

# Business Intelligence per i Big Data

---

## *Esercitazione di laboratorio N. 4*

L'obiettivo dell'esercitazione è:

- **utilizzare il software Rapid Miner per preparare i dati relativi ad una campagna promozionale e i dati testuali relativi ad un argomento specifico per analisi successive**

### **Dati strutturati**

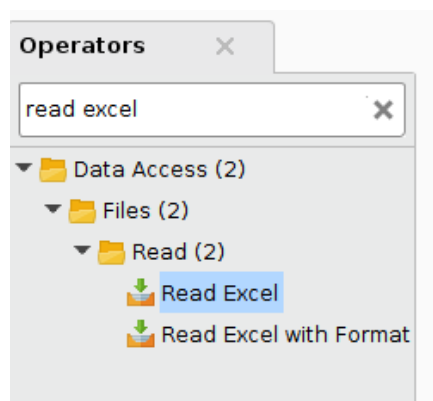
Il dataset denominato UsersSmall (UsersSmall.xls) è disponibile sul sito del corso (<http://dbdmg.polito.it/wordpress/teaching/business-intelligence/>). Esso raccoglie dati anagrafici e lavorativi relativi a circa 300 persone contattate da un'azienda per proporgli l'iscrizione ad un loro servizio. Per 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.

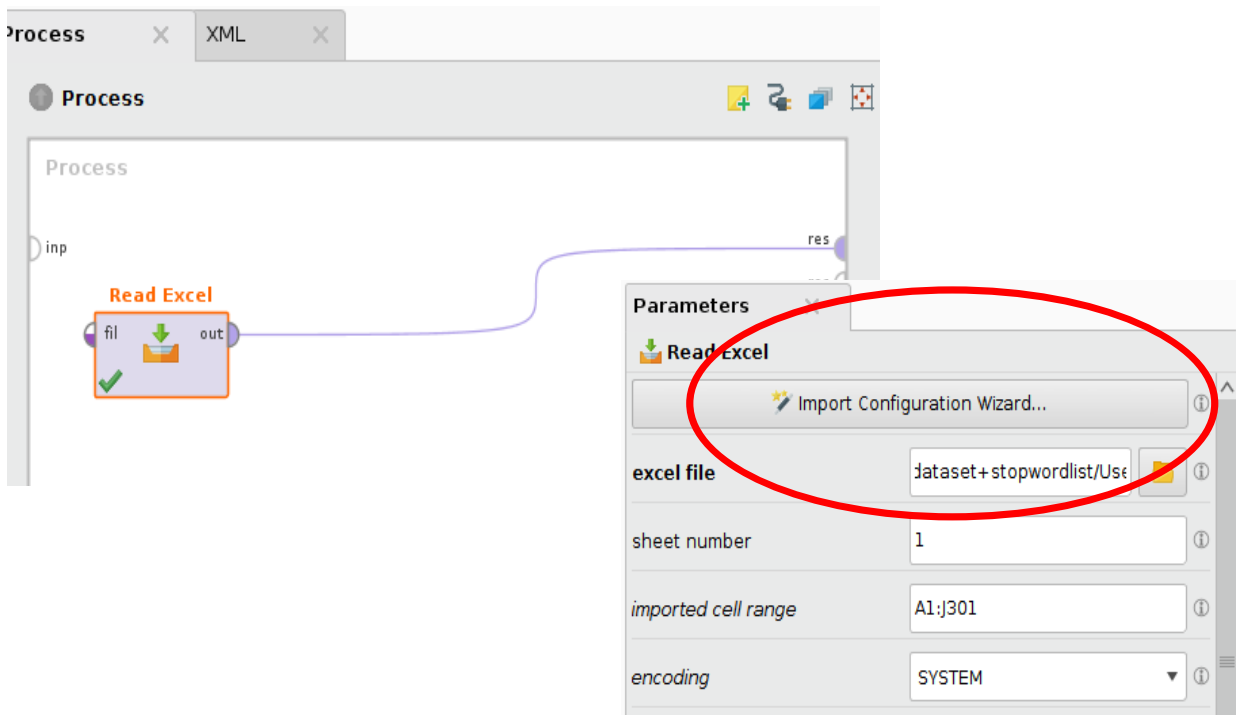
- (1) Age
- (2) Workclass
- (3) Education record
- (4) Marital status
- (5) Occupation
- (6) Relationship
- (7) Race
- (8) Sex
- (9) Hours per week
- (10) Native country
- (11) Response.

### **Obiettivo 1 - import dei dati**

- Nel pannello **Operators** cercare l'operatore **Read Excel** e trascinarlo nello spazio di lavoro.



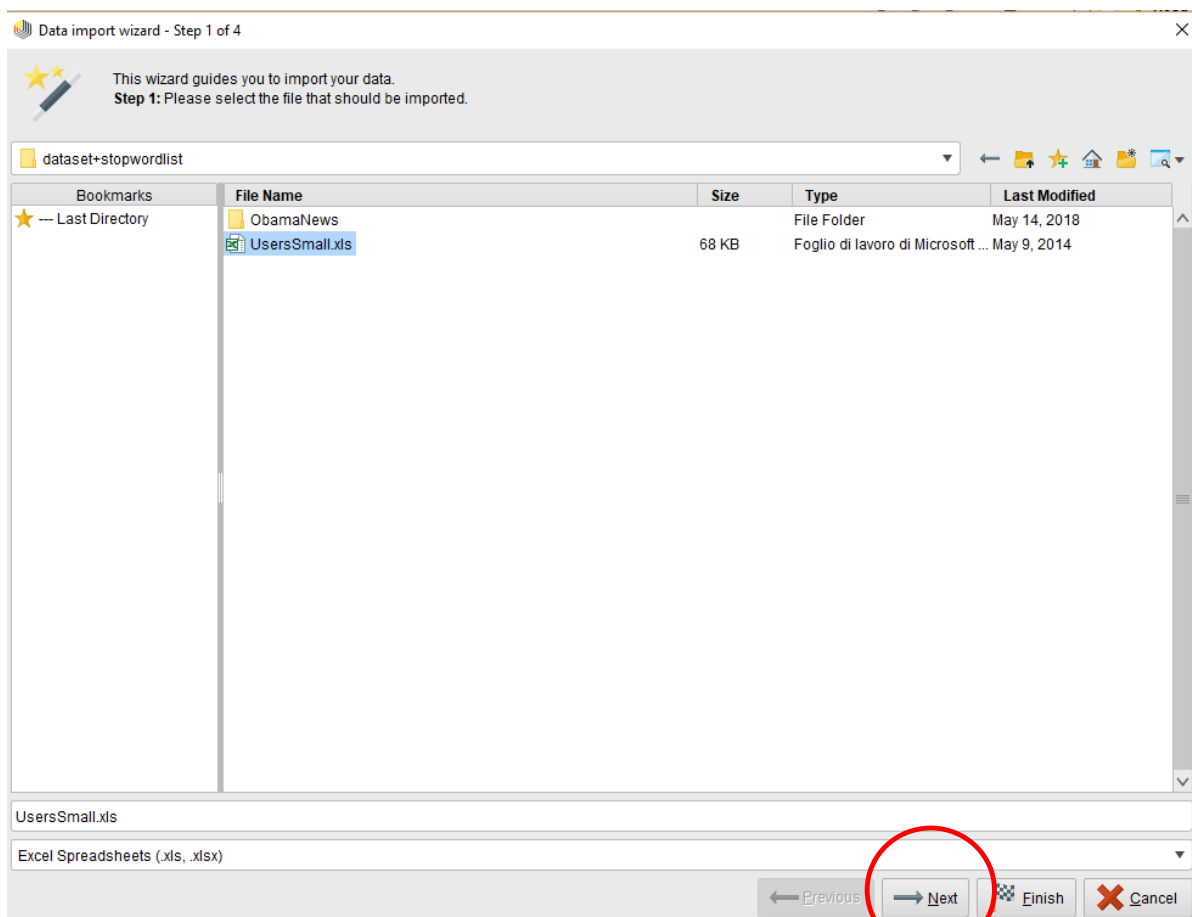
- Importare il dataset *UsersSmall.xls* utilizzando la procedura guidata *Import Configuration Wizard*.



Si aprirà l'import guidato:

Selezionare il file all'interno della cartella in cui si sono scompattati i file e selezionare *UsersSmall.xls*.

Cliccare *Next*.



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

Data import wizard - Step 2 of 4

This wizard guides you to import your data.  
**Step 2:** An Excel file can contain multiple sheets. Please select the one you want to import into RapidMiner Studio. Furthermore, you can mark a range of cells to be loaded.

Users

A	B	C	D	E	F	G	H	I	J
Age	Workclass	Education	Marital St...	Occupati...	Relation...	Race	Sex	Native C...	Response
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
37	Private	Masters	Married-...	Exec-ma...	Wife	White	Female	United-S...	Negative
49	Private	9th	Married-...	Other-se...	Not-in-fa...	Black	Female	Jamaica	Negative
52	Self-emp...	HS-grad	Married-...	Exec-ma...	Husband	White	Male	United-S...	Positive
31	Private	Masters	Never-m...	Prof-spe...	Not-in-fa...	White	Female	United-S...	Positive
42	Private	Bachelors	Married-...	Exec-ma...	Husband	White	Male	United-S...	Positive
37	Private	Some-co...	Married-...	Exec-ma...	Husband	Black	Male	United-S...	Positive
30	State-gov	Bachelors	Married-...	Prof-spe...	Husband	Asian-Pa...	Male	India	Positive
23	Private	Bachelors	Never-m...	Adm-cler...	Own-child	White	Female	United-S...	Negative
32	Private	Assoc-a	Never-m	Sales	Not-in-fa	Black	Male	United-S	Negative

← Previous   **→ Next**   🚩 Finish   ❌ Cancel

Procedere con *Next*.

Data import wizard - Step 3 of 4

This wizard guides you to import your data.  
**Step 3:** In RapidMiner Studio, each attribute can be annotated. The most important annotation of an attribute is its name - a row with this annotation defines the names of the attributes. If your data does not contain attribute names, do not set this property. If further annotations are contained in the rows of your data file, you can assign them here.

Annotat...	A	B	C	D	E	F	G	H	I
Name	Age	Workclass	Education	Marital St...	Occupati...	Relation...	Race	Sex	Native C...
-	39	State-gov	Bachelors	Never-m...	Adm-cler...	Not-in-fa...	White	Male	United-S...
-	50	Self-emp...	Bachelors	Married-...	Exec-ma...	Husband	White	Male	United-S...
-	38	Private	HS-grad	Divorced	Handler...	Not-in-fa...	White	Male	United-S...
-	53	Private	11th	Married-...	Handler...	Husband	Black	Male	United-S...
-	28	Private	Bachelors	Married-...	Prof-spe...	Wife	Black	Female	Cuba
-	37	Private	Masters	Married-...	Exec-ma...	Wife	White	Female	United-S...
-	49	Private	9th	Married-...	Other-se...	Not-in-fa...	Black	Female	Jamaica
-	52	Self-emp...	HS-grad	Married-...	Exec-ma...	Husband	White	Male	United-S...
-	31	Private	Masters	Never-m...	Prof-spe...	Not-in-fa...	White	Female	United-S...
-	42	Private	Bachelors	Married-...	Exec-ma...	Husband	White	Male	United-S...
-	37	Private	Some-co...	Married-...	Exec-ma...	Husband	Black	Male	United-S...
-	30	State-gov	Bachelors	Married-...	Prof-spe...	Husband	Asian-Pa...	Male	India
-	23	Private	Bachelors	Never-m...	Adm-cler...	Own-child	White	Female	United-S...
-	32	Private	Assoc-a	Never-m...	Sales	Not-in-fa...	Black	Male	United-S...

← Previous   **→ Next**   🚩 Finish   ❌ Cancel

Procedere con *Finish*. Si chiuderà il processo guidato.

This wizard guides you to import your data.  
**Step 4:** RapidMiner Studio uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner Studio assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.

Reload data     Guess value types    Date format:

Preview uses only first 100 rows.

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

0 errors.     Ignore errors     Show only errors

Row, Column	Error	Original value	Message
-------------	-------	----------------	---------

Collegare l'uscita dell'operatore Read Excel con res (vedere figura seguente). Usare il tasto destro del mouse.

Process

Process

inp

Read Excel

fil

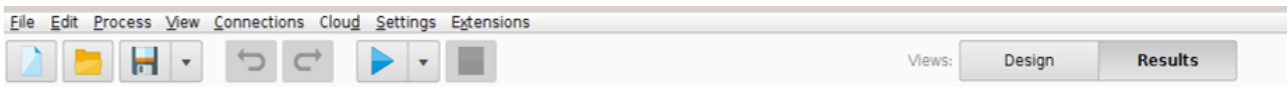
out

res

res

Per lanciare un processo in RapidMiner usare il triangolo in alto nella barra dei processi.

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

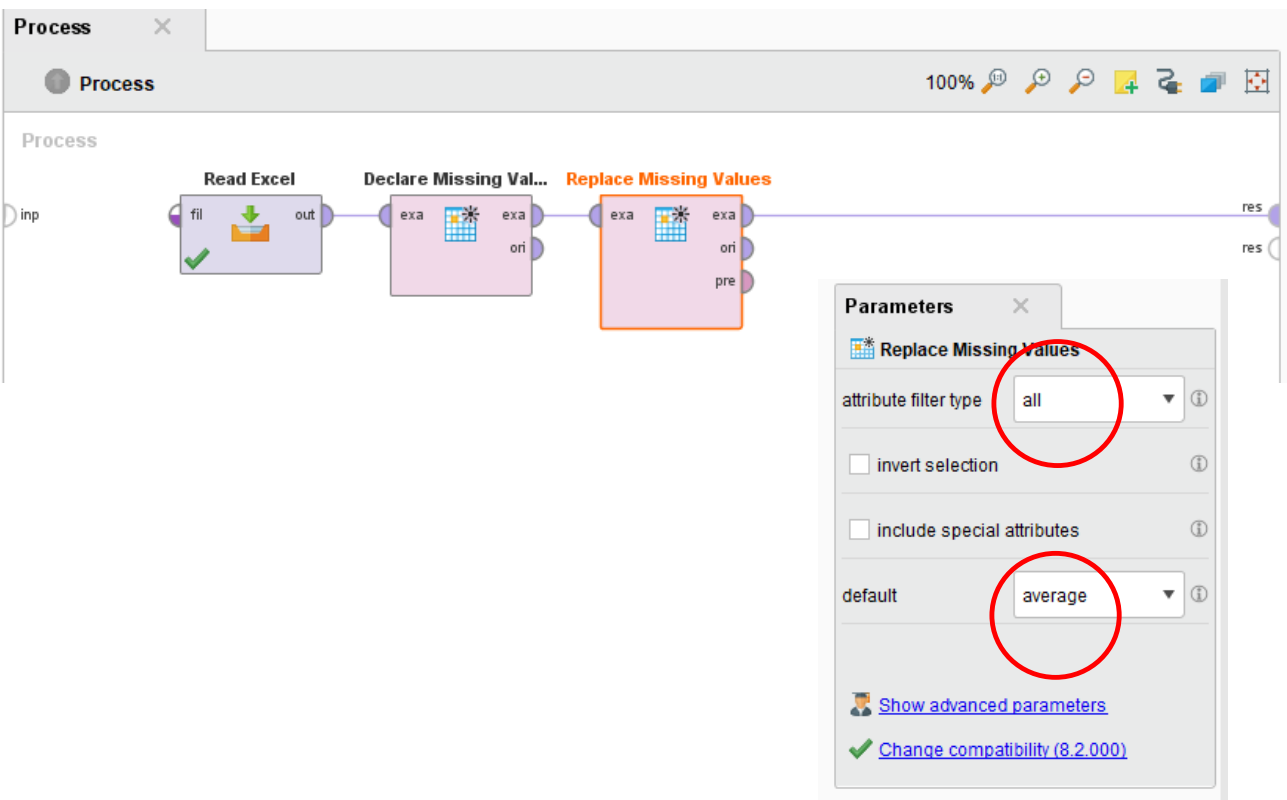
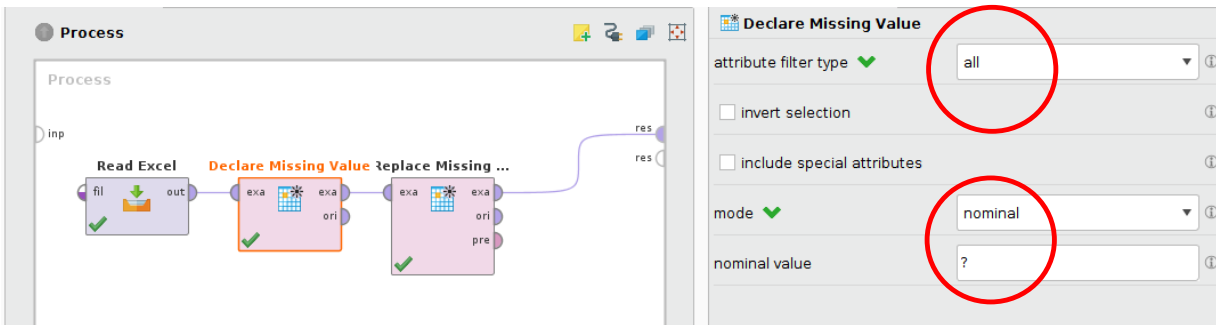


**Per tornare nel processo principale, cliccare su design.**

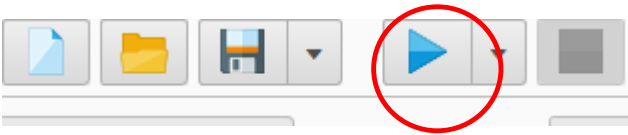
### Obiettivo 2 – Gestione dei dati mancanti

Verificare la presenza di eventuali dati mancanti e gestirli con opportuni passi di trasformazione (operatori *Declare 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**.



Per lanciare un processo in RapidMiner usare il triangolo in alto nella barra dei processi.

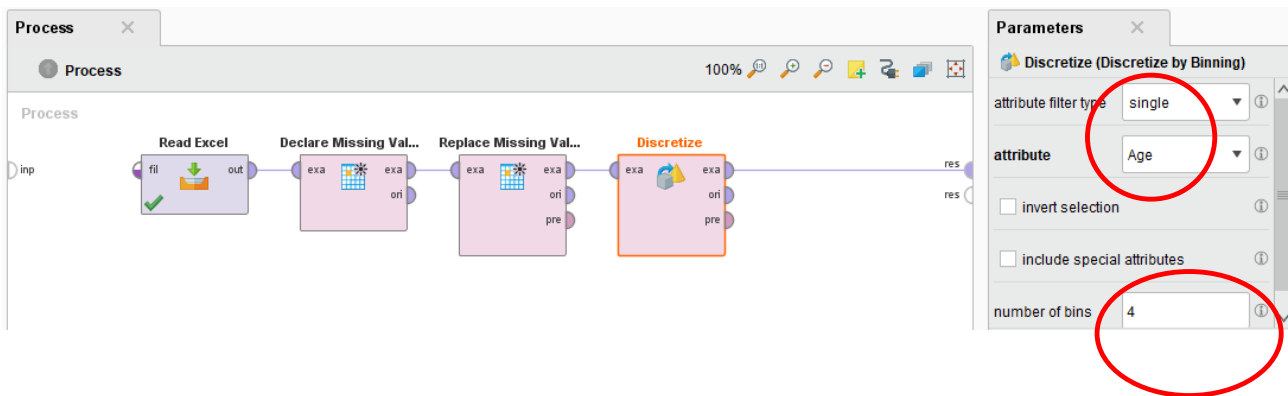


### Obiettivo 3 – Discretizzazione

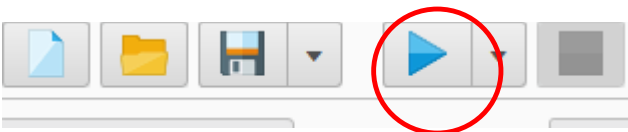
Verificare la presenza di attributi continui nei dati di origine.

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*, *Discretize by entropy*) e confrontare i risultati.



Per lanciare un processo in RapidMiner usare il triangolo in alto nella barra dei processi.



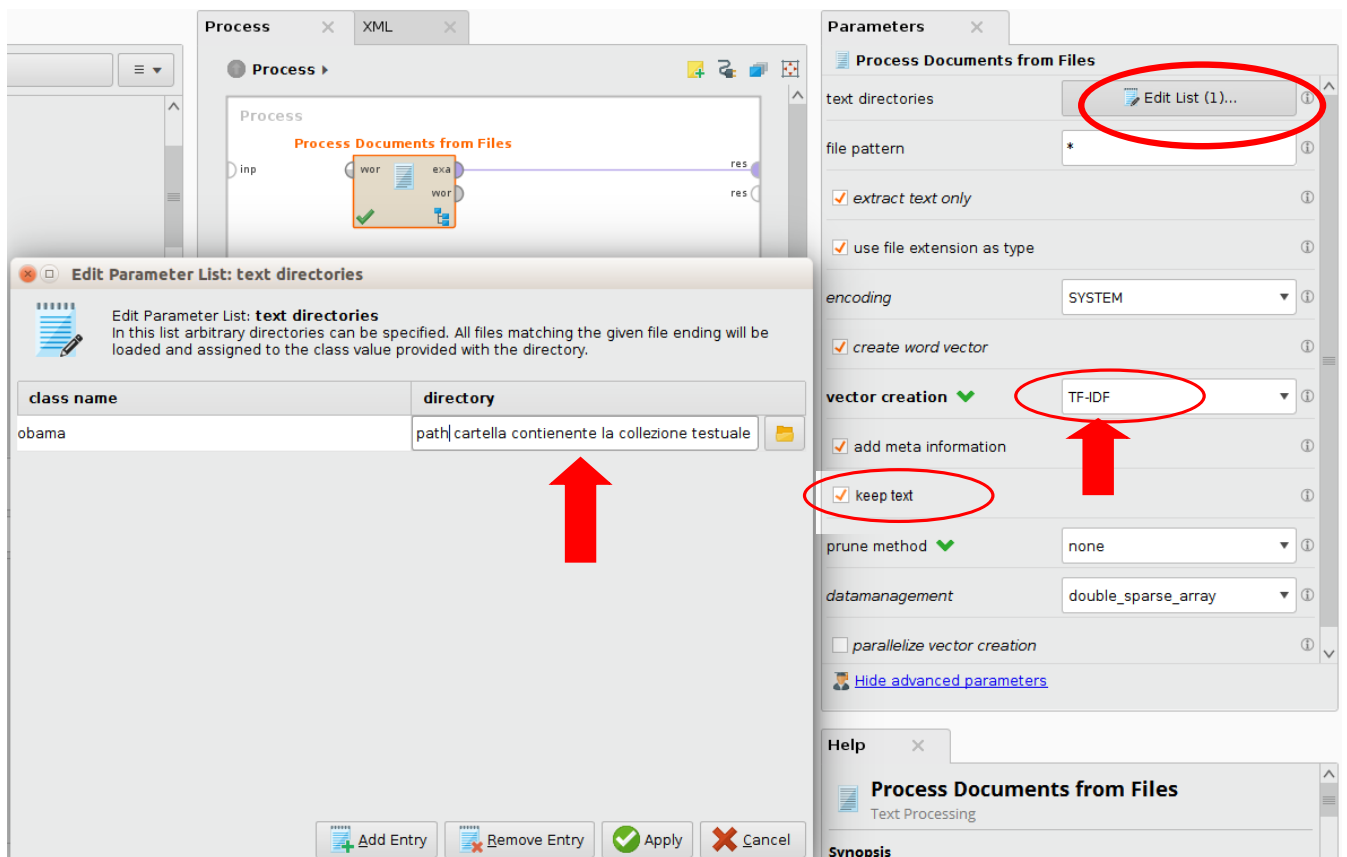
Analizzare i risultati della discretizzazione scelta.

## Dati testuali

Il dataset denominati ObamaNews (ObamaNews.zip) è disponibile sul sito del corso (<http://dbdmg.polito.it/wordpress/teaching/business-intelligence/>). Esso contiene una collezione di news scaricate mediante il servizio Google News. La collezione rappresenta l'insieme delle prime 10 news (pagine contenenti notizie) restituite da Google News a fronte della specifica della parola chiave *Obama*.

### Obiettivo 1 – Import dei dati

Importare il dataset ObamaNews in Rapid Miner (operatore *Process Documents From Files*).



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**.

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.

$$tfidf_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

$tf_{i,j}$  = total number of occurrences of  $i$  in  $j$   
 $df_i$  = total number of documents (speeches) containing  $i$   
 $N$  = total number of documents (speeches)

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.

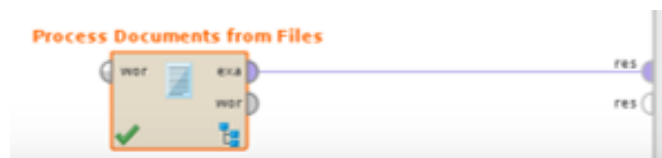
Prima di creare la matrice, guardare l'output. Nel sottoprocesso (per entrare nel sottoprocesso, fare doppio click con il tasto sinistro sull'operatore *Process Document from Files*) collegare le uscite come in figura.



Per tornare al processo principale, cliccare la freccia blu come in figura.

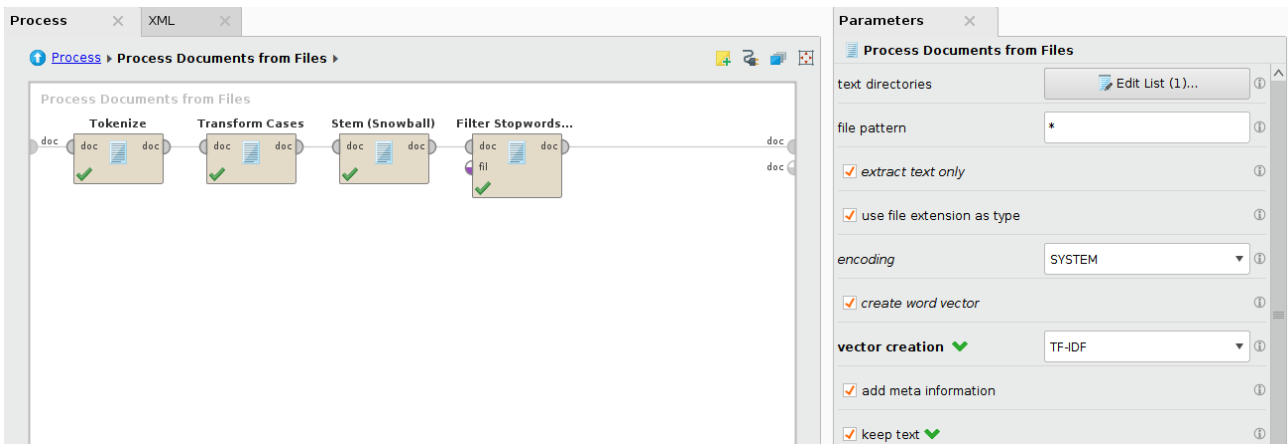


Collegare l'uscita exa con res ed eseguire il processo.





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:

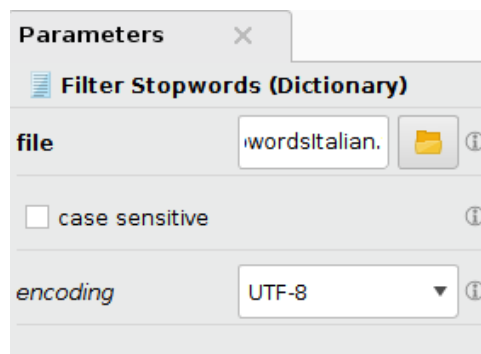


Il blocchetto **Tokenize**: splitta ogni documento della collezione Obama in un vettore di parole. L'ordine delle parole non sarà più rispettato. Secondo te ha importanza ai fini dell'analisi? (Settare il parametro non letters).

Il blocchetto **Transform Cases**: Trasforma il testo in maiuscolo o minuscolo.

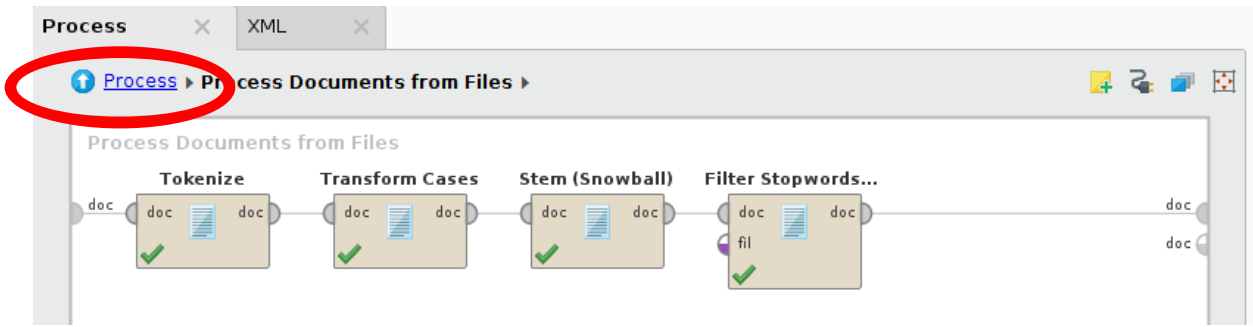
Il blocchetto **Stem (Snowball)**: Riduce le parole alla propria radice. La radice è quell'elemento linguistico irriducibile (non ulteriormente suddivisibile) che esprime il significato principale della parola. (Utilizzare la lingua italiana).

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 **stopwordsItalian.txt** presente sul sito del corso.



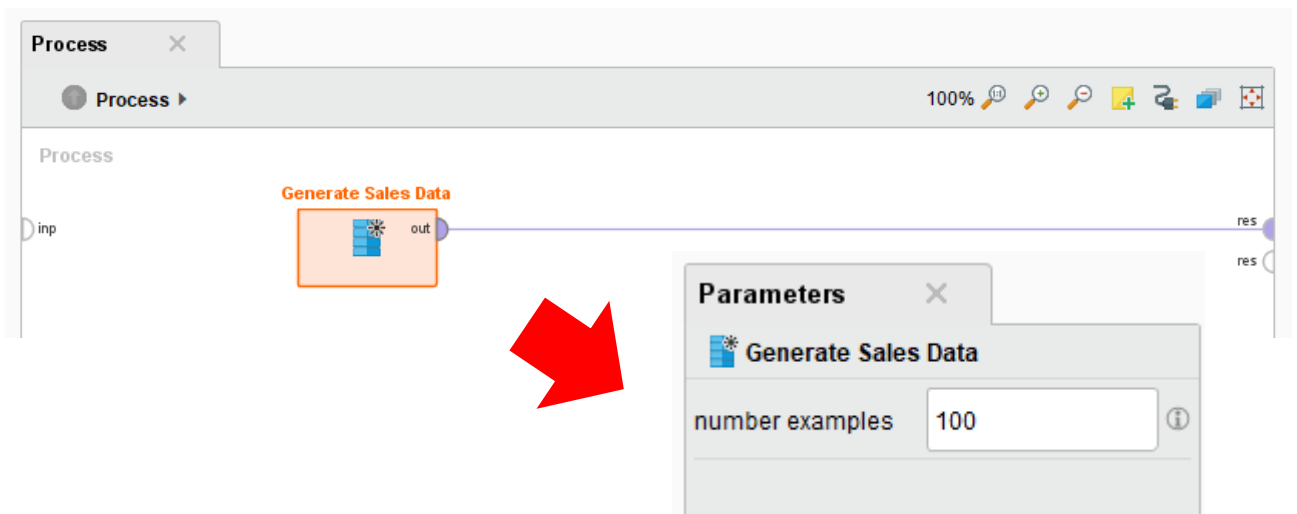
Utilizzare la codifica UTF-8 per il file delle stopwords.

Torna al processo iniziale cliccando sulla freccia blu sotto la voce Process.

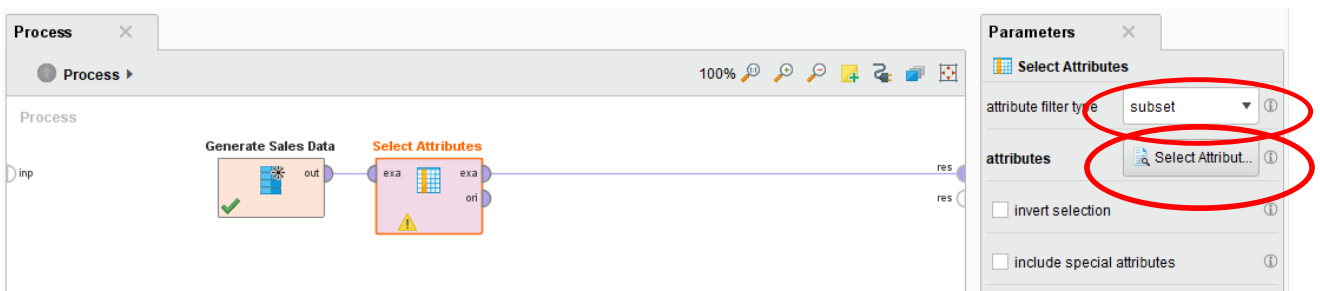


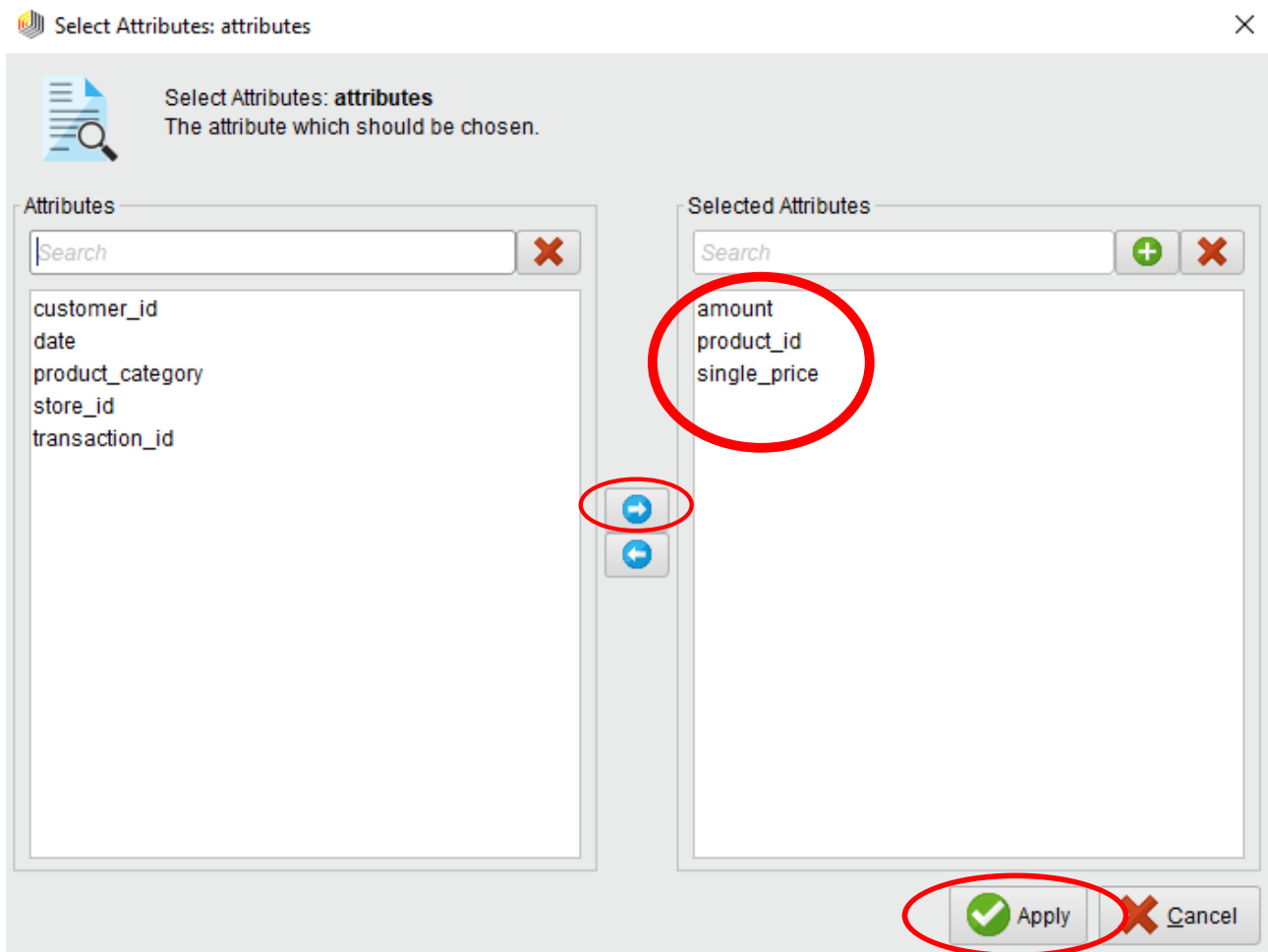
## Bonus: dati strutturati con attributi continui

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



Selezionare solo gli attributi numerici. Utilizzare l'operatore Select Attributes.

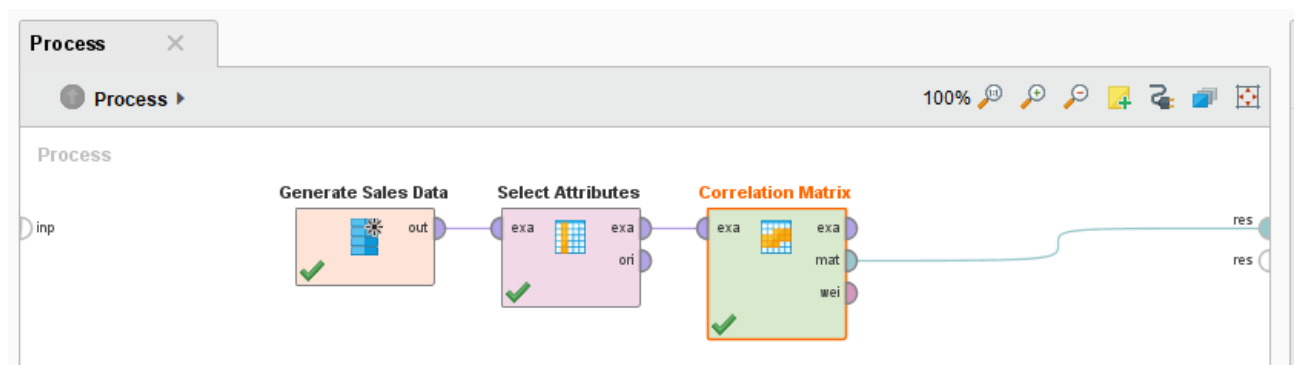




Eseguire il processo e analizzare i risultati.

### Obiettivo 1 – Analisi delle correlazioni

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:

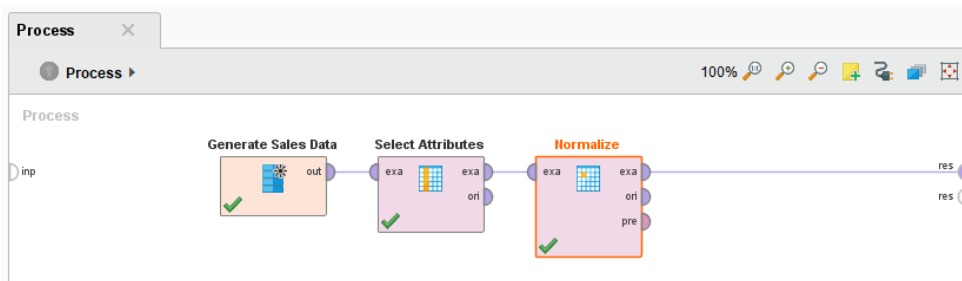


Esiste qualche correlazione elevata?

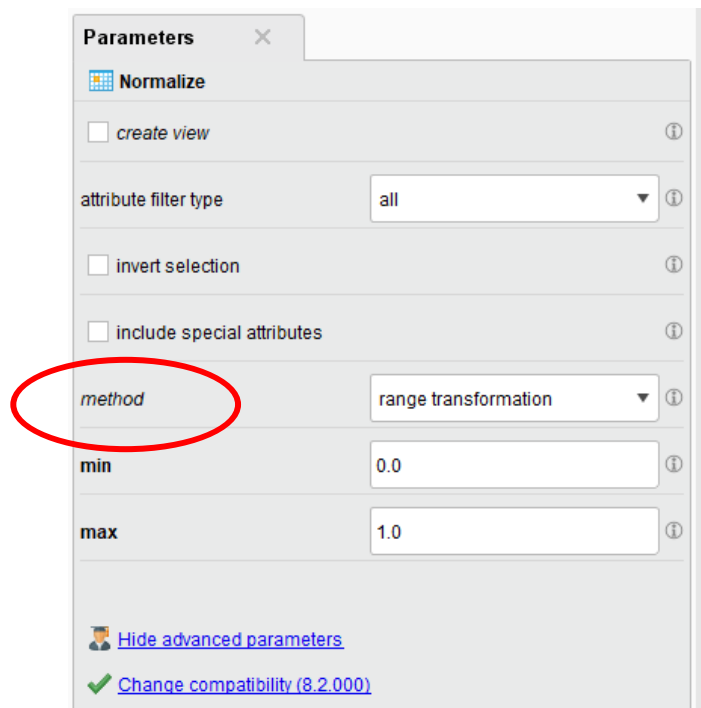
Eliminare l'operatore correlation matrix (selezionarlo con il mouse e premere canc).

## Obiettivo 2 – Normalizzazione

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*).



Quale normalizzazione scegliereste? (nel caso non compaia method, cliccare su [Show advanced parameters](#))



Quando è utile normalizzare i dati?