background\_color="white",

stopwords=stopwords,

colormap="viridis"

).generate(cleaned\_text)

# Display WordCloud

plt.figure(figsize=(12, 8))

plt.imshow(wordcloud, interpolation="bilinear")

plt.axis("off")

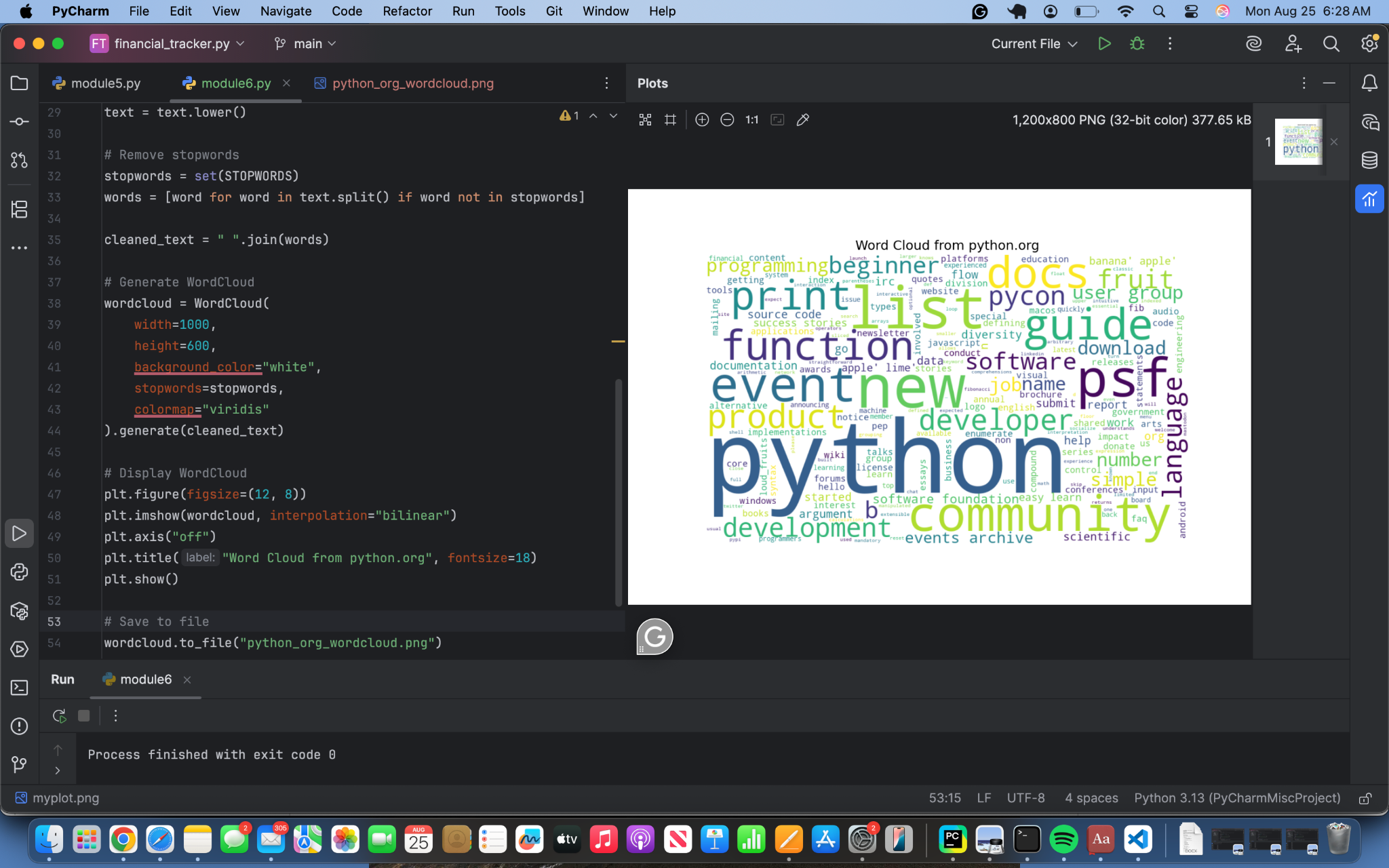
plt.title("Word Cloud from python.org", fontsize=18)

plt.show()

# Save to file

wordcloud.to\_file("python\_org\_wordcloud.png")

### **Screenshot:**

**

### **Git Repository:**

<https://github.com/ianpatricio-csuglobal/CSC484-1>