Laboratorio avanzato di informatica umanistica 23/24 (Modulo 2)

Relazione progetto finale

Indice dei contenuti

Musei dell'Emilia-Romagna

N. matricola: VR475748

Obiettivi

L'obiettivo principale di questo progetto è quello di fare un'analisi dei dati di visita dei musei dell'Emilia-Romagna. In particolare, ci si interessa a come cambia il numero dei visitatori dei musei negli anni (tra il 2018 e il 2021); si analizzano prima tutti i musei indistintamente e poi separati in base alla località in cui si trovano. Ci si concentra in seguito sul numero dei visitatori per luogo (senza quindi considerare la distinzione tra i musei) e sul numero dei musei stessi in ogni località.

Svolgimento

Acquisizione

Apro Wikidata e uso il query service. Questa è la query che ho scritto:

```
SELECT ?museo ?nome ?luogo ?nomeLuogo ?coordinate ?anno ?visitatori
#Uso SELECT e non SELECT DISTINCT perché altrimenti non avrei la possibilità di vedere come cambia il numero dei visitatori dello stesso museo nel corso degli
#anni.
WHERE {
  ?museo wdt:P31 wd:Q33506.
  #Istanza (wdt:P31) appartenente alla categoria dei musei (wd:Q33506), quindi una struttura espositiva di oggetti.
  ?museo rdfs:label ?nome.
  #Recupero l'etichetta dell'elemento museo e la chiamo "nome".
  FILTER((LANG(?nome)) = "it")
 #Applico un filtro linguistico, voglio che tutti i nomi dei musei siano in lingua italiana. In questo modo evito di ottenere nei risultati dei doppioni,
 #infatti lo stesso museo può essere memorizzato più volte in lingue diverse.
  ?museo wdt:P131 ?luogo;
  #Wdt.P131 indica l'unità amministrativa in cui è situato il singolo museo, do un nome a questa proprietà ("luogo") in modo da poterlo inserire nel
  #SELCT e vederlo nei risultati
    (wdt:P131*) wd:Q1263.
  #Indico che il museo si deve trovare in Emilia Romagna (wd:Q1263). Inserisco * dopo wdt:P131 perché garantisce la ricorsività. Cercando l'unità
  #amministrativa in cui si trova un museo, infatti, ottengo un comune o una frazione, non troverei quindi alcun risultato per l'Emilia-Romagna.
  #L'asterisco invece fa sì che si proceda a ritroso, partendo dal comune/frazione per arrivare prima alla provincia ed infine alla regione in cui
  #il museo è ubicato.
  ?luogo rdfs:label ?nomeLuogo.
  #Recupero l'etichetta dell'elemento luogo e la chiamo nomeLuogo
  FILTER((LANG(?nomeLuogo)) = "it")
  #Applico un filtro linguistico, voglio che tutti i nomi dei luoghi siano in lingua italiana. In questo modo evito di ottenere nei risultati dei doppioni,
  #infatti lo stesso luogo può essere memorizzato più volte in lingue diverse.
  OPTIONAL { ?museo wdt:P625 ?coordinate. }
  #Se disponibili ottengo anche le coordinate del museo (wdt:P625 è una proprietà che indica le coordinate).
  OPTIONAL {
    ?museo p:P1174 ?nodointermedio.
    ?nodointermedio pq:P585 ?anno;
      ps:P1174 ?visitatori.}
    #Se disponibili voglio anche vedere i visitatori annui e l'anno di riferimento di ogni singolo museo. Con p:P1174 indico i visitatori annui,
    #chiamo questo proprietà nodointermedio e la metto in relazione  con una data (pq:P585), proprietà che chiamo "anno".
    #Infine, richiamo la proprietà dei visitatori annui come "visitatori".
}
```

Trascrizione a pagina 4

Ho scritto una query per l'estrazione di un elenco dei musei situati in Emilia-Romagna contenente: il nome del museo, il luogo in cui si trova, le sue coordinate e i dati relativi ai visitatori annui insieme agli anni a cui si riferiscono.

La query restituisce come risultato una tabella con il codice del museo, il nome del museo, il codice del luogo, il nome del luogo, le coordinate del museo, l'anno di cui si conoscono i visitatori annui e il numero di visitatori. I filtri linguistici applicati (ho richiesto risultati che avessero il nome del luogo e del museo esclusivamente in italiano) mi permettono di evitare di avere nel risultato più tuple riguardanti lo stesso museo; infatti, questo (ma anche il luogo) può essere presente in più record che si differenziano esclusivamente per la lingua in cui viene salvato il nome.

Non tutte le tuple presentano i dati relativi ai visitatori annui. Nonostante questo sia il tipo di dato più saliente per la mia ricerca, ho deciso di estrarre anche i dati dei musei senza registrazione dei visitatori per avere una visione complessiva e coerente per quanto riguarda il quadro dei musei in Emilia-Romagna.

Elaborazione

Estraggo il risultato che mi è stato restituito da Wikidata in un file CSV, un formato tabellare adeguato a lavorare con dei fogli di calcolo e importo i dati in un nuovo foglio di lavoro Excel. Importando i dati seleziono l'opzione per indicare che questi sono delimitati e scelgo come delimitatore la virgola. Nomino questo foglio Dati RAW (non elaborati) e ne creo una copia, in questo modo posso lavorare sui dati mantenendo comunque una copia del risultato della query originale.

In questo secondo foglio (Tabella) converto i dati in una tabella e blocco la prima riga del foglio (la riga di intestazione). Dopodiché, ordino i dati per numero di visitatori in ordine decrescente. Ho deciso di mantenere solo i dati più significativi, per questo motivo cancello le tuple relative ai musei che hanno meno di 10000 visitatori in un anno (in questo modo cancello anche i musei di cui non sono stati registrati i visitatori annui). Scorrendo i dati mi accorgo che trovo in realtà dei doppioni. Infatti, per la Galleria Estense di Modena è presente una coppia di tuple per ogni anno, queste differiscono esclusivamente per le coordinate. Guardando la pagina Wikidata del museo scopro che è stato salvato con due set di coordinate; scelgo di usare le tuple con le coordinate registrate nel Catalogo Generale dei Beni culturali.
Resta così un elenco di 55 musei.

Creo allora alcune tabelle pivot per visualizzare i dati.

La prima tabella che creo mi permette di vedere come evolve il numero dei visitatori dei diversi musei nel corso degli anni; per crearla ho impostato come colonne gli anni, come righe i nomi dei musei e come valori la somma dei visitatori. La tabella mi permette di scegliere di visualizzare i dati relativi a tutti o solo uno o alcuni musei. Ho poi creato un grafico pivot ad istogrammi a partire da questa.

La seconda tabella mi permette di vedere come evolve il numero dei visitatori nelle città, non c'è più quindi, come nel grafico precedente, la distinzione per museo. Si ottiene, invece, una visione globale delle diverse località. Per realizzare questo grafico ho inserito come colonne gli anni, come righe i nomi dei luoghi e come valori la somma dei visitatori. Ho poi creato un grafico pivot ad istogrammi a partire da questa.

La terza tabella pivot che ho realizzato permette di vedere la situazione di un singolo luogo, mantenendo i dati separati in base ai musei a cui si riferiscono. Consultando la tabella posso cercare una città che mi interessa e vedere come evolve il numero dei visitatori dei musei che si trovano in questa località. Per realizzare questa tabella ho inserito come filtro il nome del luogo, come colonne i nomi del museo, come righe gli anni e come valori la somma dei visitatori. Ho poi creato un grafico pivot ad istogrammi a partire da questa.

Infine, ho creato una tabella pivot per contare i musei (di cui abbiamo dati riguardo i visitatori annui e che questi siano più di 10000) per città. Il dato più interessante che mi ha permesso di vedere questa tabella è come il numero di registrazioni relative i visitatori sia sceso nel 2020, è possibile intuire che la pandemia di Covid-19 sia una delle cause legate a ciò. Ho poi creato anche un grafico pivot ad istogrammi a partire da questa tabella.

Condivisione/pubblicazione

Ho deciso di condividere il progetto usando GitHub. Ho creato un nuovo repository chiamato I musei in Emilia-Romagna e ho aggiornato il file READme, dopodiché ho caricato il file Excel con i risultati. Infine, ho deciso di realizzare la pagina web del repository (https://aurosiro.github.io/Musei-in-Emilia-Romagna/), per fare ciò ho abilitato pages dalle impostazioni del repository e ho creato la cartella docs dove si trovano i file per la configura e la modifica delle pagine (_config.yml e index.md). Come ultimo passo ho modificato il file index.md per modificare il contenuto visibile dalla pagina web. Infine, ho caricato documento pdf in cui spiego passo per passo cosa ho fatto

Considerazioni conclusive

In questo progetto ho analizzato i dati relativi ai visitatori annui dei musei dell'Emilia-Romagna. Ho ottenuto i dati tramite il database di Wikidata e li ho elaborati usando Excel. Avevo ottenuto 377 risultati, alcune di queste tuple non presentavano il numero dei visitatori annui, mentre altre avevano invece un numero di visitatori troppo basso perché potesse restituire dei risultati significativi per il progetto. Ho deciso di impiegare le tuple in cui il numero dei visitatori era superiore a 10000. Ho trasformato i dati in alcune tabelle e grafici pivot in modo da poter visualizzare l'evoluzione dei visitatori negli anni per museo e per città. Ho poi contato anche il numero dei musei a cui si riferivano questi dati.

Nello svolgimento del lavoro ho dovuto cercare una soluzione principalmente a due difficoltà.

Scrivendo la query ho dovuto trovare un modo per riconoscere quali musei fossero in Emilia-Romagna; infatti, richiedendo esclusivamente la località dei musei mi veniva restituito il comune o la frazione in cui si trovano. Ho quindi dovuto individuare una soluzione che mi permettesse di capire in quale regione fossero situati i comuni. Avevo inizialmente pensato di impiegare una lista dei comuni dell'Emilia-Romagna trovata su Wikidata (Q20894773), ma consultando alcune query simili alla mia ho trovato una soluzione più veloce e mirata. Nella query Museums in Brittany, alla pagina https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples (12.2.1), infatti, viene impiegato un asterisco per trovare tutti i musei che si trovano in questa zona geografica, indipendentemente dal comune o dalla città in cui sono effettivamente ubicati, ho allora deciso di impiegare questa strategia.

Successivamente, nella fase di pulizia dei dati mi sono resa conto che la Galleria Estense di Modena compariva due volte per anno, ho allora dovuto cancellare metà delle tuple relative questo museo. È stato importante accorgersene subito perché altrimenti avrei ottenuto risultati erronei nelle tabelle e nei grafici realizzati in seguito.

Per quanto riguarda queste tabelle ho cercato di ottenere i maggiori dati possibili e di visualizzarli in modalità diverse in modo che fossero il più chiaro e il più leggibile possibile. Anche per questo motivo ho trasformato ogni tabella in un grafico, in quanto permette di avere subito un'idea generale della situazione.

Query

```
SELECT ?museo ?nome ?luogo ?nomeLuogo ?coordinate ?anno ?visitatori
#Uso SELECT e non SELECT DISTINCT perché altrimenti non avrei la possibilità di vedere come cambia il numero dei visitatori dello stesso museo nel corso degli anni.
WHERE {
  ?museo wdt:P31 wd:Q33506.
  #Istanza (wdt:P31) appartenente alla categoria dei musei (wd:Q33506), quindi una struttura espositiva di oggetti.
  ?museo rdfs:label ?nome.
  #Recupero l'etichetta dell'elemento museo e la chiamo "nome".
  FILTER((LANG(?nome)) = "it")
 #Applico un filtro linguistico, voglio che tutti i nomi dei musei siano in lingua italiana. In questo modo evito di ottenere nei risultati dei doppioni, infatti lo stesso museo può essere memorizzato più volte in lingue diverse.
  ?museo wdt:P131 ?luogo;
  #Wdt.P131 indica l'unità amministrativa in cui è situato il singolo museo, do un nome a questa proprietà ("luogo") in modo da poterlo inserire nel SELCT e vederlo nei risultati
          (wdt:P131*) wd:Q1263.
  #Indico che il museo si deve trovare in Emilia Romagna (wd:Q1263). Inserisco * dopo wdt:P131 perché garantisce la ricorsività. Cercando l'unità amministrativa in cui si trova un museo, infatti, ottengo un comune o una frazione, non troverei quindi alcun risultato per l'Emilia-Romagna. L'asterisco invece fa sì che si proceda a ritroso, partendo dal comune/frazione per arrivare prima alla provincia ed infine alla regione in cui il museo è ubicato.
  ?luogo rdfs:label ?nomeLuogo.
  #Recupero l'etichetta dell'elemento luogo e la chiamo nomeLuogo
  FILTER((LANG(?nomeLuogo)) = "it")
  #Applico un filtro linguistico, voglio che tutti i nomi dei luoghi siano in lingua italiana. In questo modo evito di ottenere nei risultati dei doppioni, infatti lo stesso luogo può essere memorizzato più volte in lingue diverse.
  OPTIONAL { ?museo wdt:P625 ?coordinate. }
  #Se disponibili ottengo anche le coordinate del museo (wdt:P625 è una proprietà che indica le coordinate).
  OPTIONAL {
    ?museo p:P1174 ?nodointermedio.
    ?nodointermedio pq:P585 ?anno;
      ps:P1174 ?visitatori.}
    #Se disponibili voglio anche vedere i visitatori annui e l'anno di riferimento di ogni singolo museo. Con p:P1174 indico i visitatori annui, chiamo questo proprietà nodointermedio e la metto in relazione  con una data (pq:P585), proprietà che chiamo "anno". Infine, richiamo la proprietà dei visitatori annui come "visitatori".
}
```