YourBook

Riferimento	
Versione	0.00.004
Data	18/02/2021
Destinatario	Prof. F. Palomba, Prof. F.Narducci
Presentato da	D'Urso Serena (0512105709)[IS - FIA], Petta Paolo (0512105013) [IS - FIA],
Approvato da	

Revision History

Data	Versione	Descrizione	Autori
18/02/2021	0.00.001	Creazione documenti	S.D'Urso
01/03/2021	0.00.002	Stesura punto 1 e 2	S.D'Urso
08/03/2021	0.00.003	Stesura punto 3	P.Petta
09/03/2021	0.00.004	Revisione	S.D'Urso, P.Petta

Contents

1 Introduzione del problema

Il sistema che si intende realizzare è un servizio che fornisce al website YourBook la gestione del suggerimento di nuovi titoli agli utenti registrati. Gli utenti presenti sulla piattaforma, i quali avranno inserito una votazione ai titoli già letti, otterranno una lista di titoli che possano apprezzare.

2 Raccolta dati - specifica dell'ambiente (PEAS)

Per la raccolta dati si è preferito usare un dataset disponibile in rete, che comprendeva informazioni quali:

Per i libri:

- ISBN;
- Titolo;
- Autore;
- Anno di pubblicazione;
- Editore;
- Immagine copertina (versione small, medium e large);

Per gli utenti:

- User id;
- Location;
- Età;

Per ratings:

- ISBN;
- User id;
- Votazione libro;

Poichè la capacità di elaborazione dei computer utilizzati per la realizzazione del progetto è limitata, si è andato ad utilizzare solo 1000 righe del dataset (ottenuto dopo varie elaborazioni) sulle oltre 200.000 presenti.

2.1 Studio dei dati

Analizzando i diagrammi ottenuti dalle prime elaborazioni sui dati, si è notato che le variabili prese in considerazione erano poco correlate tra di loro. Si è giunti alla conclusione che questo sia dovuto a una dose di inesperienza nella lavorazione dei dati e anche all'aver considerato solo una parte del dataset.

2.2 Analisi dei PEAS

L'acronimo PEAS descrive una rappresentazione schematica dell'ambiente e dell'agente. Questa descrizione è basata su quattro aspetti, quali Performance, Environment, Actuators e Sensors.
Nel progetto YourBook, preso in considerazione questi aspetti sono identificati in:

- ➔ **Performance** (misure di prestazione)**:** la misura di prestazione adottata si basa su quanto gli elementi (i libri) dell'insieme individuato dall'algoritmo vengano apprezzati da un utente;
- ➔ **Environment** (descrizione elementi dell'ambiente): l'ambiente è costituito da tutti i libri e le relative recensione fornite dall'utente. L'ambiente definito è: **dinamico**, in quanto il livello di apprezzamento per determinati autori o generi letterari cambia con le votazioni assegnate dall'utente; **sequenziale**, in quanto deve tener conto delle precedenti valutazioni per stabilire il grado di apprezzamento di un genere/autore per poter consigliarlo; **discreto**, gli input forniti dall'utente sono finiti; **deterministico**, poichè lo stato successivo dell'agente è determinato dallo stato corrente e dall'azione eseguita dall'agente; infine, è ad **agente singolo**, in quanto è necessario un solo agente che ha il compito di esplorare tutto l'ambiente;
- ➔ **Attuatori** (in che modo l'agente restituisce l'output): l'output dell'esecuzione viene restituito in una pagina .jsp appositamente creata con lo scopo di mostrare tutti i libri che l'agente crede che possano interessare all'utente;
- ➔ **Sensori** (sono la parte percettiva dell'agente)**:** l'input dell'agente è fornito tramite una pagine .jsp che permette all'utente di ricercare un titolo e fornirgli una valutazione da interi in un range(0 - 10).

3 Implementazione

Per l'implementazione del modulo di IA per il progetto YourBook si è optato per il linguaggio Python e come ambiente di sviluppo si è scelto CoLab e, in seguito, JupyterLab, visti problemi legati all'esecuzione degli script.
Si è fatto uso delle librerie (elencate le più importanti):

- pandas e numpy, per la gestione dei dati;
- sklearn, per la manipolazione del modello;
- seaborn e matplotlib, per la costruzione dei grafici;

3.1 Motivazioni scelte progettuali

Per l'implementazione dell'agente è stato scelto l'algoritmo è il K-means.

Il K-Means è un algoritmo di apprendimento non supervisionato che trova un numero fisso di cluster in un insieme di dati.

I cluster rappresentano i gruppi che dividono gli oggetti a seconda della presenza o meno di una certa somiglianza tra di loro.
Nella implementazione proposta i cluster vengono raggruppati prima per titoli che hanno una votazione uguale e numero di votazioni maggiore di 30. In sintesi, i cluster sono formati da libri con votazioni simili.

Quando si utilizza un algoritmo K-Means, per ogni cluster si definisce un centroide, ossia un punto al centro di un cluster. I centroidi sono scelti tra i libri con valutazione maggiore.

Si è scelto l'algoritmo k-means ha il vantaggio di essere abbastanza veloce, in quanto sono richiesti pochi calcoli e di conseguenza poco tempo di elaborazione computer, considerando che la sua applicazione è per un servizio web.

Il k-means ha come svantaggi che bisogna selezionare quanti gruppi si desidera visualizzare. Questo non è sempre banale in quanto non sempre è possibile farlo, soprattutto per problemi di complessità maggiore. Inoltre, K-means inizia con una scelta casuale di centroidi e pertanto può produrre risultati di clustering diversi su diverse sequenze dell'algoritmo, quindi i risultati potrebbero non essere ripetibili e mancare di coerenza.

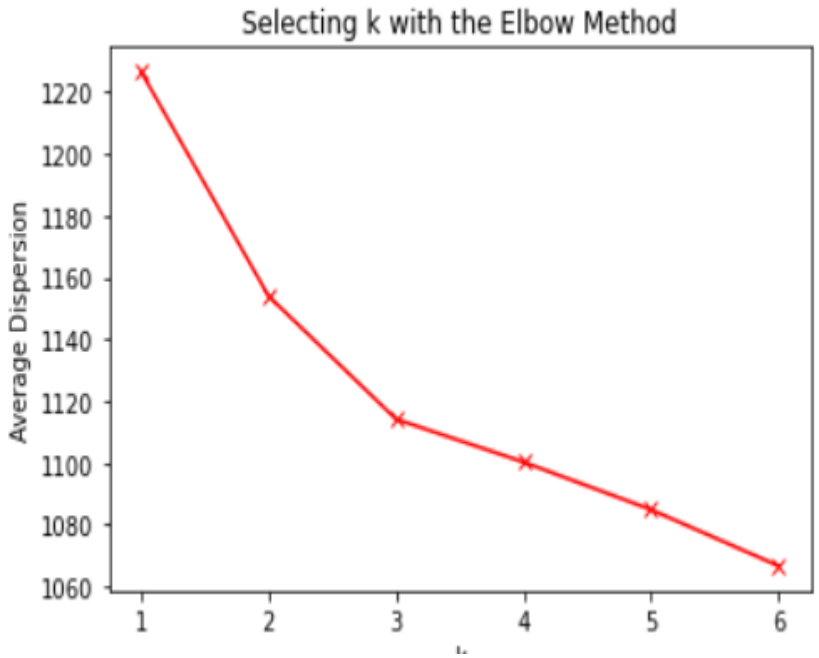
3.2 Indice di Silhouette

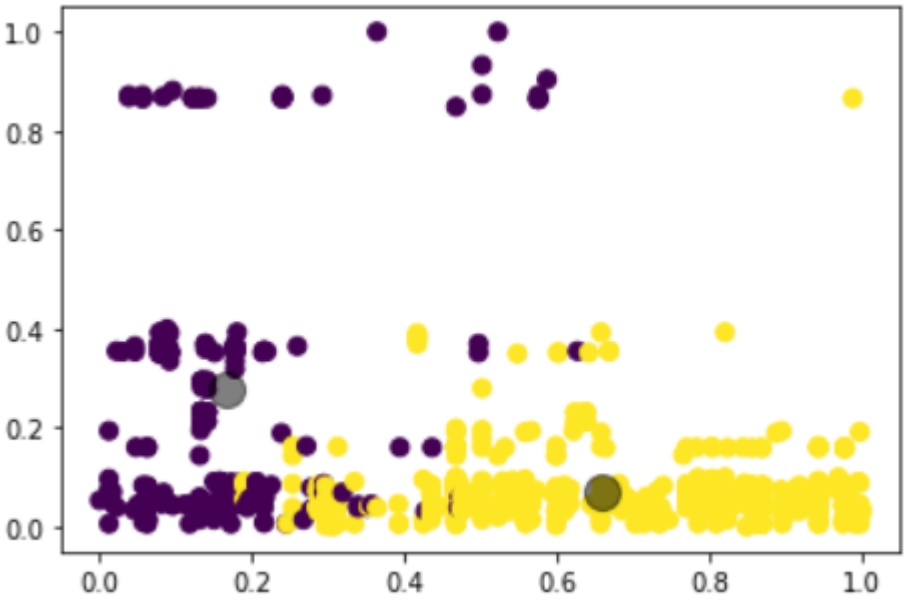
L'indice di silhouette permette di quantificare la bontà della disposizione dei dati nei cluster assegnati, quindi la distanza da qualsiasi altro cluster e quanto sono ben raggruppati nel cluster di appartenenza. L'indice oscilla tra [-1, 1], dove 1 indica che i dati fanno un giusto match con il proprio cluster, mentre un valore vicino allo 0 o addirittura negativo indica che la configurazione dei cluster può aver troppi o troppi pochi cluster.

Il valore dell'indice di silhouette ottenuto è molto basso (circa 0.07), sintomo di ciò che è appena stato spiegato, probabilmente.

3.3 Elbow point

La costruzione del grafico che rappresenta l'andamento del valore della somma degli errori quadratici al variare del numero di cluster, non ha risolto la questione relativa al numero di cluster da adottare. Infatti, a seconda delle esecuzioni veniva mostrato un grafico diverso, con un punto di elbow diverso. Si è optato infine per l'uso di 2 cluster, dopo vari tentativi, per ottenere un indice di silhouette più alto.




1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0