Huấn Luyện Mô Hình Gợi Ý Phim

Tổng Quan Dự Án

Script này xây dựng và huấn luyện mô hình gợi ý phim sử dụng thư viện LightFM. Dữ liệu phim và thông tin diễn viên được xử lý, mô hình được huấn luyện, ma trận tương đồng được tính toán và kết quả được lưu trữ để sử dụng trong ứng dụng chính.

Giải Thích Quy Trình

1 Tải Dữ Liệu

```
def load_data():
    movies_df = pd.read_csv('./data/tmdb_5000_movies.csv')
    credits_df = pd.read_csv('./data/tmdb_5000_credits.csv')
    return movies_df, credits_df
```

- **Mục Đích:** Tải dữ liệu phim và thông tin diễn viên.
- **Tệp Đầu Vào:** `tmdb_5000_movies.csv` và `tmdb_5000_credits.csv`
- **Đầu Ra:** DataFrames `movies_df` và `credits_df`

---

2 Chuẩn Bị Dữ Liệu

```
def prepare_data(movies_df, credits_df):
    user_ids = movies_df['id'].unique().tolist()[:]
    item_ids = credits_df['movie_id'].unique().tolist()[:]

    dataset = Dataset()
    dataset.fit(users=user_ids, items=item_ids)

    interactions, _ = dataset.build_interactions([(user, item) for user, item in zip(user_id
    return dataset, interactions
```

- **Mục Đích:** Chuẩn bị dữ liệu tương tác giữa người dùng và phim.
- **Bước Chính:**
  - Trích xuất `user_ids` và `item_ids` duy nhất
  - Tạo `LightFM Dataset`
  - Xây dựng ma trận tương tác
- **Đầu Ra:** Ma trận tương tác

---

3 Huấn Luyện Mô Hình

```
def train_model(interactions):
    model = LightFM(loss='warp')
    model.fit(interactions, epochs=10, num_threads=2)
    return model
```

- **Mục Đích:** Huấn luyện mô hình gợi ý.
- **Thuật Toán:** WARP (Weighted Approximate-Rank Pairwise)
- **Tham Số:**
  - `epochs`: 10 vòng lặp huấn luyện
  - `num_threads`: 2 luồng xử lý
- **Đầu Ra:** Mô hình LightFM đã huấn luyện

---

4 Tính Ma Trận Tương Đồng

```
def compute_similarity_matrix(model, interactions):
    item_embeddings = model.get_item_representations()[1]
    similarity_matrix = np.dot(item_embeddings, item_embeddings.T)
    return similarity_matrix
```

- **Mục Đích:** Tính toán độ tương đồng giữa các bộ phim.
- **Bước Chính:**
  - Trích xuất embedding của phim
  - Tính toán độ tương đồng bằng phép nhân ma trận
- **Đầu Ra:** Ma trận tương đồng

---

5 Lưu Dữ Liệu Mô Hình

```
def save_model_data(movies_df, similarity_matrix):
    with open('model/movie_list.pkl', 'wb') as f:
        pickle.dump(movies_df[['id', 'title']], f)
    with open('model/similarity.pkl', 'wb') as f:
        pickle.dump(similarity_matrix, f)
```

- **Mục Đích:** Lưu dữ liệu phim và ma trận tương đồng.
- **Tệp Được Tạo:**
  - `movie_list.pkl`: Danh sách ID và tiêu đề phim
  - `similarity.pkl`: Ma trận tương đồng

---

6 Tạo Dữ Liệu

```
def generate_data():
    movies_df, credits_df = load_data()
```

```
    dataset, interactions = prepare_data(movies_df, credits_df)
    model = train_model(interactions)
    similarity_matrix = compute_similarity_matrix(model, interactions)
    save_model_data(movies_df, similarity_matrix)

generate_data()
```

- **Mục Đích:** Điều phối quy trình làm việc tổng thể.
- **Bước Chính:**
  1. Tải dữ liệu
  2. Chuẩn bị tập dữ liệu
  3. Huấn luyện mô hình
  4. Tính toán ma trận tương đồng
  5. Lưu kết quả

Công Nghệ Sử Dụng

- **Python**
- **Pandas**
- **NumPy**
- **LightFM**
- **Pickle**

Tệp Đầu Ra

- `movie_list.pkl`: Chứa ID và tiêu đề phim.
- `similarity.pkl`: Ma trận tương đồng giữa các bộ phim.

Liên Hệ

- Mọi thắc mắc, vui lòng liên hệ: `phutc04@gmail.com`

Huấn Luyện Mô Hình Hoàn Tất!