Documentație

Am primit 3 fișiere CSV cu o multitudine de date, având ca obiectiv final unirea celor 3 fișiere în unul complet.

1. Funcția read\_large\_csv pentru citirea CSV-urilor mari:

Citirea fișierului CSV: Funcția utilizează funcția pandas.read\_csv pentru a citi un fișier CSV în bucăți mari (chunks) de câte 100.000 de rânduri, utilizând parametrii chunksize și delimiter.

Gestionarea erorilor: În cazul în care există rânduri cu probleme în fișier (de exemplu, cu un format incorect sau incomplet), aceste rânduri sunt omise cu ajutorul opțiunii on\_bad\_lines='skip'.

Concatenarea bucăților de date (chunks): Datele citite în bucăți sunt adăugate la un DataFrame gol folosind pd.concat(). Astfel, se construiește treptat întregul set de date.

2. Citirea fișierelor CSV:

Trei fișiere sunt citite:

facebook\_dataset.csv

google\_dataset.csv

website\_dataset.csv

Pentru fișierul website\_dataset.csv, se folosește separatorul “;”, deoarece acesta este separatorul folosit în acel fișier.

3. Verificarea coloanelor:

După ce datele din fiecare fișier au fost citite și stocate în DataFrames, se verifică capurile de tabel (columns) pentru a se asigura că au fost citite corect și pentru a confirma existența coloanei domain în fișierele Facebook și Google și root\_domain în fișierul Website.

4. Join-ul pe baza domeniului:

Îmbinarea între Facebook și Google: Se realizează un inner join pe coloana domain, care există în ambele fișiere (facebook\_data și google\_data). Îmbinarea cu Website: După primul join, se face un al doilea inner join cu datele din website\_data, folosind domain din setul de date anterior și root\_domain din fișierul website.

5. Verificarea duplicatelor

Se verifică dacă există coloane duplicate.

6. Afișarea și salvarea rezultatului final:

Setul de date rezultat din îmbinare este salvat într-un nou fișier CSV numit merged\_dataset.csv fără index.