Progetto
Metodi Informatici Gestione Aziendale
a.a.2022/2023

1 Descrizione progetto

Obiettivo del progetto: sviluppare un Recommendation System utilizzando i diversi set di dati indicati

Di seguito verrà riportata la descrizione di tre tipologie di progetti, così denominati:

- Progetto base
- Progetto intermedio
- Progetto avanzato

Ogni tipologia di progetto, come indicato dal nome, corrisponde a un diverso livello di difficoltà e complessità, sia in base al tipo di dati da elaborare che alla tipologia di analisi da applicare.

Il progetto può essere svolto singolarmente o da gruppi di massimo 2 persone per gli appelli di Gennaio e Febbraio 2023. 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	**Progetto singolo**	**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 2022-2023.

1.1 Informazioni di base sui set di dati da utilizzare

I dataset da analizzare fanno riferimento al sito "*Recommender Systems and Personalization Datasets*".

Alcuni sono in formato json, si consiglia di storicizzarli in un dataframe Pandas per poi utilizzarli per le successive analisi. Di seguito un semplice script per effettuare questo:

```
import pandas as pd
import gzip

def parse(path):
  g = gzip.open(path, 'rb')
  for l in g:
    yield eval(l)

def getDF(path):
  i = 0
  df = {}
  for d in parse(path):
    df[i] = d
    i += 1
  return pd.DataFrame.from_dict(df, orient='index')

df = getDF('reviews_Video_Games.json.gz')
```

2 PROGETTO BASE:

2.1 Dati da analizzare:

Dataset Name	User	Items	Link
Behance Community Art	63K	178k	https://cseweb.ucsd.edu/~jmcauley/datasets.html#behance
Clothing Fit Data (ModCloth)	47K	1k	https://cseweb.ucsd.edu/~jmcauley/datasets.html#clothing_fit
Marketing Bias (ModCloth)	44K	1k	https://cseweb.ucsd.edu/~jmcauley/datasets.html#market_bias

Note: Per i dataset Clothing Fit Data (ModCloth) e Marketing Bias (ModCloth) si ricorda di selezionare il dataset riguardante ModCloth, tralasciando gli altri. Per completezza sotto vengono riportati i link diretti ai set di dati interessati:

- Clothing Fit Data: https://jmcauley.ucsd.edu/data/modcloth/modcloth_final_data.json.gz
- Marketing Bias: https://github.com/MengtingWan/marketBias/raw/master/data/df_modcloth.csv

2.2 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: distanza, 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. Creare per ogni utente la lista degli *n* 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

3 PROGETTO INTERMEDIO:

3.1 Dati da analizzare progetto intermedio:

Dataset Name	User	Items	Link
Clothing Fit Data (Renttherunway)	105k	5.8K	https://cseweb.ucsd.edu/~jmcauley/datasets.html#clothing_fit
Marketing Bias (Amazon Electronics)	1.1M	9K	https://cseweb.ucsd.edu/~jmcauley/datasets.html#market_bias
Multi-aspect Reviews (BeerAdvocate)	33k	66K	https://cseweb.ucsd.edu/~jmcauley/datasets.html#multi_aspect
Multi-aspect Reviews (RateBeer)	40k	110K	https://cseweb.ucsd.edu/~jmcauley/datasets.html#multi_aspect

Note:

Dal momento che ogni dataset presenta diverse versioni, di seguito viene indicata quella da utilizzare con il relativo link per il download diretto:

- Dataset Clothing Fit Data: selezionare il set di dati *RentTheRunway* (https://jmcauley.ucsd.edu/data/renttherunway/renttherunway_final_data.json.gz)
- Dataset Marketing Bias : selezionare il set di dati Amazon Eletronics (https://github.com/MengtingWan/marketBias/raw/master/data/df_electronics.csv)
- Dataset Multi-aspect Reviews: selezionare il set di dati BeerAdvocate (https://jmcauley.ucsd.edu/data/beer/beeradvocate.json.gz) oppure il set RateBeer (https://jmcauley.ucsd.edu/data/beer/ratebeer.json.gz)

3.2 Passi di analisi progetto intermedio

Per questa tipologia di progetto devono essere eseguiti tutti i passi del progetto base e i seguenti:

1. Formulazione del problema multi-obiettivo (accuratezza, novelty e coverage) per creare la lista degli *n* items da raccomandare per ogni utente.
2. Risoluzione del problema attraverso gli algoritmi MOEA/D e NSGA-II della libreria Pymoo.
3. Confronto dei risultati ottenuti dai due algoritmi attraverso l'analisi delle metriche di performance quali Hypervolume, C-metric.
4. Visualizzazione delle frontiere Paretiane.

Librerie suggerite in aggiunta alle precedenti: Pymoo.

4 PROGETTO AVANZATO:

4.1 Dati da analizzare:

Dataset Name	User	Items	Link
Twitch	100k	162k	https://cseweb.ucsd.edu/~jmcauley/datasets.html#twitch
Food.com	226k	232K	https://cseweb.ucsd.edu/~jmcauley/datasets.html#foodcom

Per questa tipologia di progetto l'individuazione del dataset dipende dalla scelta della tipologia di analisi da effettuare, così come dettagliato nella sezione successiva.

4.2 Tipologie di analisi per progetto avanzato

In questo tipo di progetto, vengono proposte due tipologie alternative di analisi con i relativi set di dati.

4.2.1 Tipologia A

Eseguire tutti i passi del progetto base ed intermedio sul dataset proposto di seguito. Tenere in considerazione i suggerimenti riportati.

Dataset:

Twitch (100K) https://cseweb.ucsd.edu/~jmcauley/datasets.html#twitch. Un dataset di utenti che guardano contenuti in streaming su Twitch. Contiene gli streamer e i dati degli utenti connessi di 43 giorni.

Suggerimenti:

1- Considerare il "Watch Time" come rating (Time stop – time start).
2- Raccomandare streamers a utenti.

4.2.2 Tipologia B

Eseguire tutti i passi del progetto base ed intermedio sul dataset sotto proposto di seguito. In aggiunta eseguire le analisi riportate sotto.

Dataset:

Food.com Recipe & Review Data
(https://cseweb.ucsd.edu/~jmcauley/datasets.html#foodcom). Il dataset contiene dettagli di ricette e recensioni di Food.com. Contiene i seguenti dati:

- **Review**:
  - Ratings
  - Reviews (non deve essere utilizzato perché campo testuale)
- **Receipt**:
  - Recipe Name (non deve essere utilizzato perché campo testuale)
  - Recipe Description (non deve essere utilizzato perché campo testuale)
  - Recipe Ingredients (non deve essere utilizzato perché campo testuale)

- N. ingredients
- Recipe Directions (non deve essere utilizzato perché campo testuale)
- N.steps
- Recipe Categories (Tags)
- Recipe Nutrition Information

Analisi da sviluppare:

- Eseguire tutte le analisi fatte in precedenza (progetto base + intermedio) in termini “content based”, come visto a esercitazione.

5 ORGANIZZAZIONE DEI RISULTATI, REPORT FINALE E PRESENTAZIONE

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.