Progetto
Metodi Informatici Gestione Aziendale
a.a.2023/2024

1 Descrizione progetto

Obiettivo generale del progetto: sviluppare diverse tipologie di Recommendation System (collaborative filtering e content based) utilizzando il set di dati indicato

Le attività finalizzate all'obiettivo generale si articoleranno in 3 livelli, così denominati:

- Livello base
- Livello intermedio
- Livello avanzato

Ogni livello di progetto, come indicato dal nome, corrisponde a un diverso grado di difficoltà e complessità, in base alla tipologia di analisi da applicare.

Il progetto può essere svolto singolarmente o da gruppi di massimo 2 persone per gli appelli di Giugno e Luglio 2024. Negli appelli successivi il progetto dovrà essere svolto singolarmente.

- *Progetti singoli*: requisito minimo è svolgere tutti i passi di analisi riportati nel "progetto base".
- *Progetti di gruppo* (max 2 persone): requisito minimo è svolgere tutti i passi di analisi riportati almeno nel "progetto intermedio".

Valutazione:

La tipologia di progetto scelta e il numero di componenti del gruppo (singolo o due), saranno tenuti in considerazione in fase di valutazione finale di progetto.

Tipologia	Votazione Progetto singolo	Votazione Progetto di gruppo
Base	Max. 25	Non applicabile
Intermedio	Max. 30	Max. 28
Avanzato	Max. 30L	Max. 30L

Il testo del progetto è valido per tutti gli appelli dell'anno accademico 2023-2024.

2 SET DI DATI

Il set di dati da analizzare è relativo alle recensioni di prodotti Amazon raggruppate per categorie merceologiche, quali:

"All_Beauty, Amazon_Fashion, Appliances, Arts_Crafts_and_Sewing, Automotive, Baby_Products, Beauty_and_Personal_Care, Books, CDs_and_Vinyl, Cell_Phones_and_Accessories, Clothing_Shoes_and_Jewelry, Digital_Music, Electronics, Gift_Cards, Grocery_and_Gourmet_Food, Handmade_Products, Health_and_Household, Health_and_Personal_Care, Home_and_Kitchen, Industrial_and_Scientific, Kindle_Store, Magazine_Subscriptions, Movies_and_TV, Musical_Instruments, Office_Products, Patio_Lawn_and_Garden, Pet_Supplies, Software, Sports_and_Outdoors, Subscription_Boxes, Tools_and_Home_Improvement, Toys_and_Games, Video_Games, Unknown"

Link alla pagina del dataset:
https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

Vista la dimensione dell'intero dataset, nella pagina trovate anche dei sottoinsiemi di dati raggruppati in categorie merceologiche (https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023#grouped-by-category). Per le diverse tipologie di progetti deve essere analizzata una di queste categorie.

In particolare, per ogni categoria vi sono due dataset:

- **user reviews**: con le informazioni sulle recensioni di prodotti

Field	Type	Explanation
rating	float	**Rating of the product (from 1.0 to 5.0).**
title	str	Title of the user review.
text	str	Text body of the user review.
images	list	Images that users post after they have received the product. Each image has different sizes (small, medium, large), represented by the small_image_url, medium_image_url, and large_image_url respectively.
asin	str	**ID of the product.**
parent_asin	str	Parent ID of the product. Note: Products with different colors, styles, sizes usually belong to the same parent ID. The "asin" in previous Amazon datasets is actually parent ID.
user_id	str	**ID of the reviewer**
timestamp	int	Time of the review (unix time)
verified_purchase	bool	User purchase verification
helpful_vote	int	Helpful votes of the review

- **Item Metadata**: con le informazioni sui prodotti

Field	Type	Explanation
main_category	str	Main category (i.e., domain) of the product.
title	str	**Name of the product.**
average_rating	float	Rating of the product shown on the product page.
rating_number	int	Number of ratings in the product.
features	list	Bullet-point format features of the product.
description	list	**Description of the product.**
price	float	Price in US dollars (at time of crawling).
images	list	Images of the product. Each image has different sizes (thumb, large, hi_res). The "variant" field shows the position of image.
videos	list	Videos of the product including title and url.
store	str	Store name of the product.
categories	list	Hierarchical categories of the product.
details	dict	Product details, including materials, brand, sizes, etc.
parent_asin	str	Parent ID of the product.
bought_together	list	Recommended bundles from the websites.

3 PROGETTO BASE

Obiettivo progetto base: sviluppo di un sistema di raccomandazione basato su collaborative filtering partendo da un set di dati di review di prodotti Amazon.

3.1 Dati da analizzare progetto base:

Per il progetto base è sufficiente utilizzare una sola categoria di prodotti, prendendo in considerazione solo il file relativo alle recensioni (*user reviews*).

I rating contenuti dovranno essere utilizzati per lo sviluppo del sistema di raccomandazione collaborative filtering.

3.2 Passi di analisi progetto base

Di seguito i principali step da eseguire:

1. Analisi Esplorativa (statistiche descrittive, analisi correlazione)
2. Identificazione della configurazione ottimale dell'algoritmo K-NN per la predizione dei rating. In questo punto dovranno quindi essere testate le diverse combinazioni: similarità, valore di K, user/item based. Tramite le diverse metriche di performance (MSE e RMSE) individuare di conseguenza la configurazione ottimale.
3. Filling della matrice di rating con la configurazione ottimale
4. Segmentazione degli utenti in base alle preferenze: algoritmo di clustering K-MEANS con cosine similarity.
5. Creazione per ogni utente della lista degli *n* items (top k items) da consigliare (es. considerando il rating predetto).
6. Filling della matrice di rating attraverso Matrix Factorization in aggiunta a K-NN e confronto dei risultati ottenuti in termini di MSE e RMSE.

Librerie suggerite: Surprise e Scikit-Learn

4 PROGETTO INTERMEDIO:

Obiettivo progetto intermedio: sviluppo di un sistema di raccomandazione collaborative filtering (progetto base) e content based, partendo da un set di dati di review di prodotti Amazon.

4.1 Dati da analizzare progetto intermedio:

Per il progetto intermedio è sufficiente utilizzare una sola categoria di prodotti, prendendo in considerazione sia il file relativo alle recensioni (*user reviews*) che quello relativo ai prodotti (*items metadata*).

Per lo sviluppo del sistema di raccomandazione andranno utilizzati almeno i campi *title* e *description* del file *items metadata*.

4.2 Passi di analisi progetto intermedio

Per questa tipologia di progetto devono essere eseguiti tutti i passi del progetto base (sviluppo del sistema di raccomandazione collaborative filtering) e i seguenti:

1. Processamento degli attributi testuali dei diversi prodotti (almeno i campi *title* e *description*) con le tecniche di Natural Language Processing viste in laboratorio.
2. Embedding dei campi con una tecnica basata sulla frequenza (bag-of-words o TFIDF) e una tecnica neurale (transformers).
3. Effettuare la predizione dei rating attraverso l'algoritmo K-NN per ogni utente usando gli embedding ottenuti con le due tecniche del punto 2.
4. Valutazione critica dei risultati ottenuti con le due diverse tecniche di embedding
5. Valutazione critica dei risultati ottenuti con il sistema di raccomandazione collaborative filtering (progetto base) e quello content-based (punto 1-3)

Librerie suggerite Scikit-Learn, NLTK, Hugging-face

5 PROGETTO AVANZATO

Obiettivo: sviluppare un sistema di raccomandazione collaborative filtering (progetto base), content based (progetto intermedio) ed effettuare una sentiment analysis sulle recensioni dei prodotti.

5.1 Dati da analizzare progetto avanzato

Per il progetto avanzato è sufficiente utilizzare una sola categoria di prodotti, prendendo in considerazione il file relativo alle recensioni (*user reviews*).

5.2 Passi di analisi progetto avanzato

Per questa tipologia di progetto devono essere eseguiti tutti i passi del progetto base (sviluppo del sistema di raccomandazione collaborative filtering), intermedio (sviluppo del sistema di raccomandazione content based) e i seguenti:

1. Processamento degli attributi testuali relativi alle review dei diversi prodotti (almeno i campi *title* e *text* del file *user reviews*) con tecniche di Natural Language Processing.
2. Embedding dei campi con una tecnica basata sulla frequenza (bag-of-words o TFIDF) e una tecnica neurale (transformers).
3. Effettuare la predizione del sentiment (rating 1-2: sentiment negativo, rating 3: sentiment neutro, rating 4-5: sentiment positivo) utilizzando gli algoritmi di classificazione di Scikit-Learn (quelli visti in laboratorio).

Librerie suggerite : Scikit-Learn, NLTK, hugging-face

6 ORGANIZZAZIONE DEI RISULTATI, REPORT FINALE E PRESENTAZIONE

Una componente importante della valutazione del progetto è basata su come viene rappresentata l'organizzazione dei risultati, la loro rappresentazione e valutazione.

Ogni studente/gruppo di studenti deve produrre:

- un **report** strutturato come segue:
  - Un breve riassunto (executive summary) con i principali obiettivi e risultati ottenuti (max 1 pagina).
  - Un'introduzione al problema (descrizione dei dati, obiettivi dell'analisi e risultati dell'analisi esplorativa) (da 5 a 10 pagine).
  - Diverse sezioni che riassumono i risultati dei diversi step del progetto (raggruppati per step) (da 15 a 20 pagine).
  - Conclusioni e interpretazione sintetica dei risultati (max 1 pagina).
- una **presentazione** per la discussione d'esame
  - Durata 10 minuti.
  - Max 10 slides che riprendono i punti principali del report.

Si ricorda che l'organizzazione dei risultati sarà un elemento integrante della valutazione finale.

Modalità di consegna

Ogni studente/gruppo di studenti deve inviare a Ilaria Giordani (ilaria.giordani@unimib.it) seguendo le scadenze su Moodle il report prodotto e il codice Python. La presentazione sarà utilizzata in fase di discussione esame e non deve essere inviata al docente.