Netflix Content Clustering and Recommendation

Name: Amaan Hussain I Mirza

Subject: AI / ML Project

```
import pandas as pd
import numpy as np

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import TruncatedSVD
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.metrics.pairwise import cosine_similarity

df = pd.read_csv("/content/NetflixSimple.csv") # Load Netflix CSV

text_cols = ["description", "listed_in", "cast", "director"] # Select
text columns and fill missing values

for col in text_cols:
    if col not in df.columns:
        df[col] = ""
    df[col] = df[col].fillna("")

df["bag_of_content"] = (
    df["description"] + " " +
    df["listed_in"]   + " " +
    df["cast"]        + " " +
    df["director"]
)

# Combine text columns into one
df = df[df["bag_of_content"].str.strip() != ""]

tfidf = TfidfVectorizer(
    stop_words="english", # Convert text to TF-IDF numbers
    max_features=10000  # control dimensionality
)
tfidf_matrix = tfidf.fit_transform(df["bag_of_content"])

svd = TruncatedSVD(
    n_components=100,   # reduce data size
    random_state=42
)
reduced_features = svd.fit_transform(tfidf_matrix)
```

```
k = 2 0  # Divide content into similar clusters
kmeans = KMeans(
    n_clusters=k,
    random_state=42,
    n_init="auto"
)
cluster_labels = kmeans.fit_predict(reduced_features)
df["cluster"] = cluster_labels

score = silhouette_score(reduced_features, cluster_labels) # Check
clustering score
print("Silhouette score:", score)
```

```
Silhouette score: 0.07009307629448445
```

```
# Basic analysis of content types and countries
print(df["type"].value_counts())
print(df["country"].value_counts().head(10))
```

```
type
Movie      5377
TV Show    2410
Name: count, dtype: int64
country
United States     2555
India              923
United Kingdom     397
Japan              226
South Korea        183
Canada             177
Spain              134
France             115
Egypt              101
Turkey             100
Name: count, dtype: int64
```

```
similarity_matrix = cosine_similarity(reduced_features)

title_col = "title"  # Check how similar titles are

# Map title to index
title_to_index = {t.lower(): i for i, t in
enumerate(df[title_col].astype(str))}

def recommend_similar(title, top_n=10):
    t = title.lower()
    if t not in title_to_index:
        print("Title not found.")
        return []

    idx = title_to_index[t]
    sim_scores = similarity_matrix[idx]

    # Get similar results but remove same item
    similar_indices = np.argsort(sim_scores)[::-1]  # descending
    similar_indices = [i for i in similar_indices if i != idx][:top_n]

    return df.iloc[similar_indices][[title_col, "cluster"]]

# Example usage:
print(recommend_similar("3 Idiots", top_n=10))
```

```
                     title  cluster
1758       Dil Dhadakne Do        1
4872                    PK        1
1757        Dil Chahta Hai        1
4485              No Entry        1
7371              Upstarts        1
5097       Rang De Basanti        1
3362                Khushi        1
4961  Prem Ratan Dhan Payo        1
3128       Jatt James Bond        1
3287          Kai Po Che!       13
```

```
print(recommend_similar("PK", top_n=10))
print(recommend_similar("Dangal", top_n=10))
```

```
                                    title  cluster
271                             Aarakshan        1
4276                      Mumbai Cha Raja        1
7371                             Upstarts        1
4721                      Paan Singh Tomar        1
4634  Once Upon a Time in Mumbai Dobaara!        1
663                               Baazaar        1
2006               English Babu Desi Mem        1
100                              3 Idiots        1
5579                              Shorgul       13
1758                      Dil Dhadakne Do        1
                            title  cluster
7157                      Torbaaz        1
1739  Dhobi Ghat (Mumbai Diaries)        1
4721             Paan Singh Tomar        1
3812        Lucky: No Time for Love        1
4276              Mumbai Cha Raja        1
2859           Hum Aapke Hain Koun        1
1896                    Duplicate        1
5722                       Soorma       13
3477                    Kya Kehna        1
1257                      Chaahat        1
```