Big Data – MongoDB

Sommario

Tabelle

Codice

Presentazione del dataset

Per questa consegna è stato utilizzato un dataset riguardante la diffusione del COVID-19 in Corea del Sud. Si tratta di un set di dati strutturato basato sui materiali dei rapporti di KCDC[1] e dei governi locali che fornisce la possibilità di analizzare l'evoluzione della pandemia.

I dati presenti sul sito[2] sono strutturati su più file csv contenenti varie informazioni: dalle informazioni sui singoli casi alle cause di contagio fino alle ricerche effettuate sui motori di ricerca prima e dopo lo scoppio della pandemia.

Per questa consegna sono stati presi in considerazione i seguenti set di dati:

- **Case**: informazione sui casi di infezione da COVID-19;
- **Region**: luoghi e informazioni riguardanti le regioni della Corea del sud.
- **Patient_info**: Dati epidemiologici dei pazienti infetti da COVID-19 in Corea del Sud.

I dati presi in considerazione hanno la seguente struttura:

Field	Description	Type
code	Codice identificativo Regione	int
province	Provincia	string
city	Città(-si) / Provincia (-gun) / Distretto (-gu)	string
elementary_school_count	Numero di scuole elementari	int
kindergarten_count	Numero di asili nido	int
university_count	Numero di università	int
nursing_home_count	numero di RSA	int
academy_ratio	Rapporto popolazione che frequenta le scuole	float
elderly_population_ratio	Rapporto popolazione anziana	float
elderly_alone_ratio	Rapporto popolazione anziana che vive da sola	float

Tabella 1: Schema dati Region

[1] Korea Centers for Disease Control & Prevention

[2] https://www.kaggle.com/kimjihoo/coronavirusdataset

Field	Description	Type
case_id	codice_regione(5) + case_number(2)	int
province	Provincia	string
city	Città	string
group	True se il contagio è avvenuto in gruppo, Falso altrimenti	boolean
infection_case	Modalità di contagio. (eventualmente il nome del gruppo)	string
confirmed	Numero di persone contagiate	string

Tabella 2:Struttura dati Case

Field	sex	age
patient_id	codice_regione(5)+patient_number(5)	string
sex	Sesso del paziente	string
age[3]	Età	int
country	Nazione	string
province	Provincia	string
city	Città	string
infection_case	Tipo di contagio	string
infected_by	Eventuali persone contagiate	string
contact_number	Numero di persone esposte al paziente	string
symptom_onset_date	Data dei primi sintomi	string
confirmed_date	Data di avvenuta positività	string
released_date	Eventuale data di dimissione	string
deceased_date	Eventuale data di decesso	string
state	Dimesso, Isolato, Deceduto	string

Tabella 3: Struttura dati Patient_info

[3] Le informazioni memorizzate nel campo *age* sono vaghe e incomplete. I valori sono stati modificati con numeri casuali da 0 a 100.

La modellazione del Dataset

Il dataset è stato modellato cercando di semplificare le query più comuni. Si è fatto uso di **embedding** e **linking**.

È stato creato un database **covid** e tre collezioni:

- Region_info;
- Patient_info;
- Case.

Per la collezione region_info la modellazione dei documenti è stata molto semplice in quanto non è stata effettuata nessuna modifica alla struttura presente nel file csv.

```
{
    "_id": ObjectId("5fcf45e7aabf47ad374b4a99"),
    "code": 10000,
    "province": "Seoul",
    "city": "Seoul",
    "elementary_school_count": 607,
    "kindergarten_count": 830,
    "university_count": 48,
    "nursing_home_count": 22739,
    "academy_ratio": 1.44,
    "elderly_population_ratio": 15.38,
    "elderly_alone_ratio": 5.8
}
```

Codice 1: Esempio di documento presente in Region_info

Per la collezione patient_info è stata effettuata una modellazione dei documenti più complessa con l'uso di embedding per le informazioni possedute in Region e la creazione dell'oggetto *date_history* per tenere traccia delle date relative ai primi sintomi e all'eventuale dimissione/decesso.

Si è ritenuto opportuno effettuare un embedding delle informazioni di Region in quanto sono risultate informazioni accedute spesso con quelle riguardanti il paziente.

```
{
    "_id": ObjectId("5fcf45e7aabf47ad374b4b8d"),
    "patient_id": "1000000001",
    "sex": "male",
    "age": 24,
    "region": {
        "code": 10040,
        "province": "Seoul",
        "city": "Gangseo-gu",
        "elementary_school_count": 36,
        "kindergarten_count": 56,
        "university_count": 1
    },
    "infection_case": "overseas inflow",
    "date_history": {
        "symptom_onset_date": ISODate("2020-01-22T00:00:00.000Z"),
        "confirmed_date": ISODate("2020-01-23T00:00:00.000Z"),
        "released_date": ISODate("2020-02-05T00:00:00.000Z")
    },
    "state": "released"
}
```

Codice 2: Esempio di documento presente in patient_info

Per i documenti presenti nella collezione Case è stato effettuato un linking alla collezione region_info.

Per ogni caso di contagio è importante sapere in quale città è avvenuta la trasmissione del virus.

```
{
    "_id": ObjectId("5fcf45faaabf47ad374b5fba"),
    "case_id": 1000001,
    "group": true,
    "infection_case": "Itaewon Clubs",
    "confirmed": 139,
    "region": {
        "_id": ObjectId("5fcf45e7aabf47ad374b4aae"),
        "city": "Yongsan-gu"
    }
}
```

Codice 3: Esempio di documento presente in Case