# Crawler Coding Challenge #

from urllib import request

from bs4 import BeautifulSoup

from collections import Counter

import threading

# Parse the HTML and extracts text from History section.

def extract\_text(url):

    page = request.get(url)

    soup = BeautifulSoup(page.content, "html.parser")

    history\_section = soup.find(id="History")

    text = " ".join([p.text for p in history\_section.find\_all("p")])

    return text

# Iterates through parsed words and counts them

# and converts characters to lower case

def count\_words(text):

    words = text.lower().split()

    return Counter(words)

# Removes punctuation

def process\_text(url):

    text = extract\_text(url)

    word\_counts = count\_words(text)

    return word\_counts.most\_common(10)

# Creates a main method to point the program where to go and what to parse

def main():

    url = "https://en.wikipedia.org/wiki/Microsoft"

    results = []

    threads = []

    for i in range(5):

        t = threading.Thread(target=process\_text, args=(url,))

        t.start()

        threads.append(t)

    # Multithreads for scalability

    for t in threads:

        t.join()

        results.append(process\_text(url))

    # Counts how many times the Top Ten words are used

    word\_counts = Counter()

    for result in results:

        for word, count in result:

            word\_counts[word] += count

    # Prints the results

    print(word\_counts.most\_common(10))

if \_\_name\_\_ == "\_\_main\_\_":

    main()

Second line, where the BeautifulSoup import is, I was having trouble with it on my machine after a Python update. Note that when you run the code. I’m not sure it will compile with older versions of Python or if you have bs3 or earlier versions.

All methods are commented out with what they’re intended to be.

This .doc file is intended as a representation, the code written in Python 3.11.1 and is zipped in a .py file, written in VSCode 1.74.3.

Have a marvelous day!