iLike

Università degli Studi di Salerno

Documentazione progetto
Fondamenti di Intelligenza Artificiale
a.a. 2022/2023
prof. Fabio Palomba

Autore	*Matricola*
Costante Luigina	0512110457
Lo Conte Simona	0512110922
Napolillo Marta	0512109836

Indice

1 Introduzione **2**

2 Business Understanding **3**
2.1 Obiettivi di business 3
2.2 PEAS 3
2.3 Proprietà dell'ambiente 3
2.4 Analisi del problema 4

3 Data Understanding **5**
3.1 Acquisizione dei dataset 5
3.2 Analisi dei dataset 5

4 Data Preparation **6**
4.1 Data Cleaning 6
4.2 Feature Scaling 6
4.3 Feature Selection 6
4.4 Data Balancing 6

5 Modeling **7**
5.1 Scelta dell'algoritmo da utilizzare 7
5.2 Fase di addestramento 7

6 Evaluation **8**
6.1 Elbow point 8
6.2 Silhouette coefficient 8
6.3 MoJo distance 8

7 Deployment **9**

8 Glossario **10**

1 Introduzione

Con l'avvento del digitale è in costante crescita il numero di persone - di ogni fascia di età - che scelgono di guardare un film nel proprio tempo libero. Di conseguenza, aumenta la voglia di scoprire sempre nuovi contenuti in base alle proprie preferenze e rimanere costantemente aggiornati sulle ultime novità. La maggior parte degli utenti, però, è spesso indecisa su quale film scegliere e passa gran parte del tempo a navigare tra i contenuti disponibili. A tal proposito *iLike*, oltre a realizzare una piattaforma unificata che consente di recensire contenuti, offre la possibilità di interagire con un Conversational Agent, la quale permette di visualizzare i film richiesti dall'utente, personalizzati sulla base delle sue preferenze. Il Conversational Agent è di fondamentale importanza poichè permette agli utenti indecisi di ricevere consigli personalizzati evitando di passare ore ed ore nella scelta di un film.

2 Business Understanding

2.1 Obiettivi di business

L'obiettivo principale di *iLike* è la realizzazione di un Conversational Agent, che permetterà all'utente di interagirvi per richiedere consigli su film da guardare. Lo scopo è quello di consentire una facile interazione degli utenti con la nostra applicazione, permettendo a utenti indecisi di ricevere consigli personalizzati su film che potrebbero essere di interesse personale, in base a quelli appartenenti alle liste personali create oppure in base ad un genere scelto dall'utente.

2.2 PEAS

Specifica PEAS dell'ambiente.

Performance	Capacità dell'agente di suggerire all'utente film che rispecchiano i suoi gusti.
Environment	L'ambiente in cui l'agente opera è rappresentato da iLike, un'applicazione in cui gli utenti possono scrivere recensioni ed esprimere preferenze sui contenuti che si trovano all'interno di essa.
Actuators	Risposta del Conversational Agent.
Sensors	Utterances (messaggi in linguaggio naturale dati in input al CA da un utente umano).

2.3 Proprietà dell'ambiente

L'ambiente possiede le seguenti proprietà:

- **Completamente osservabile**: l'agente ha accesso all'elenco dei contenuti presenti nell'applicazione e alle preferenze degli utenti in qualsiasi momento;
- **Stocastico**: lo stato dell'ambiente varia indipendentemente dall'azione intrapresa dall'agente;

- **Sequenziale**: le decisioni prese dall'agente dipendono dalle azioni passate dell'utente;
- **Statico**: nel momento in cui l'agente sta elaborando la sua decisione l'utente non può modificare le sue preferenze;
- **Discreto**: i suggerimenti dati dall'agente dipendono dalla combinazione di contenuti preferiti di cui l'utente dispone o da un genere stabilito ed esistono un numero limitato di possibili combinazioni;
- **Singolo-agente**: esiste un unico agente che opera nell'ambiente.

2.4 Analisi del problema

Il problema che l'agente intelligente dovrà risolvere consiste nel suggerire film da vedere in base ai contenuti presenti nei dataset dell'applicazione e soprattutto in merito alle preferenze espresse dagli utenti (in base ai contenuti delle liste personali o ad un genere scelto). Il problema in esame può essere risolto con un algoritmo di apprendimento in quanto consiste nel migliorare l'esecuzione di un task (T=fornire suggerimenti personalizzati) rispetto ad una misura di prestazione (P= numero di suggerimenti accettati dall'utente) e sulla base dell'esperienza (E= database di contenuti non etichettati). Inoltre l'algoritmo di apprendimento in questione è di tipo non supervisionato in quanto non si dispone di un database contenente dati già etichettati, bensì dovrà essere l'agente capace di apprendere il valore reale della variabile dipendente sulla base delle conoscenze di cui dispone. Nello specifico il problema in esame può essere risolto tramite l'utilizzo di un algoritmo di clustering. Una volta che l'utente ha espresso le sue preferenze riguardanti contenuti presenti nell'applicazione, l'algoritmo creerà, in base ad una misura di similarità (che verrà definita in seguito), dei cluster contenenti film dotati di un certo grado di omogeneità. Procederà quindi a consigliare nuovi film in base alla clusterizzazione effettuata.

I suggerimenti verranno dati solo qualora l'utente ne faccia richiesta ed il tutto avviene in maniera automatica tramite l'utilizzo di un Conversational Agent.

3 Data Understanding

3.1 Acquisizione dei dataset

Durante la scelta dai dati da fornire al machine learning le possibili scelte da seguire erano sostanzialmente due:

- Creare un dataset contenente gli utenti di iLike ed analizzare il loro comportamento, al fine di creare cluster di utenti i quali hanno preferenze similari;
- Cercare dataset con le informazioni relative ai film e creare cluster di film.

I problemi riscontrati sono:

- La disponibilità di dati era maggiore nei dataset già esistenti;
- Ogni utente ha gusti differenti, quindi la similarità tra utenti, rappresentata come il numero di contenuti uguali appartente alle proprie liste, può non essere sempre veritera;
- Individuare dataset con un numero ottimale di istanze e le giuste informazioni sui film richiede un'accurata analisi.

Al seguito di un trade-off tra le due alternative abbiamo preferito utilizzare dataset già esistenti relativi ai film, poichè la disponibilità di dati e la giusta similarità di elementi in un cluster agevola le prestazioni dell'algoritmo di machine learning.

3.2 Analisi dei dataset

Il dataset utilizzato riguardo i Film è reperibile sulla piattaforma Kaggle.

4 Data Preparation

In fase di Data Understanding abbiamo scelto l'utilizzo di un dataset già esistente, quindi è opportuno effettuare Data Preparation. Lo scopo di questa fase è pulire i dati al fine di passarli all'algoritmo di Machine Learning.

4.1 Data Cleaning

Lo scopo del Data Cleaning è gestire dati rumorosi e/o nulli. A tal proposito, è stata effettuata inizialmente un'analisi dei dati al fine di evidenziare dati rumorosi. Si è notato che le colonne 'voto_medio' e 'voti_totali' rispecchiano esattamente la media e la somma delle colonne 'voto_critica' e 'voto_pubblico'. Dunque abbiamo deciso di elimnare le colonne 'voto_critica' e 'voto_pubblico' e non eliminare le colonne 'voto_medio' e 'voti_totali', rimandando tale scelta nel Feature Selection, per valutare quale delle due ha una maggiore corelazione tra le altre variabili.

In seguito è riportata una tabella contenente il nome della colonna, il numero di elementi mancanti e la scelta di Data Imputation effettuata.

Nome Colonne	Numero Elementi	Scelta Data Imputation
genere	95	Eliminazione Riga
paese	11	Eliminazione Riga
registi	33	Eliminazione Riga
attori	2.027	Eliminazione Colonna
descrizione	1.449	Eliminazione Riga
note	21.628	Eliminazione Colonna

Le scelte sopra riportate sono state effettuate per evitare di eliminare un eccessivo numero di righe. Si fa eccezione per la colonna descrizione in quanto tale colonna è necessaria all'interno dell'applicazione iLike, per altre funzionalità non riguardanti il modulo di Intelligenza Artificiale.

4.2 Feature Scaling

Lo scopo del Feature Scaling è normalizzare i valori numeri del dataset, portandoli tutti nello stesso range. La tabella sotto riportata confronta il range iniziale e i range ottenuti prima con la tecnica del Min-Max normalizzation e in seguito con lo Z-Score Normalizzation.

Nome Colonne	Range Iniziale	Media Min-Max	Media Z-Score
erotismo	0-4	0.076504	-0.354508
tensione	0-5	0.188511	-0.354313
impegno	0-5	0.139378	-0.354420
ritmo	0-5	0.278423	-0.354193
humor	0-5	0.120133	-0.354415
voti_totali	1-1052	0.035554	-0.345445
voto_medio	1-10	0.531023	-0.353328
durata	41-924	0.067289	-0.333306
anno	1911-2023	0.741849	0.061447

4.3 Feature Selection

4.4 Data Balancing

5 Modeling

5.1 Scelta dell'algoritmo da utilizzare

5.2 Fase di addestramento

6 Evaluation

6.1 Elbow point

6.2 Silhouette coefficient

6.3 MoJo distance

7 Deployment

8 Glossario

Conversational Agent/CA	É un bot che interpreta e risponde alle dichiarazioni fatte dagli utenti in un linguaggio naturale, attraverso la generazione di una conversazione simil-umana.
Lista di contenuti	Sottoinsieme di contenuti offerti da iLike, scelti dagli utenti secondo i loro gusti e inseriti nelle proprie liste disponibili sul proprio profilo personale.
Cluster	Sottoinsieme di contenuti con caratteristiche simili.
Machine Learning	É la branca dell'Intelligenza Artificiale che include tutti gli algoritmi che possano imparare dai dati e sulla base di questi fare previsioni.
Data Imputation	É l'insieme di tecniche che possono stimare il valore di dati mancanti sulla base dei dati disponibili oppure mitigare il problema dei dati mancanti.