

```
import os
import csv
import sys
import shutil

# CSV limit fix
csv.field_size_limit(10**8)

# 1 Yeni hedef klasör
OUTPUT_ROOT = "dataset_cleaned"

# Orijinal dataset
DATASET_ROOT = "dataset"

# 4 ana klasör
CATEGORIES = ["contracts", "invoices", "notifications", "reports"]

os.makedirs(OUTPUT_ROOT, exist_ok=True)

# 1 CONTRACTS: Part_I + Part_II + Part_III hepsi tek klasöre
contracts_src = os.path.join(DATASET_ROOT, "contracts")
contracts_dst = os.path.join(OUTPUT_ROOT, "contracts")
os.makedirs(contracts_dst, exist_ok=True)

for part in ["Part_I", "Part_II", "Part_III"]:
```

```
part_dir = os.path.join(contracts_src, part)

for folder, _, files in os.walk(part_dir):

    for f in files:

        if f.endswith(".txt"):

            src = os.path.join(folder, f)

            dst = os.path.join(contracts_dst, f)

            shutil.copy2(src, dst)

print(f"\u25aa Contracts merged: {contracts_dst}")

# 2\u25aa INVOICES: alt klas\u011felerin hepsi tek klas\u011fere

invoices_src = os.path.join(DATASET_ROOT, "invoices")

invoices_dst = os.path.join(OUTPUT_ROOT, "invoices")

os.makedirs(invoices_dst, exist_ok=True)

for folder, _, files in os.walk(invoices_src):

    for f in files:

        if f.endswith(".txt"):

            src = os.path.join(folder, f)

            dst = os.path.join(invoices_dst, f)

            shutil.copy2(src, dst)

print(f"\u25aa Invoices merged: {invoices_dst}")

# 3\u25aa NOTIFICATIONS: emails.csv \u25aa her sat\u011f bir txt

notifications_csv = os.path.join(DATASET_ROOT, "notifications", "emails.csv")

notifications_dst = os.path.join(OUTPUT_ROOT, "notifications")

os.makedirs(notifications_dst, exist_ok=True)

with open(notifications_csv, encoding="utf-8") as f:
```

```
reader = csv.reader(f)

header = next(reader) # varsa başlık atla

for i, row in enumerate(reader):

    text = " ".join(row)

    with open(os.path.join(notifications_dst, f"email_{i}.txt"), "w", encoding="utf-8") as out_f:

        out_f.write(text)

print(f"\n Notifications extracted: {notifications_dst}")

# 4 REPORTS: Synthetic_Financial...csv \ her satır bir txt

reports_csv = None

for f in os.listdir(os.path.join(DATASET_ROOT, "reports")):

    if f.endswith(".csv"):

        reports_csv = os.path.join(DATASET_ROOT, "reports", f)

reports_dst = os.path.join(OUTPUT_ROOT, "reports")

os.makedirs(reports_dst, exist_ok=True)

with open(reports_csv, encoding="utf-8") as f:

    reader = csv.reader(f)

    header = next(reader)

    for i, row in enumerate(reader):

        text = " ".join(row)

        with open(os.path.join(reports_dst, f"report_{i}.txt"), "w", encoding="utf-8") as out_f:

            out_f.write(text)

print(f"\n Reports extracted: {reports_dst}")

print("\n DONE! Clean dataset ready:", OUTPUT_ROOT)
```

