Spectral Clustering: analisi della teoria e applicazione allo studio della Visitor Experience alla Pinacoteca Tosio Martinengo

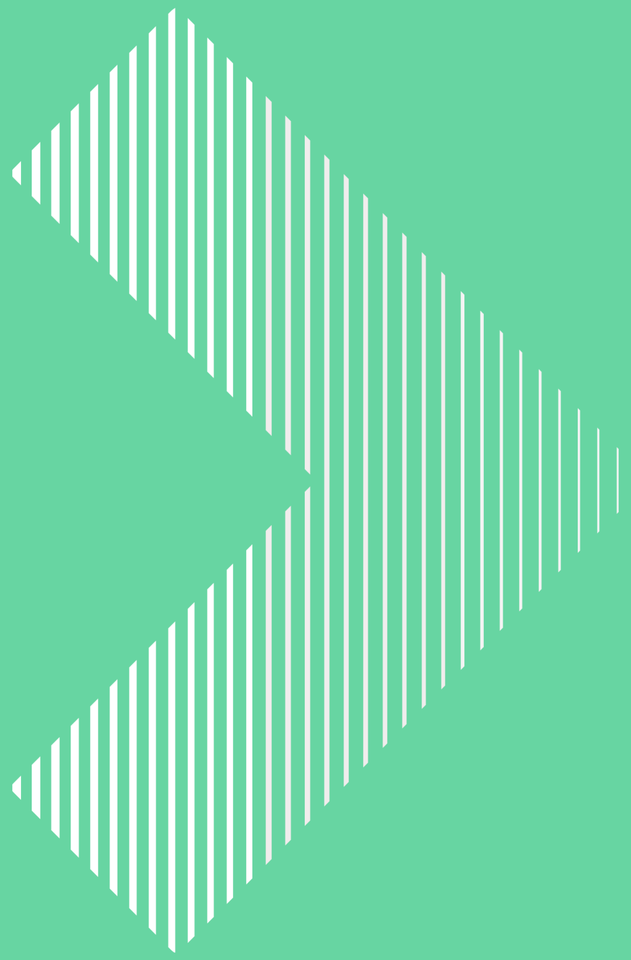
UNIVERSITÀ DEGLI STUDI DI BRESCIA

Elena Romano

DIPARTIMENTO DI ECONOMIA E MANAGEMENT
CORSO DI LAUREA MAGISTRALE IN MANAGEMENT

Relatore:
Chiar.ma Prof.ssa Paola Zuccolotto
Correlatore:
Dott. Matteo Ventura

Indice



1. INTRODUZIONE

Cos'è lo Spectral Clustering?

E’ un metodo di clustering basato sui **grafi** che utilizza gli **autovalori** e gli **autovettori** di una matrice di similarità per raggruppare oggetti simili.

L’idea di base è quella di rappresentare i punti come nodi di un grafo e collegarli in base alla loro **somiglianza** a coppie. Il grafo risultante viene quindi suddiviso in cluster utilizzando **tecniche spettrali**.

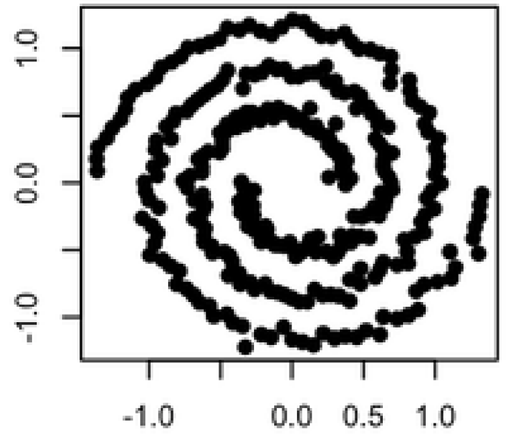
2. VANTAGGI DELLO SPECTRAL CLUSTERING

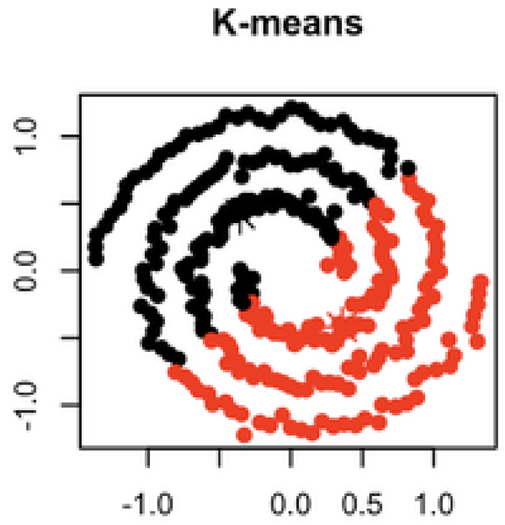
01 Flessibilità nella determinazione del numero di cluster.

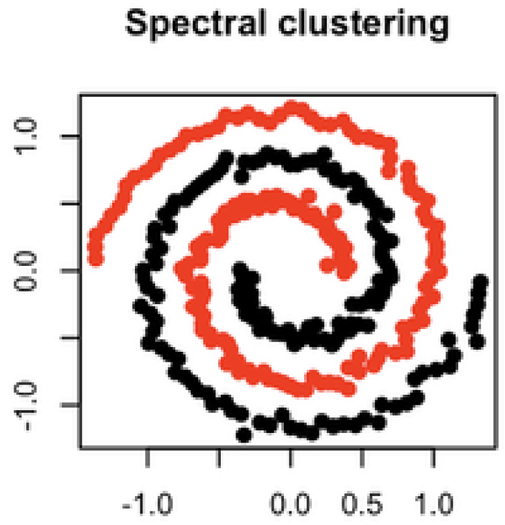
02 Capacità di gestire forme di cluster complesse.

03 Robustezza al rumore e ai dati anomali.

04 Scalabilità per grandi dimensioni.







3. ANALISI DELLA TEORIA

I

DETERMINAZIONE DELLA MATRICE DI SIMILARITÀ

II

COSTRUZIONE DEL GRAFO DI SIMILARITÀ

III

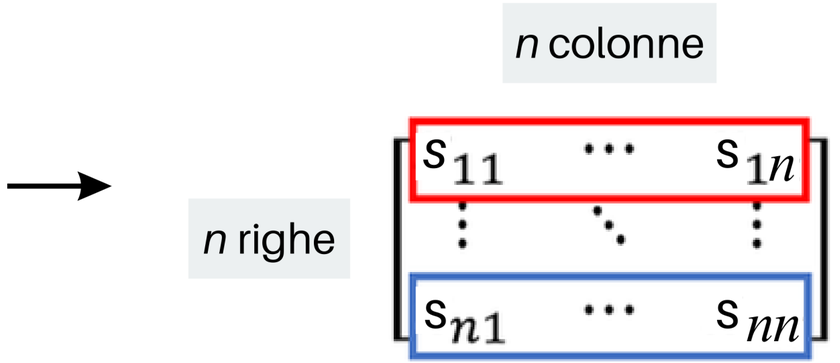
PROIEZIONE IN UNO SPAZIO A DIMENSIONI RIDOTTE

IV

CLUSTERING DEI DATI PROIETTATI

I. Matrice di similarità

La **matrice di similarità** ($n \times n$) riflette le somiglianze tra le unità statistiche, rivelando strutture latenti e agevolando l'individuazione di gruppi omogenei nei dati.



La misura di similarità può variare a seconda del tipo di dati e del contesto del problema. Seguono alcune alternative per **variabili ordinali**:

$$s(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Gaussian Kernel

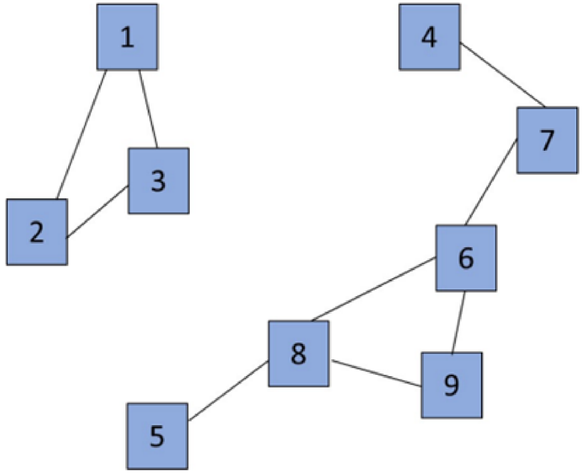
$$s(i,j) = 1 - \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

Indice di Gower

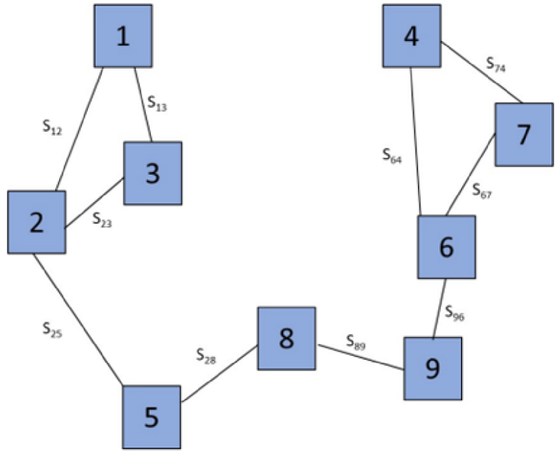
II. Grafo di similarità

Il **grafo di similarità** cattura le relazioni di similarità tra le osservazioni e ricopre un ruolo essenziale nell'identificazione dei cluster.

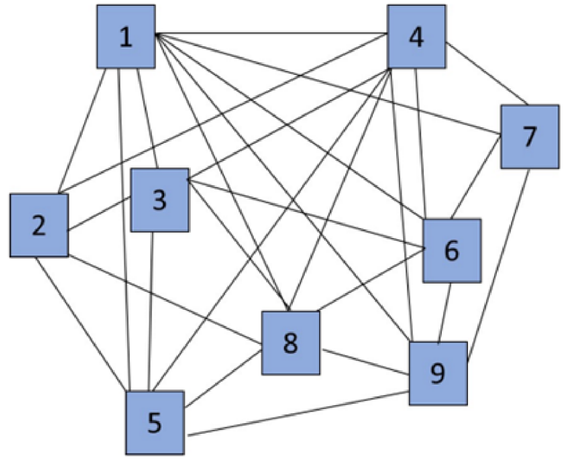
La costruzione di questo grafo può avvenire in diversi modi, tra cui l'utilizzo di grafi quali:



ε-neighborhood graph



k-nearest neighbor graph



Grafo completamente connesso

III. Matrice di adicenza e Matrice Laplaciana

La **matrice di adiacenza** riflette le relazioni tra i punti nel grafo di similarità, rivelando la struttura nascosta dei dati in uno spazio a dimensioni ridotte.

A partire dalla matrice di adiacenza si ottiene una matrice di fondamentale importanza nello Spectral Clustering: la **matrice Laplaciana del grafo**.

In letteratura, esistono diverse definizioni della matrice Laplaciana:

- matrice Laplaciana non normalizzata;
- matrice Laplaciana normalizzata simmetrica;
- matrice Laplaciana normalizzata *random walk*.

E' stato adottato l'algoritmo di Ng et al. (2001), basato sulla **matrice Laplaciana normalizzata simmetrica** per l'analisi.

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

III. Calcolo degli autovalori e degli autovettori

La matrice Laplaciana è cruciale per la **decomposizione degli autovalori**, generando sia gli autovalori che gli autovettori.

Questi **autovettori** consentono la rappresentazione dei dati in uno spazio di dimensione inferiore, mentre gli **autovalori** forniscono indicazioni preziose sulla struttura a cluster dei dati.

Questa decomposizione si basa sull'**equazione caratteristica**:

$$L_{sym}\lambda_i = u_i$$

III. Selezione autovettori e Matrice dei punti U

Calcolo primi *k* autovalori e autovettori dalla matrice Laplaciana

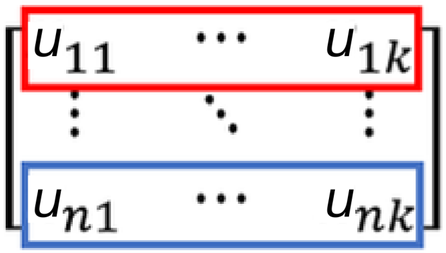
Costituisce un passo cruciale nell'implementazione di *Spectral Clustering*. Esistono diverse strategie con diversi livelli di efficacia, a seconda del contesto e delle assunzioni sul modello.

Calcolo matrice dei punti *U*

Si crea la matrice dei punti ***U*** (*nxk*) dai primi *k* autovettori della matrice normalizzata simmetrica, catturando le informazioni di raggruppamento.

k colonne

n righe



→ **Spectral Embedding**

IV. Matrice normalizzata T e k-means

Calcolo matrice normalizzata *T*

Per ottenere la divisione in *k* cluster, si effettua un processo di "arrotondamento" di ***U*** mediante la ri-normalizzazione delle righe, ottenendo una nuova **matrice *T*** (*n*×*k*).

Per ogni elemento t_*ij* in ***T***, si imposta: $$t_{ij} = \frac{u_{ij}}{(\sum_k u_{ik}^2)^{\frac{1}{2}}}$$

per *i*=1,...,*n* e *j*=1,...,*k*. Dove, *u_ij* è l'elemento corrispondente nella matrice ***U***.

k colonne

n righe

$$\begin{bmatrix} t_{11} & \cdots & t_{1k} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nk} \end{bmatrix}$$

Algoritmo *k-means*

In accordo con Ng et al., si applica l'algoritmo di clustering ***k-means*** sulle righe della matrice ***T*** per suddividere gli *n* punti in ***k* cluster distinti**.

4. APPLICAZIONE

Indagine di Marketing Sensoriale – Pinacoteca Tosio Martinengo

Obiettivo Principale

Analizzare l'**esperienza dei visitatori** e individuare gruppi con esperienze sensoriali simili per ottimizzare la *Visitor Experience* e le strategie di marketing del museo.

Metodologia

Raccolta di **1.024 questionari sensoriali** in stanze museali progettate per suscitare sensazioni uniche. I partecipanti hanno valutato l'intensità di **sei emozioni** su una scala da 1 a 5 (scala di Likert).

Utilizzo dell'algoritmo di **Spectral Clustering** per identificare gruppi di visitatori con esperienze sensoriali simili.

3. Indichi quanto percepisce dentro di sé le seguenti emozioni nella visita di questa sala su una scala da 1 a 5, in cui *1=Per nulla* e *5=Moltissimo*

	1 (Per nulla)	2	3	4	5 (Moltissimo)
Gioia	O	O	O	O	O
Tristezza	O	O	O	O	O
Rabbia	O	O	O	O	O
Paura	O	O	O	O	O
Sorpresa	O	O	O	O	O
Disgusto	O	O	O	O	O

Implementazione dell’algoritmo su R

Matrice di similarità

Calcolo della matrice di similarità con **Gaussian Kernel** ($\sigma = 1.5$) per rappresentare le relazioni non lineari tra variabili e ottenere dati flessibili e precisi.

$$s(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

`sim()`

Grafo di similarità

Costruzione del grafo basata sul metodo dei ***k-nearest neighbors***, con ***k* = 32**.

`make.similarity()`

Matrice di affinità

Creazione della matrice di affinità basata sul metodo dei ***k-nearest neighbors***.

`make.affinity()`

Matrice Laplaciana

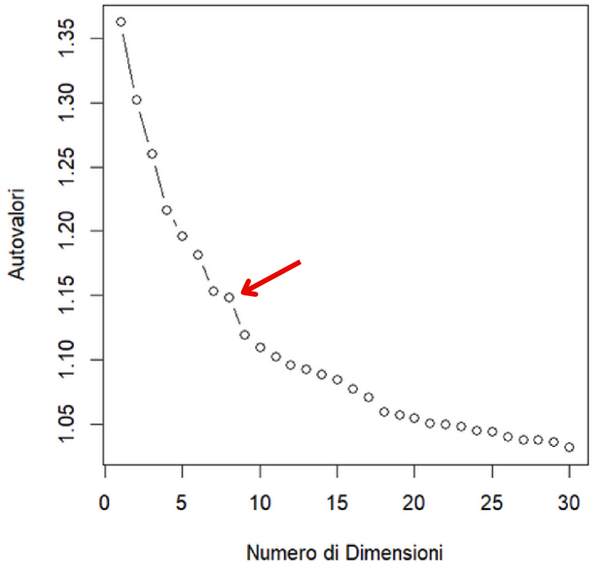
Utilizzo della **matrice Laplaciana normalizzata simmetrica** nell'algoritmo di ***Normalized Spectral Clustering*** per misurare la connettività tra i visitatori.

Calcolo autovalori e autovettori

Calcolo degli autovettori e degli autovalori per analizzare la struttura dei dati e identificare i cluster.

Selezione primi k autovettori

Selezione di **8 autovalori** significativi attraverso il **grafico degli autovalori**.



Implementazione dell'algoritmo su R

`matrixLaplacian()`

`eigen()`

`plot()`

Implementazione dell’algoritmo su R

Matrice dei punti *U*

Generazione della matrice dei punti ***U***, contenente gli autovettori associati ai primi *k* autovalori selezionati.

Matrice normalizzata *T*

Ri-normalizzazione di ***U*** per creare la nuova matrice ***T,*** le cui righe costituiscono nuovi punti in uno spazio *k*-dimensionale.

Algoritmo k-means

Clusterizzazione dei visitatori con l’algoritmo ***k-means*** per analizzare i comportamenti nei cluster.

Estrazione k autovettori

`matrix()`

`kmeans()`

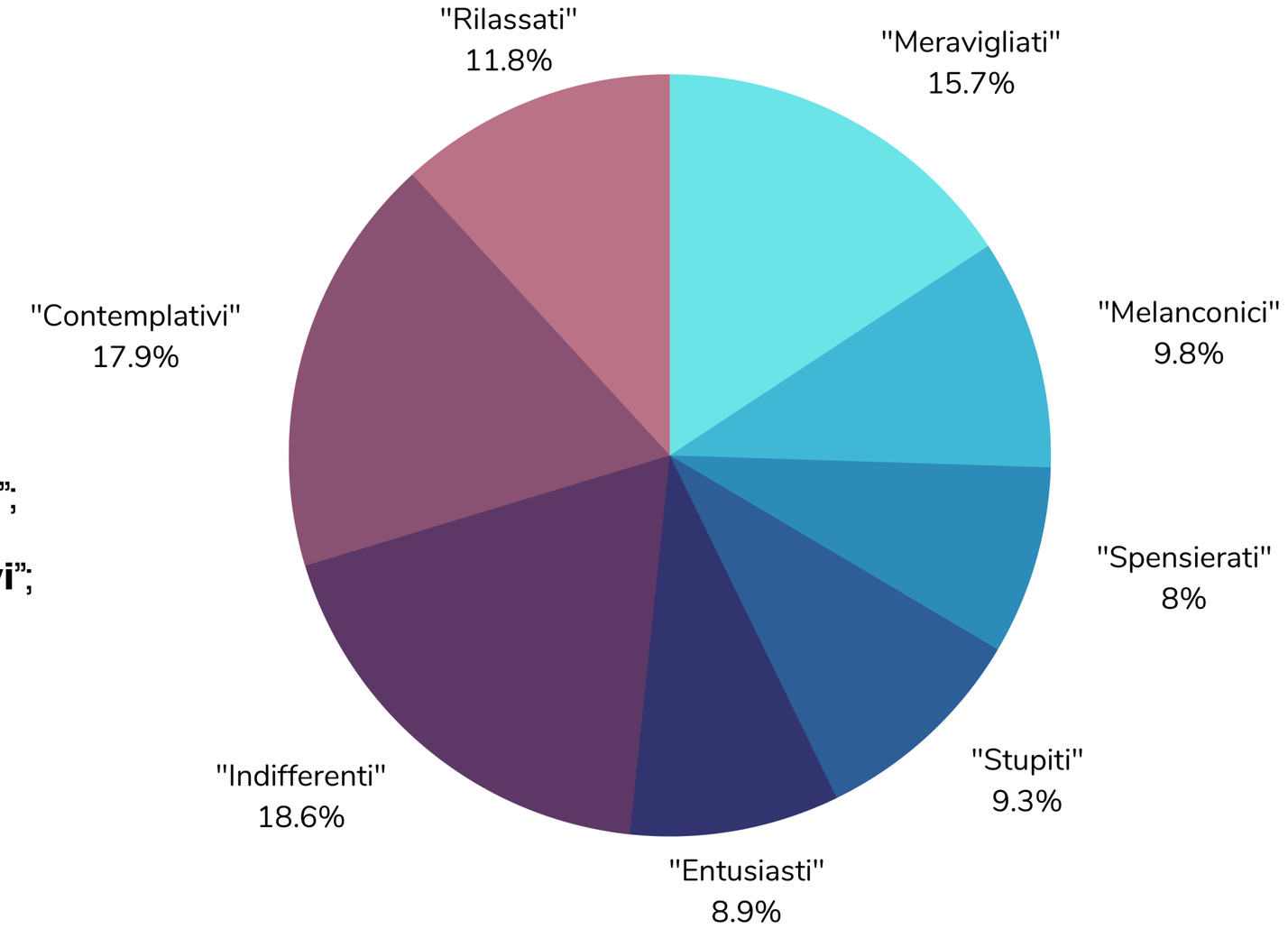
Pacchetto	Funzione	Uso
Funzioni di nostra elaborazione	sim()	Calcola la similarità utilizzando il Kernel gaussiano.
Funzioni di nostra elaborazione	make.similarity()	Crea la matrice di similarità.
Cluster	daisy()	Calcola la matrice delle distanze Euclidee per variabili quantitative.
Cluster	daisy()	Calcola la matrice delle distanze di Gower per variabili ordinali.
Funzioni di nostra elaborazione	make.affinity()	Crea la matrice di affinità utilizzando il metodo dei *k-nearest neighbor*.
MatrixLaplacian	matrixLaplacian()	Calcola la matrice Laplaciana di un grafo rappresentato dalla matrice di affinità W.
Stats	kmenas()	Esegue l'algoritmo di *k-means* sui dati.

Tabella riassuntiva degli strumenti utilizzati in R

Risultati e interpretazione

L’analisi ha portato all’identificazione di **8 cluster**, denominati:

- i “**Meravigliati**”;
- i “**Melanconici**”;
- gli “**Spensierati**”;
- gli “**Stupiti**”;
- gli “**Entusiasti**”;
- gli “**Indifferenti**”;
- i “**Contemplativi**”;
- i “**Rilassati**”.



Rappresentazione percentuale dei cluster

Variabili		1	2	3	4	5	6	7	8
Gioia	Per nulla		31,00%	4,88%	67,37%	4,40%	3,66%	36,61%	
	Poco	3,73%	36,00%	4,88%	14,74%	7,69%	38,74%	44,81%	0,83%
	Abbastanza	21,12%	26,00%	68,29%	16,84%	9,89%	40,31%	16,39%	38,02%
	Molto	45,34%	6,00%	20,73%	1,05%	49,45%	13,61%	2,19%	60,33%
	Moltissimo	29,81%	1,00%	1,22%		28,57%	3,66%		0,83%
Tristezza	Per nulla	77,64%	6,00%	60,98%	54,74%	61,54%	95,29%	44,26%	95,87%
	Poco	16,77%	30,00%	31,71%	32,63%	19,78%	3,66%	28,42%	4,13%
	Abbastanza	4,35%	41,00%	3,66%	6,32%	15,38%	1,05%	19,13%	
	Molto	1,24%	21,00%	3,66%	4,21%	2,20%		7,10%	
	Moltissimo		2,00%		2,11%	1,10%		1,09%	
Rabbia	Per nulla	92,55%	71,00%	89,02%	88,42%	90,11%	97,38%	84,70%	97,52%
	Poco	6,21%	20,00%	7,32%	8,42%	4,40%	2,62%	7,10%	1,65%
	Abbastanza	0,62%	5,00%	2,44%	3,16%	5,49%		6,01%	0,83%
	Molto		4,00%	1,22%				1,09%	
	Moltissimo	0,62%						1,09%	
Paura	Per nulla	90,06%	79,00%	90,24%	89,47%	81,32%	95,29%	88,52%	93,39%
	Poco	4,97%	12,00%	3,66%	2,11%	15,38%	3,14%	6,01%	4,96%
	Abbastanza	4,35%	9,00%	3,66%	7,37%	3,30%	1,57%	3,83%	1,65%
	Molto			2,44%	1,05%			1,64%	
	Moltissimo	0,62%							
Sorpresa	Per nulla		40,00%	4,88%	32,63%	2,20%	21,99%	21,86%	15,70%
	Poco		25,00%	4,88%	11,58%	1,10%	12,57%	38,80%	30,58%
	Abbastanza	12,42%	20,00%	2,44%	26,32%	6,59%	52,36%	18,58%	52,89%
	Molto	40,99%	15,00%	82,93%	18,95%	43,96%	3,66%	18,58%	0,83%
	Moltissimo	46,58%		4,88%	10,53%	46,15%	9,42%	2,19%	
Disgusto	Per nulla	93,17%	64,00%	90,24%	83,16%	89,01%	95,81%	84,70%	97,52%
	Poco	3,73%	15,00%	4,88%	7,37%	7,69%	3,14%	7,10%	1,65%
	Abbastanza	1,24%	10,00%	2,44%	1,05%	2,20%	0,52%	4,92%	
	Molto	0,62%	9,00%	2,44%	8,42%	1,10%	0,52%	2,19%	
	Moltissimo	1,24%	2,00%					1,09%	0,83%

Cluster

Conteggio	>=
	0,95
	0,8
	0,65
	0,5
	0

Risultati e interpretazione

Heatmap

Riflette il numero di visitatori di ciascun cluster che ha risposto a ciascuna delle possibili combinazioni di emozioni e modalità.

Considerazioni conclusive

Lo *Spectral Clustering* è un approccio che va oltre la statistica, consentendo di acquisire una **comprensione più approfondita del pubblico** e di **personalizzare le interazioni** con esso.

Questa applicazione pratica dimostra come la teoria possa tradursi in **risultati tangibili**, evidenziando l'importanza dell'analisi dei dati in svariati contesti.

5. PROSPETTIVE FUTURE

Dati simulati

Utilizzare dati simulati per semplificare l'applicazione dello *Spectral Clustering* e **superare le sfide dei dati reali**, come la complessità nella determinazione del numero di cluster e le osservazioni anomale.

Interpretazione cluster non sferici

Esplorare la letteratura alla ricerca di metodi consigliati da esperti per interpretare cluster non sferici e migliorare la comprensione, specialmente in presenza di **variabili ordinali**.

Applicazioni interdisciplinari

Esaminare le diverse applicazioni dello *Spectral Clustering*, promuovendo **collaborazioni interdisciplinari** per nuove scoperte e approfondimento della conoscenza.

Università degli Studi di Bescia

Grazie per l’attenzione!

Elena Romano