Università degli Studi di Palermo

Dipartimento di Ingegneria
Direttore: prof. Antonino Valenza

dipartimento di ingegneria unipa

CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA

Prova scritta del 6 giugno 2023

Si consideri la seguente Utility Matrix:

Utility Matrix	Il Gladiatore	I Guardiani della Galassia	Star Trek	Lo Hobbit: la desolazione di Smaug	Il Signore degli Anelli: il Ritorno del Re	Transformers: La Vendetta del Caduto	Spiderman 3	Avengers	Il Cavaliere Oscuro	Iron Man
John	5		3				2	2		3
Jack		4			4	1				
Mary				1				3	3	5
Sue	2		1		5		5		1	
George	3	4		3		2		1		
Travis			5		3		5			4
Elen	2			2			4	2	3	
Violet		2	3			4			2	3
Robin	2				3			4		2
Albert		2	4	2		3	3	4	4	

1. Suggerire i rating non specificati per gli utenti Violet e Sue sulla base dei migliori tre utenti più simili. Ai fini del calcolo, si utilizzino le medie dei rating di ogni utente su tutti i film e non solo su quelli comuni con gli altri. Calcolare i rating come media pesata, tramite i valori di correlazione, dei rating dei tre utenti migliori.

   punti ___/ 8

2. Suggerire i rating sui film ('Il Gladiatore', 'Avengers') per Violet e ('Star Trek', 'Iron Man') per Sue usando la item based similarity con i tre migliori item. Ai fini del calcolo, si utilizzi la distanza coseno tra i valori normalizzati di rating, ovvero: $\text{rating}_{u,i} = rating_{u,i} - avg(rating_u)$

   punti ___/ 8

3. Implementare una routine K-means per effettuare il clustering tra le righe di una Utility Matrix, in cui si gestiscano esplicitamente vettori in cui non tutte le dimensioni sono specificate (Suggerimento: usare `np.nan`)

   punti ___/ 10

4. Usando la routine sviluppata al punto 3. Implementare una classe `CollaborativeFiltering` che istanzi un modello in grado di accettare una utility matrix e un numero di cluster di utenti predeterminato al fine di creare i gruppi di controllo per i rating. Il metodo `fit` eseguirà il clustering, mentre il metodo `predict` accetterà un utente specificato come riga di una matrice di utilità e restituirà i rating medi degli utenti del cluster su tutti gli item non specificati dell'utente di test.

   punti ___/ 4

TOTALE: punti ___/ 30

Data: _____ Allievo: _____ Matricola: _____

Regole della prova scritta

Di seguito si riportano le regole da seguire e le caratteristiche della prova ai fini della valutazione:

1. La durata complessiva della prova è pari a due ore e prevede una serie di quesiti che approfondiscono diversi aspetti dello stesso problema: ognuno sarà libero di dedicare ad ogni quesito il tempo che vorrà.
2. La prova si svolge ***interamente*** al calcolatore.
3. L'ambiente di sviluppo predisposto è un environmet conda/mamba dotato di editor Spyder e dei seguenti pacchetti:
    a. Pandas
    b. Matplotlib
    c. Seaborn
    d. Numpy
    e. Scipy
    f. Scikit-learn
    g. Pyspark
    h. Keras
    i. Tensorflow
4. Ai fini dell'avvio dell'ambiente, aprire il prompt dei comandi e digitare i due seguenti comandi:
    ```
    $ mamba activate spyder-env
    $ spyder
    ```
5. Sarà consentito consegnare dopo la prima ora di prova.
6. Sarà necessario spegnere e consegnare i dispositivi mobili (smartphone, smartwatch e tablet) alla cattedra prima dello svolgimento della prova.
7. La navigazione internet dalle postazioni sarà bloccata, in generale, e consentita solo verso i siti di documentazione delle librerie ed il repository delle slide in pdf.
8. Il docente distribuirà copia digitale del compito ed eventuali data set direttamente dalla propria postazione ovvero tramite penna USB e allo stesso modo raccoglierà gli elaborati di programmazione.
9. Il candidato consegnerà comunque il presente foglio datato e con l'indicazione del nome e del numero di matricola.
10. Ai fini del calcolo del voto finale della prova, il *valore massimo* di ciascun quesito è riportato in calce allo stesso. La prova riceverà una valutazione pari alla somma dei voti riportati in ciascun quesito. Si precisa che il docente attribuirà ad ogni quesito una votazione *non binaria*, cioè non tutto il valore oppure 0, ma valuterà la correttezza formale dell'elaborato, il rigore metodologico dell'approccio teorico e l'originalità delle soluzioni proposte per attribuire una votazione nel range definito dal valor massimo del quesito.

Data: _____ Allievo: _____ Matricola: _____