Nicolò Giovannacci

**EPICODE – DATA ANALYST 02-23 PT**

**Esercitazione M1-D3**

***Open Data***

***Consegna***

***Parte 1 - Individuare almeno 10 fonti dati diverse e, per ogni fonte dati, individuare almeno 5 dataset riportando:***

* *Classificazione 5 stelle*
* *I metadati e la loro classificazione*
* *Il tipo di file ed eventuali alternative disponibili con numero maggiore di stelle*

*Suggerimenti*

*per la metrica fate riferimento a https://5stardata.info/en/*

*esempi di sorgenti dati Kaggle, Wikipedia, GitHub, qualsiasi portale di una pubblica amministrazione*

*esempio di fonte dati (potete farlo per il comune che preferite o per più comuni ovviamente): https://opendata.comune.bologna.it/pages/home/*

*esempio di dataset estratto da questa fonte dati (che ne ha più di 600...):*

*https://opendata.comune.bologna.it/explore/dataset/rilevazione-flusso-veicoli-tramite-spire-anno-2023/table/?disjunctive.codice\_spira&disjunctive.tipologia&disjunctive.nome\_via&disjunctive.stato&sort=data (uri alla sorgente, dato strutturato, scaricabile, accessibile, integrabile, in formato non proprietario e linkato)*

*il portale open data di Bologna è fatto davvero bene*

**Parte 2 - Produci tre dataset con argomento a piacere (3 argomenti diversi) nel formato CSV.**

Due dei dataset prodotti dovranno essere composti obbligatoriamente da un header, tutti i file dovranno avere almeno 10 righe. Inoltre, per campi a propria scelta sarà necessario (a patto che abbiano senso) inserire almeno una virgola (punto e virgola nel caso del punto e virgola come separatore), più caratteri di nuova linea, valori vuoti, valori nulli, i doppi apici, i singoli apici. Infine, in un file bisogna utilizzare la virgola come separatore, in un altro file il punto e virgola e nel terzo file la tabulazione. Carica i 3 file prodotti *utilizzando uno strumento tra Excel, VisulaStudio, Notepad, Notepad++*

**Svolgimento**

Secondo quanto comunicato da Francesco nel corso della esercitazione del 12/04/2023 si procede svolgendo l’esercizio riportando 5 dataset anziché 10.

Si prende come riferimento la classifica delle prime 20 città del mondo per numero di abitanti. Si riportano di seguito 5 differenti Dataset di questa classifica, uno per ogni categoria di stelle, illustrandone successivamente le caratteristiche che ne riprendono la definizione.

**DATASET 1 (1 STELLA)**

Immagine che contiene tavolo

Descrizione generata automaticamente

Tale dataset è costituito da una immagine, ovvero un dato non strutturato, non modificabile e che non può essere oggetto di elaborazione.

**DATASET 2 (2 STELLE)**



Tale Dataset è strutturato, modificabile tramite un software proprietario e può essere convertibile in formato aperto (es. CSV). Per semplicità qui si riporta l’immagine, ma il dataset si riferisce ad una tabella Excel che sarà allegata al suddetto file word.

**DATASET 3 (3 STELLE)**

Immagine che contiene testo

Descrizione generata automaticamente

Tale Dataset è strutturato e codificato in un formato aperto. Si può elaborare senza la necessità di possedere un software proprietario. Per semplicità si riporta l’immagine, ma il dataset si riferisce ad un file CSV aperto con Notepad++ che sarà allegato al suddetto file word.

**DATASET 4 (4 STELLE)**

Immagine che contiene tavolo

Descrizione generata automaticamente

<https://it.wikipedia.org/wiki/Citt%C3%A0_del_mondo_per_popolazione>

Tale dataset è condiviso sul web e dotato di URI (riportato sopra) che costituisce anche l’URL, il quale rende il dataset accessibile online. Anch’esso è un dataset strutturato codificato in un formato aperto. Per semplicità si riporta l’immagine, ma il dataset NON è costituito dell’immagine, bensì dalla tabella della pagine web visualizzabile tramite l’URL.

**DATASET 5 (5 STELLE)**

Immagine che contiene tavolo

Descrizione generata automaticamente

<https://it.wikipedia.org/wiki/Citt%C3%A0_del_mondo_per_popolazione>

Tale dataset ha la stesse caratteristiche del precedente con l’aggiunta di collegamenti ad altri dataset: cliccando su ognuna delle città infatti, si visualizza la pagina wikipedia relativa alla città stessa, che in questo caso sarà un dataset semi-strutturato (combinazione di dati strutturati e non)

Per quanto concerne i metadati essi sono informazioni strutturate che descrivono, spiegano e rendono più semplice il recupero, l’uso e la gestione delle risorse informative a patto che siano interpretabili da un calcolatore. Un altro contenuto tipico dei metadati può essere la fonte o l'autore dell'insieme di dati descritto, oppure le modalità d'accesso con le eventuali limitazioni.

Per questo esercizio si decide di riportare come metadato la fonte dell’insieme dei dati:

**Fonte**: esa.un.org, [*World Urbanization Prospects: The 2007 Revision Population*, Database](https://web.archive.org/web/20100822053414/http:/esa.un.org/unup/index.asp?panel=6) URL consultato il 26 luglio 2010 (archiviato dall'url originale il 22 agosto 2010)

Per ogni dataset ne è stato indicato il tipo e implicitamente, ne sono state indicate le successive versioni corrispondenti al dataset con più stelle, essendo riferite alla stessa fonte.

**Parte 2**

In questo primo CSV si riporta un elenco dei primi 10 film in ordine decrescente in termini di Oscar Vinti. Procedo con non indicare il genere (dato mancante), per i film per i quali non ne sono a conoscenza

Immagine che contiene testo

Descrizione generata automaticamente

In questo secondo CSV si riportano le frasi di celebri di alcuni film con il rispettivo anno di uscita nelle sale. Si utilizzano gli apici per riportare titoli contenenti le virgole (le quali costituiscono il separatore) e si utilizza il carattere “new line” per poter andare a capo nelle frasi.

Text

Description automatically generated

In questo terzo CSV si riportano alcuni messaggi di errore indicando il relativo identificativo numerico. Si utilizzano i doppi apici per riportare il testo tra apici. Si utilizza il TAB come separatore.

Graphical user interface, text, application

Description automatically generated