OPENCOESIONE

Verso un migliore uso delle risorse: scopri, segui, sollecita.

# OPENCOESIONE HACKATHON VERSIONE B

# LUOGHI DI CULTURA E COESIONE

# LUOGHI DI CULTURA E COESIONE

Il patrimonio culturale italiano vanta quasi cinquemila musei e istituti similari, pubblici e privati, aperti al pubblico. Si tratta di musei, gallerie, collezioni, aree e parchi archeologici, monumenti e complessi monumentali.
Sono migliaia i progetti finanziati in ambito culturale grazie alle politiche di coesione.
Lo scopo di questo progetto è garantire ad ogni regione investimenti proporzionali a livello di svantaggio.

- Quali sono i musei che beneficiano dei fondi per realizzare questi progetti?

- Dove sono e che caratteristiche hanno?

## I DATASET UTILIZZATI

- OpenCoesione: Focus Turismo, Cultura e Natura

- ISTAT: Indagine sui musei e istituzioni similari

La task principale proposta è di  costruire un sistema di comparazione di stringhe testuali tra titolo del progetto, descrizione sintetica del progetto e soggetto beneficiario del finanziamento (OpenCoesione) e denominazione della struttura (ISTAT).

# METODO DI RISOLUZIONE PROPOSTO

Per creare una matrice di raccordo dei due dataset, contenente le chiavi univoche di ciascuno di essi ("COD_LOCALE_PROGETTO" per i progetti OpenCoesione e "OC_COD_MUSEO" per l'anagrafica dei musei) il metodo principale su cui basarsi da noi proposto è l'utilizzo del TF-IDF

## TF-IDF

**TF-IDF** (*term frequency–inverse document frequency*) è una funzione utilizzata per misurare l'importanza di un termine rispetto ad un documento o ad una collezione di documenti dando più importanza ai termini che compaiono nel documento, ma che in generale sono poco frequenti.

$$\text{TF-IDF} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = frequenza di i in j
$df_i$ = numero di documenti contenente i
*N* = Numero totale dei documenti

*Considerando ad esempio un documento contenente 100 parole, in cui il termine x compare 5 volte, il fattore TF per il termine x è* $\frac{5}{100}$ *= 0.05. Assumiamo ora di avere una collezione di 1000 documenti e x compare in 10 di questi allora il termine IDF è log* $\frac{1000}{10} = 2$*. Da questo possiamo calcolare il valore della parola x nel documento iniziale che non è altro che TF*IDF = 0.05*2 = 0.1*

# RISULTATI RILEVANTI

Dei cinquemila musei e istituti similari, pubblici e privati, aperti al pubblico la percentuale rilevata di associazione con i finanziamenti corrispondenti è:

- Tra il **49**% e il **55**% con 40% di match impostando una soglia dello 0.3
- Tra il **59**% e il **67**% con 30% di match impostando una soglia dello 0.4

La misura della probabilità sopra descritta è stata misurata nel primo caso con un indagine a campione svolta manualmente su 200 elementi, e nel secondo caso 200.

# RISULTATI

Dai risultati abbiamo potuto appurare che in ordine di frequenza questi sono i campi in cui sono state trovate più associazioni :

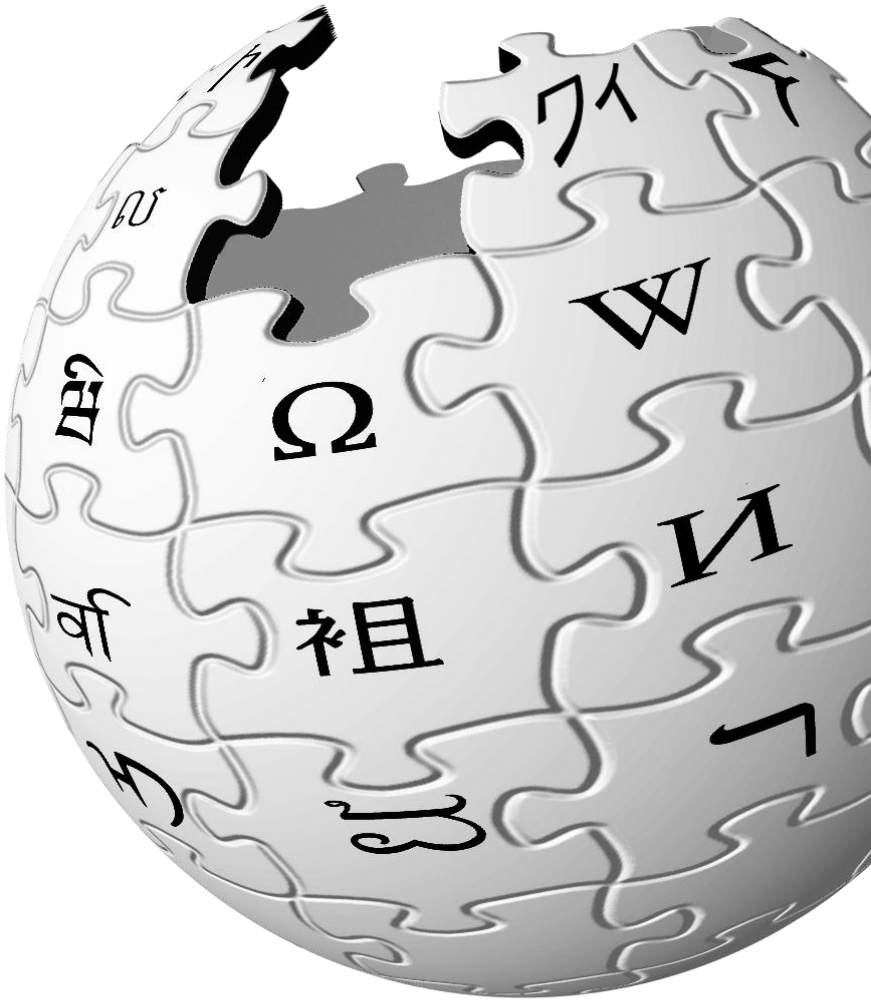
- ARTE da mediovale ad oggi
- Parchi archeologici
- Archeologia
- Etnografia e Antropoligia

# DATA LINKAGE – WIKIPEDIA

A partire dall'anagrafica delle strutture Istat, è stato effettuato un data linkage con le pagine wikipedia dei corrispettivi musei e istituti similari.

È stato stimato che la probabilità di correttezza dell'associazione dei musei con le rispettive pagine wikipedia si trova tra il **70%** e **l'84%**

La misura della probabilità sopra descritta è stata misurata con un indagine a campione svolta manualmente su 40 elementi.



WIKIPEDI
*The Free Encyclope*

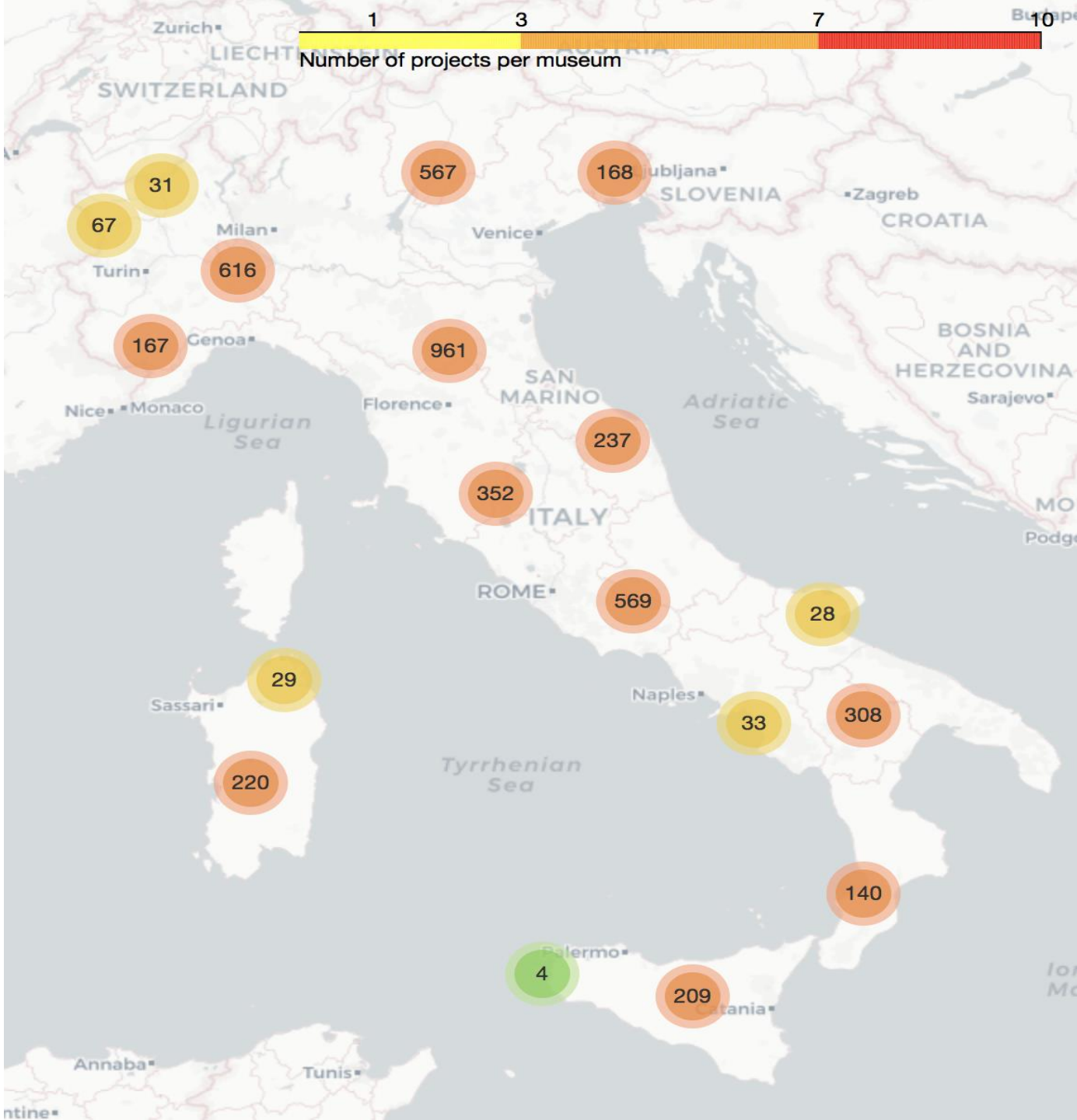
# GEOCODING

È Stato effettuato un Geocoding delle strutture museali e loro mappatura a partire dall'indirizzo stradale contenuto nel dataset fornito da Istat.

Da GoogleMaps api, inserendo la denominazione del museo e il comune di appartenenza abbiamo ottenuto la mappa (interattiva), col numero dei musei, raggruppati per cluster, suddivisi per regione.

Clickando sulle regioni ci si può via via avvicinare fino ad arrivare alla suddivisione dei musei, categorizzati per codice museo .

Di 4889 elementi, 161 non sono rappresentati nella mappa per mancanza delle coordinate geografiche, quindi un 3% del totale.



# SUGGERIMENTI ANALISI FUTURE

- Progettare un custom tf-idf
- Clustering
- Utilizzare tecniche NLP più avanzate (NN, World’s bandwidth)
- Dividere la mappa interattivain base ai fondi stanziati per il progetto stesso.
- Ripetere la stessa analisi dando meno peso ai nomi dei comuni
- Per la task su Wikipedia, Utilizzare più motori di ricerca contemporaneamente e utilizzare ai fini della ricerca solo i risultati presenti su ognuno di questi