Business Intelligence per i Big Data

Esercitazione di laboratorio N. 4

L'obiettivo dell'esercitazione è:

- utilizzare il software Rapid Miner per effettuare i preprocessing di dati strutturati (relativi ad una campagna promozionale) e di dati non strutturati (ad esempio, dati testuali relativi ad un argomento specifico) per analisi successive.
- applicare i principali algoritmi di clustering disponibili in RapidMiner per segmentare gli utenti della campagna in funzione delle loro caratteristiche anagrafiche e lavorative e le news in base alla similarità dei termini in esse contenuti.

Dati strutturati

Il dataset denominato UsersSmall (UsersSmall.xls) è disponibile sul sito del corso (<u>http://dbdmg.polito.it/wordpress/teaching/business-intelligence/</u>). Esso raccoglie dati anagrafici e lavorativi relativi a circa 300 persone contattate da un'azienda per proporgli l'iscrizione ad un loro servizio. Per tali utenti è noto se, dopo essere stati contattati, si sono iscritti al servizio proposto oppure no (valore del campo Response).

La lista completa degli attributi del dataset a disposizione (UsersSmall.xls) è riportata di seguito.

- (1) Age
- (2) Workclass
- (3) Education record
- (4) Marital status
- (5) Occupation
- (6) Relationship
- (7) Race
- (8) Sex
- (9) Hours per week
- (10) Native country
- (11) Response.

Dati testuali

Il dataset denominato Wikipedia (Wikipedia.zip) è disponibile sul sito del corso (http://dbdmg.polito.it/wordpress/teaching/business-intelligence/). Esso contiene una collezione di 12 articoli di Wikipedia, appartenenti a 3 differenti categorie. In particolare, i documenti appartengono ai seguenti argomenti: matematica, cibo, sport.

Preparazione dei dati strutturati

Obiettivo 1 – Import dei dati

- Nel pannello Operators cercare l'operatore Read Excel e trascinarlo nello spazio di lavoro.
- Importare il dataset UsersSmall.xls utilizzando la procedura guidata Import Configuration Wizard.

Process X XML X			
Process	i 🕹 😜	• 🖂	
Process			
Dinp	r	es	
Read Excel	Parameters X		
fil 🔮 out	🛓 Read Excel		
1	🌮 Import	Configuration Wizard	1
	excel file	Jataset+stopwordlist/Use	D
	sheet number	1	0
	imported cell range	A1:J301	D
	encoding	SYSTEM	0

• Selezionare il file all'interno della cartella in cui si sono scompattati i file e selezionare

ı	Ise	rsSn	nal	l vl	s
L	JJCI	3511	iui	1.71	з.

\vartheta Data import wizard - St	ep 1 of 4			>
This wizard Step 1: Ple	f guides you to import your data. ase select the file that should be imported.			
dataset+stopwordlist			•	+ 📪 🛊 🏦 🖼 🗔 -
Bookmarks	File Name	Size	Туре	Last Modified
🚖 Last Directory	ObamaNews	68 KB	File Folder Foglio di lavoro di Micros	May 14, 2018 / off May 9, 2014
UsersSmall.xis				
Excel Spreadsheets (xis,	.xisx)			,
				Einish X Cancel

• Cliccare Next.

• Controllare che l'intera matrice di dati sia selezionata, in caso contrario selezionare **tutte le colonne**.

Procedere con Next

• Procedere con Finish. Si chiuderà il processo guidato.

🤳 Data imp	ort wizard - Ste	ep 4 of 4								×
	This wizard Step 4: Rap Furthermore operators. 1	guides you to oidMiner Studi e, RapidMiner These roles ca	i import your da o uses strongl Studio assign an be also defi ue types Da	ata. ly typed attribu is roles to the ined here. Fin ate format	ites. In this ste attributes, def ally, you can re	p, you can de ining what the ename attribui	fine the data ey can be use tes or desele	types of your at ad for by the ind act them entirely	tributes. ividual /.	
Preview	uses only first	100 rows.								
V	V	7	-	1	1	√	1	1	√	
Age	Workclass	Education	Marital Statu	Occupation	Relationship	Race	Sex	Native Coun	Response	
integer 🔻	polyno 🔻	polyno 🔻	polyno 🔻	polyno 🔻	polyno 🔻	polyno 🔻	polyno •	polyno 🔻	polyno 🔻	
attribute 💌	attribute 🔻	attribute 💌	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🖪	attribute 🔻	attribute 🔻	
39	State-gov	Bachelors	Never-m	Adm-cler	Not-in-fa	White	Male	United-S	Negative	^
50	Self-emp	Bachelors	Married	Exec-ma	Husband	White	Male	United-S	Negative	
38	Private	HS-grad	Divorced	Handler	Not-in-fa	White	Male	United-S	Negative	
53	Private	11th	Married	Handler	Husband	Black	Male	United-S	Negative	
28	Private	Bachelors	Married	Prof-spe	Wife	Black	Female	Cuba	Negative	~
< O errors.							🗾 Igno	ore errors	; Show only <u>e</u> rr	> ors
Row, Colur	nn	Err	or		Original v	alue		Message		
					←	Previous	→ <u>N</u> ext	Einish	Cano	cel

• Collegare l'uscita dell'operatore Read Excel con res. Usare il tasto destro del mouse.

100% 🔎	P	P	4	2	-	1.A.I
			-	-		100
						res d
						res

- Per lanciare un processo in RapidMiner usare il triangolo in alto nella barra dei processi.
- Per tornare nel processo principale, cliccare su design.

	• •			
File Edit Process View Connections Cloud Settings	Extensions			
🗋 🔚 • 🕤 ਦੇ 🕨 •		Views:	Design	Results

• Analizzare la semantica degli attributi e il loro ruolo a seconda degli obiettivi dell'analisi svolta.

Obiettivo 2 – Gestione dei dati mancanti

Verificare la presenza di eventuali dati mancanti e gestirli con opportuni passi di trasformazione (operatori Decleare Missing values e Replace Missing Values).

• Dichiarare per tutti gli attributi il '?' come valore NULL attraverso l'operatore Declare Missing Value.

• Sostituire i valori nulli dichiarati al punto precedente con il valore più frequente usando l'operatore **Replace Missing Values**.

er type 💙 election special attributes	all nominal	•	© © ©
election special attributes Je	nominal ?	•	0
special attributes Je	nominal ?	v	©
Je	nominal	•	Ð
e	?		
			Œ
100	1% 🖗 🔍 🔍 🙀	3	
Parameters	×		
Replace Miss	ing Values		
attribute filter type	all 🔻	Ð	
invert selection	n	Ð	
include specia	al attributes	Ð	
default	average 🔻	Ð	
	Parameters Parameters Replace Miss attribute filter type invert selection include specia default	100% Parameters Replace Missing Values attribute filter type all invert selection include special attributes default average	100% Parameters Replace Missing Values attribute filter type all invert selection include special attributes default

Obiettivo 3 – Outlier detection

Verificare la presenza di outlier all'interno del dataset, utilizzando una strategia univariata.

- Nel tab Risultati, visualizzare le statistiche calcolate per il dataset da analizzare.
- Nel tab R*isultati*, plottare il grafico *Quartiles* per l'attributo *Age*, selezionandolo tra i diversi Charts disponibili.

Sono presenti possibili outlier per il dataset in questione? Quali?

Verificare la presenza di outlier all'interno del dataset, utilizzando una strategia multivariata.

• Utilizzare il blocchetto DBScan per eseguire l'analisi multivariata. Settare i seguenti parametri: epsilon = 10, minpoints = 3.

Process							
Process >	100% 🔎	₽	P	4	2	7	<u>o</u>
Process							
Read Excel Clustering							
Dip fit de out de angel chu							res
							res
							res (

- Collegare entrambe le uscite del DBScan all'uscita.
- Alcuni record sono stati identificati come outlier (Cluster 0)? Visualizzare i sample appartenenti al Cluster 0, sia in forma tabellare che in forma di grafico (Scatter chart, assi x e y uguali a *Age* e *Education* rispettivamente, color column uguale a *cluster*).



Confrontare i risultati dell'analisi univariata e dell'analisi multivariata.

Rimuovere (per t	utte le analisi	successive) gli	outlier	identificati,	utilizzando	l'operatore	Filter	Exmaples.
------------------	-----------------	-----------------	---------	---------------	-------------	-------------	--------	-----------



Obiettivo 4 – Discretizzazione

Verificare la presenza di attributi continui nei dati di origine.

- Discutere l'eventuale necessità di applicare un processo preliminare di discretizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati.
- Applicare diverse tecniche di discretizzazione (operatori *Discretize by binning, Discretize by frequency, Discretize by size*) e confrontare i risultati.

Process	ess	-			100	% / ^D / ^D	۶ 📮	2		
Process inp	4	Read Excel	Declare Missing Val	Replace Missing Values	F		_		1.0	res
				pre	Parameters Replace Missi	x ng Values				
					attribute filter type	all	٠	Ð		
					invert selection			Ð		
					include special	lattributes		Ð		
					default	average	٠	Ð		
					Show advance	d parameters atibility (8.2.0	L 00)			

Analizzare i risultati della discretizzazione scelta.

Bonus: dati strutturati con attributi continui

• Utilizzare l'operatore **Generate Sales Data**. Nella barra dei parametri, settare il parametro number examples a 100.

Process >		100% 🔎	0	ρ 📮	2		E
Process Generate Sales Data	Parameters	×					res
	Generate Sales	Data					
	number examples	100			Œ)	

• Selezionare solo gli attributi numerici. Utilizzare l'operatore Select Attribues.

			Parameters	×
Process >		100% 🔎 🔎 📮 🚰 🔠	Select Attribut	tes
Process			attribute filter type	subset 🔻 🗊
inp Generate Sales Data Select Attributes		res	attributes	Select Attribut 🛈
ori A		res	invert selection	ı D
			include special	l attributes ①
Select Attributes: attributes				
Select Attributes: attributes				
The attribute which should be cho	osen.			
Attributes		Selected Attributes		
Search	X	Search		OX
	and the second sec	250.535.50.0		and the second sec
and a second state		[
customer_id		amount		
customer_id date product category		amount product_id single_price		
customer_id date product_category store_id		amount product_id single_price		
customer_id date product_category store_id transaction_id	0	amount product_id single_price		
customer_id date product_category store_id transaction_id	0	amount product_id single_price		
customer_id date product_category store_id transaction_id	0	amount product_id single_price		
customer_id date product_category store_id transaction_id	0	amount product_id single_price		

• Analizzare la correlazione tra coppie di attributi (operatore *Correlation Matrix*). Inserire l'operatore "Correlation Matrix" in coda al processo e visualizzare la rispettiva matrice collegando il plug-in del blocco denominato "mat" al plug-in "Result" sulla destra della finestra del processo principale. Il processo così generato sarà analogo al seguente:

Process ×				4000	0	0	-	2	-	121
Process >				100%	p	P	4	4		M
Process										
	Generate Sales Data	Select Attributes	Correlation Matrix							
Dinp	tuo 📲	(eta (eta)	era			-				res
12.112	J =	ori	mat)		-	1				res (
		1	wei							
			1							

- Esiste qualche correlazione elevata?
- Eliminare l'operatore correlation matrix (selezionarlo con il mouse e premere canc).
- Discutere l'eventuale necessità di applicare un processo preliminare di normalizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati (operatore *Normalize*).

Process ×								
Process >				100% 🔎	Ð	ρ.	2	$\overline{\bigcirc}$
Process								
	Generate Sales Data	Select Attributes	Normalize					
) inp	out 🖉	exa exa	exa exa					 res
	-	ori	ori					res (
		V	pre					

- Quale normalizzazione scegliereste? (nel caso non compaia method, cliccare su *Show advanced parameters*)
- Quando è utile normalizzare i dati?

Preparazione dei dati testuali

Obiettivo 1 – Import dei dati

• Importare il dataset Wikipedia in Rapid Miner (operatore Process Documents From Files).

Process X X	ML ×		Parameters ×		
≡ ▼ Process >		🗵 🗣 🗧	Process Documents f	rom Files	
Process		^	text directories	🌄 Edit List (1)	
Process Do	cuments from Files	res	file pattern	*	
	wor	res	<pre>extract text only</pre>		
			✓ use file extension as typ	e	
Edit Parameter List: text directories	-		encoding	SYSTEM	•
In this list arbitrary directories can be loaded and assigned to the class val	e specified. All files matching the g ue provided with the directory.	given fil <mark>e</mark> ending will be	create word vector		
class name	directory		vector creation 💙	TF-IDF	•
pama	path cartella contienente la	a collezione testuale	☑ add meta information		
			🗌 keep text 💙		
			prune method 💙	none	•
			datamanagement	double_sparse_array	•
			parallelize vector creation	n	
			Hide advanced paramet	ers.	
			Help X		
			Process Docum	ents from Files	
A	ld Entry	Apply X Cancel	Synameie		

• Se volete avere l'informazione del testo all'interno dei risultati, spuntate la voce **Keep Text** nel pannello dei parametri dell'operatore **Process Documents from Files**.

Obiettivo 2 – Generazione dei token, stopwords e stemming

Il **TF-IDF** (*Term Frequency–Inverse Document Frequency*) è una funzione nota nel text mining utilizzata per misurare l'importanza di un termine rispetto ad una collezione di documenti. Il TF-IDF aumenta

proporzionalmente al numero di volte che il termine è contenuto nel documento, ma cresce in maniera **inversamente proporzionale** con la frequenza del termine all'interno della collezione. In questo modo si possono penalizzare le parole molto frequenti che non danno rilevanza alla collezione e dare più importanza ai termini che in generale sono poco frequenti ma più rilevanti per l'analisi.

$$\mathsf{tfidf}_{i,j} = \mathsf{tf}_{i,j} \times \log\left(\frac{\mathbf{N}}{\mathbf{df}_{i}}\right)$$

 $\begin{aligned} \mathbf{f}_{ij} &= \text{ total number of occurrences of } i \text{ in } j \\ \mathbf{df}_i &= \text{ total number of documents (speeches) containing } i \\ \mathbf{N} &= \text{ total number of documents (speeches)} \end{aligned}$

L'operatore *Process Document from Files* ammette un sottoprocesso per poter pulire il dataset e trasformarlo in una tabella chiamata matrice documenti*termini. La tabella avrà una riga per ogni documento della collezione presente nella cartella letta e una colonna per ogni termine presente all'interno della collezione.

• Prima di creare la matrice, guardare l'output. Nel sottoprocesso (per entrare nel sottoprocesso, fare doppio click con il tasto sinistro sull'operatore *Process Document from Files*) collegare le uscite come in figura.

Process ×		
Process Process Documents from Files	100% 🔑 🔑 📮 🍓 💣 🗄	E
Process Documents from Files		
doc	dor	1
	doc	Q

- Per tornare al processo principale, cliccare la freccia blu di fianco a *Process*.
- Collegare l'uscita exa con res ed eseguire il processo.



- Applicare i passi di pre-processing visti nell'esercitazione precedente. Doppio click sull'operatore **Process Documents from Files**. Verrà aperto un sottoprocesso. Utilizzare i seguenti blocchi:
 - Il blocchetto **Tokenize**: splitta ogni documento della collezione Obama in un vettore di parole. L'ordine delle parole non sarà più rispettato. Secondo te ha importanza ai fini dell'analisi? (Settare il parametro non letters).
 - \circ $\;$ Il blocchetto Transform Cases: Trasforma il testo in maiuscolo o minuscolo.
 - Il blocchetto Stem (Snowball): Riduce le parole alla propria radice. La radice è quell'elemento linguistico irriducibile (non ulteriormente suddivisibile) che esprime il significato principale della parola. (Utilizzare la lingua italiana).
 - Il blocchetto Filter Stopwords (Dictionary): Permette di eliminare le parole definite Stopword, parole che non hanno un particolare significato se isolate dal testo e quindi vengono ignorate dai programmi. Sono parole poco significative perché possono essere usate spesso all' interno delle frasi. Ad esempio articoli, congiunzioni e preposizioni non

caratterizzano il significato di un testo, possono essere eliminate a monte di una analisi text mining. Carica il file **stopwordsEnglish.txt** presente sul sito del corso.

Process X XML X			Parameters X		
Process ▶ Process Documents from Files ▶		图 🖷 🗲 📮	Process Documents	from Files	
Process Documents from Files			text directories	🍃 Edit List (1)	D
Tokenize Transform Cases Stem	(Snowball) Filter Stopwords	doc	file pattern	*	D
		doc	<pre>extract text only</pre>		٢
			✓ use file extension as ty	pe	Ð
			encoding	SYSTEM	•
			✓ create word vector		1
			vector creation 💙	TF-IDF	• 1
			add meta information		D
			📝 keep text 💙		٢
	Parameters	×			
	📕 Filter Stopw	ords (Dictionary)			
	file	wordsItalian.	D		
	case sensitive		Ð		
	encoding	UTF-8	•		

- Utilizzare la codifica UTF-8 per il file delle stopword.
- Torna al processo iniziale cliccando sulla freccia blu sotto la voce Process.

Clustering di dati strutturati

L'obiettivo dell'analisi è raggruppare le persone in gruppi omogenei, tali che persone appartenenti al medesimo gruppo abbiano caratteristiche simili mentre persone appartenenti a gruppi diversi siano dissimili. I gruppi possono rappresentare segmenti di clientela verso cui mirare specifiche promozioni o campagne pubblicitarie.

Obiettivo 1 – Import e preprocessing dei dati

Eseguire i diversi step di preprocessing, come imparato nella prima parte dell'esercitazione.

- In particolare, eseguire i gli step:
 - o Import dati
 - o Declare and replace missing values

Process			🖸 🗣 🍒
Process			
) inp Read Excel Declare M	Missing Replace Missing S	elect Attributes exa exa ori	res
	attribute filter type 💙	subset	•
	attributes	🗟 Selec	t Attributes
	invert selection 💙		D
	include special attributes		0

• Escludere l'attributo *Response* dall'analisi usando l'operatore *Select Attributes*.

Select Attributes: attributes Select Attributes: attributes The attribute which should be chosen. Attributes	s	elected Attribu	tes			
Search 🗶		Search			0	×
Response		Age Education Marital Status Vative Country Occupation Race Relationship Sex Workclass				
				O Apply	×	ance

• **Normalizzare** i valori degli attributi numerici indicando come intervallo di valori [0-1] utilizzando l'operatore **Normalize**. L'unico attributo che verrà normalizzato è l'attributo età.

rocess X XML X	Parameters X				
🕒 Process > 🛛 📮 🍹 😰	Normalize				
Process	create view		(j)		
) inp res (attribute filter type 💙	all	١		
Read Excel Declare Missing Replace Missing Select Attributes	invert selection		(j)		
fil to out exa to exa t	include special attributes				
pre l	method 💙	range transformation 🔻	1		
	min	0.0	1		
exa exa	max	1.0	٩		
pre					

• Quando avete bisogno di utilizzare lo stesso input per diversi algoritmi, utilizzate l'operatore **Multiply**. Nei prossimi step verranno comparati diversi algoritmi di clustering.

Obiettivo 2 – Clustering dei dati

• Applicare l'algoritmo di clustering *k-Medoids* (quali differenze ha rispetto al K-means?) settando a K=2 il numero di cluster. Esegui il processo e analizza i risultati. Come sono distribuiti i due cluster trovati?

Process >	📓 Clustering (k-Medo	ids)	
Process	✓ add cluster attribute		٢
	add as label		١
Read Excel Declare Missing Replace Missing Select Attributes	remove unlabeled		
exa exa exa ori exa ori exa ori exa	k 💙	2	1
pre	max runs	10	1
	max optimization steps	100	١
Normalize Multiply Clustering	use local random seed	d	١
pre clu	measure types 💙	MixedMeasures •	1
	mixed measure	MixedEuclideanDista 🔻	1

• Applicare l'algoritmo di clustering *Agglomerative (Agglomerative Clustering)*. Selezionare due cluster dal risultato dell'algoritmo di clustering *Agglomerative* utilizzando l'operatore *Flatten Clustering*. Esegui il processo e confrontare il risultato ottenuto con quello prodotto dall'algoritmo *k-Medoids* (numero di cluster k=2) svolto precedentemente. Come sono distribuiti gli elementi per cluster?



Obiettivo 3 – Valutazione oggettiva dei cluster generati per l'algoritmo K-Medoids

• Calcolare l'**SSE** (Sum of Squared Errors) dei cluster generate con l'algoritmo *K-Means*. Per il calcolo usare lo script **sse.script** disponibile sul sito del corso; copia e incolla il codice dello script all'interno dell'opportuna textbox dell'operatore **Execute Script**. (NB. Settare all'interno del codice il corretto algoritmo (KMEANS o DBSCAN)).

ocess X XML X		Parameters X			
🕼 Process + 🧧 🧯	- 13	Execute Script			
Process		script	📿 Edit Tex	t (2629 charact	ters)
Read Excel Declare Missing Replace Missing Select Attributes	4Q	Standard imports			
	Edit Pa	arameter Text: script dt Parameter Text: script he script to execute.		_	
Hormalize Huitiply exc and any exc a	This Inpu inpu Inpu Outp The ort co	script permits to calculate the t: t(0): the cluster model coming of t[1]: the example set of the clu ut: SSE value of the clustering will m.rapidminer.operator.clustering m.rapidminer.operator.clustering	e SSE measure of a g out from the cluster ustering L be displayed in lo g.ClusterModel; g.Cluster;	operator g console,	ing.
Clustering (3) Flatten Clustering 20 int (eva () () () () () () () () () (DBSCA KMEAN Set t ALGO sterMo mpleSe	N = 1; S = 0; he current clustering algorithm = KMENS; End setting del clustering = input[0]; t clusteringSet = input[1];	••/		>
ecommended Operators			Enlarge	Apply	

- Eseguite il programma. Il valore di SSE viene riportato nel tab Log della pagina dei risultati.
- Rieseguire il processo di valutazione precedente per **differenti valori di K** per l'algoritmo **K-medoids**. Come scegliere il valore ottimale di K? (Suggerimento: utilizza il metodo del ginocchio o del gomito visto durante la spiegazione del clustering in classe).
- Eliminare il ramo del Clustering agglomerativo (o disabilitarlo con il tasto destro).

Obiettivo 4 – Visualizzazione/validazione del risultato di un processo di clustering tramite l'uso di tecniche di riduzione delle dimensioni dei dati - SVD (Singular Value Decomposition)

- Analizzare la qualità del clustering generato mediante una tecnica di riduzione della dimensionalità dei dati, nota come *Singular Value Decomposition* (SVD). SVD permette di proiettare dei dati a N-dimensioni in uno spazio a K-dimensioni, con K scelto dall'utente e minore di N.
- Applicate l'operatore *SVD (Singular Value Decomposition)* sul dataset generato dal processo di clustering realizzato al passo precedente. Eseguire il processo impostando K=3 e visualizzare su un grafico di tipo scatter i dati rispetto alle tre dimensioni individuate dall'operatore SVD. Usare l'attributo cluster come attributo per la **scelta dei colori dei punti**. I cluster sono ben definiti?



• Moltiplica l'uscita del clustering K-medoids e utilizza l'operatore **Nominal to Numerical**. Collegalo all'operatore **SVD** e esegui il programma. Salva il processo sul tuo computer.



Clustering di dati testuali

La seconda parte di questa esercitazione prevede l'analisi attraverso l'algoritmo K-Means della collezione di articoli denominata *Wikipedia*. Scompatta la cartella presente sul sito del corso ed esegui i passi seguenti.

Obiettivo 1 – Import e preprocessing dei dati

• Trasforma la collezione di documenti nella matrice **document*term**. Per eseguire questa trasformazione, eseguire i diversi step di preprocessing, come imparato nella prima parte dell'esercitazione.

Obiettivo 2 – Clustering dei dati

• Utilizzare l'algoritmo di **K-Means** per dividere la collezione in gruppi omogenei di documenti che parlino di uno stesso topic.

Process X XML X		Parameters ×		
Process >	🙀 🌏 💣 🐼	📓 Clustering (2) (k-N	1eans)	
Process	^	✓ add cluster attribute	э (D ^
Process Docume Clustering (2) Execute Script		add as label	(D
wor clu inp		remove unlabeled	(D
		k 💙	2	D =
		max runs	10	D
		determine good star	rt values 💙	D
		measure types 💙	NumericalMeasur 🔻	D
C Decommonded Operators	>	numerical measure	CosineSimilarity 🔻	Ð
Retrieve status Retrieve status Select Attributes status	ribu 📌 31%	The advanced para	meters	~

Per i dati testuali la misura per calcolare la distanza tra punti (in questo caso tra due documenti) è la **Cosine Similarity**. Come viene calcolata? Rispetto a cosa si differenzia dalla norma Euclidea?

- Riesegui il K-means diverse volte per cercare il parametro K migliore. Utilizza l'SSE come metrica di valutazione (guarda la prima parte dell'esercitazione).
- Visualizza i cluster trovati attraverso la tecnica SVD.
- Ha senso calcolare la correlazione prima di eseguire il K-means? Perché?