Nama : Teuku Tamam Al Fatah

NPM : 2008107010071

Tugas 4 TWM Penulisan laporan (10 poin).

1. **Penjelasan umum tentang tahapan yang dilakukan dalam membagi bagian isi atau konten**

   Program dimulai dengan membaca setiap file HTML dalam direktori yang telah ditentukan sebelumnya. Selanjutnya, program mencari dan mengekstrak judul dari setiap file HTML dengan menggunakan pola regex yang sesuai. Setelah itu, isi dari file HTML diekstrak menggunakan modul HTML::ExtractContent, dimana hasil ekstraksi tersebut disimpan dalam variabel $content. Selanjutnya, isi tersebut dibagi menjadi tiga bagian yang memiliki jumlah kalimat yang relatif sama. Setiap bagian isi yang telah dipisahkan disimpan dalam file output, dengan penambahan tag <atas>, <tengah>, dan <bawah> untuk menandai bagian-bagian tersebut. Dengan demikian, program ini memungkinkan untuk melakukan pemrosesan lebih lanjut pada setiap bagian isi HTML secara terpisah, memfasilitasi analisis yang lebih mendalam.

2. **Penjelasan dan perintah yang digunakan untuk mengunduh 8000 file HTML tersebut (15 poin).Tampilan 5 file yang telah dibersihkan (15 poin).**

   Untuk mengunduh 8000 file HTML dari situs web Detik, saya menggunakan skrip Python yang memanfaatkan modul requests untuk membuat permintaan HTTP ke halaman web, dan modul BeautifulSoup untuk melakukan scraping web dan mengekstrak konten HTML. Skrip ini memiliki fungsi bernama DownloadBerita, yang menerima parameter seperti kategori berita, indeks halaman, bulan, tanggal awal, tanggal akhir, dan tahun. Dalam fungsi ini, kami melakukan pengulangan untuk setiap tanggal dalam rentang tanggal yang ditentukan. Setiap iterasi melibatkan pengulangan untuk setiap halaman indeks berita hingga tidak ada artikel berita lagi untuk tanggal tersebut. Kami menggunakan permintaan HTTP untuk mendapatkan halaman HTML untuk setiap halaman indeks berita, dan kemudian menggunakan BeautifulSoup untuk melakukan parsing HTML dan mengekstrak daftar artikel berita. Untuk setiap artikel berita, kami mendapatkan tautan artikel, unduh isi artikel menggunakan permintaan HTTP, dan simpan isi artikel ke dalam file teks dengan nama berdasarkan judul artikel. Proses ini diulang hingga semua berita pada tanggal yang ditentukan telah diunduh.

3. **Tampilan 5 file yang telah dibersihkan**

   https://github.com/TeukuTamamAlfatah/Tugas_4_TWM_2008107010071/tree/main/clean

4. **Lampiran script (dengan aturan penulisan code yang baik) yang dibuat untuk menyelesaikan tugas 1 ini**

   https://github.com/TeukuTamamAlfatah/Tugas_4_TWM_2008107010071