Business Intelligence per i Big Data

Esercitazione di laboratorio N. 2

Gli obiettivi dell'esercitazione sono:

- **utilizzare i principali algoritmi di estrazione di pattern frequenti disponibili in RapidMiner ed estrarre regole di associazione per dataset strutturati e non strutturati.**

Dati strutturati

Abbiamo a disposizione le playlist di un sottoinsieme di utenti del sistema Spotify. Ogni playlist è composta da un insieme di brani musicali. Ogni playlist è stata definita manualmente da un utente di Spotify. Le informazioni sulle playlist sono memorizzate in un file Excel (in una situazione reale tali informazioni sono memorizzate in una base di dati). Il file si chiama Playlists_Spotify.xlsx.
La prima riga di Playlists_Spotify.xlsx contiene i nomi delle canzoni che fanno parte di almeno una playlist, mentre le righe successive contengono le playlist degli utenti (ogni riga contiene una playlist). In ogni riga la cella relativa a un brano musicale assume il valore "*true*" o "*false*" in funzione del fatto che in quella playlist ci sia oppure no il brano musicale cui è associata la colonna.

Dati testuali

Il dataset denominati ObamaNews (ObamaNews.zip) 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*.

Caratterizzazione di playlist del sistema Spotify

Obiettivo 1 – Analisi esplorativa "manuale" delle playlist

Il primo obiettivo del laboratorio si focalizza su un'analisi preliminare dei dati disponibili. In particolare, dovete creare un processo RapidMiner che legge il file excel e provare a rispondere alle seguenti domande preliminari sulle caratteristiche dei dati:

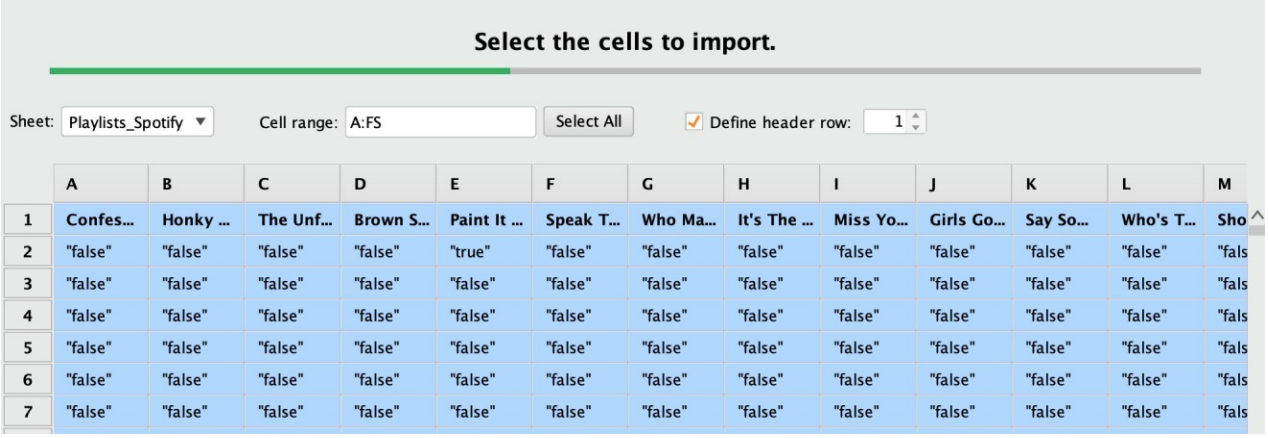
- Quante playlist sono disponibili?
- Quanti brani musicali sono presenti?
- Qual è il brano musicale più popolare?

Passi per creare il processo di RapidMiner:

- Creare un nuovo processo vuoto cliccando sul pulsante  in alto a sinistra.
- Selezionare il componente "Read Excel" tramite la casella di ricerca presente in alto a sinistra (digitare "Read Excel") e trascinarlo sull'area di lavoro centrale.
- Connettere l'uscita "out" del componente "Read Excel" con il connettore **res** del processo che si sta realizzando.

- Configurare il componente “Read Excel” cliccando sul pulsante 

    - Selezionare il file *Playlists_Spotify.xlsx* e premere Next.
    - Appena compare il contenuto del file premere nuovamente Next.
    - Verificare che la prima riga visualizzata contenga i nomi dei brani musicali e procedere premendo Next.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Confes...	Honky ...	The Unf...	Brown S...	Paint It ...	Speak T...	Who Ma...	It's The ...	Miss Yo...	Girls Go...	Say So...	Who's T...	Sho
2	"false"	"false"	"false"	"false"	"true"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"fals
3	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"fals
4	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"fals
5	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"fals
6	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"fals
7	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"false"	"fals

    - Se il numero di errori è pari a 0, premere Finish per terminare la procedura di configurazione del componente “Read Excel”.

Il processo generato è il seguente:

Ora potete eseguire il processo che avete appena creato premendo la freccia blu presente nella barra dei comandi.

Analizzando l’output del processo appena realizzato, provate a rispondere alle seguenti domande:

- Quante playlist sono disponibili?
- Quanti brani musicali sono presenti?
- Qual è il brano musicale più popolare?

Riuscite a rispondere a tutte le domande?

Obiettivo 2 – Analisi esplorativa delle playlist per l’identificazione di brani popolari e combinazioni ricorrenti di brani

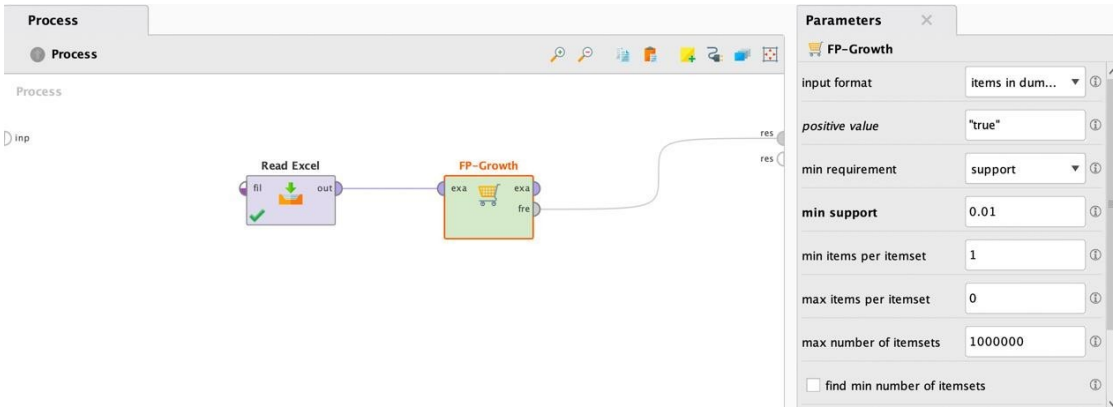
Siamo interessati a capire con quale frequenza si presentano i singoli brani musicali nelle playlist che stiamo analizzando e quali sono le combinazioni di brani musicali (combinazioni composte da due o più brani) che compaiono frequentemente insieme e di conseguenza cercare di capire quali sono le preferenze degli utenti di Spotify che stiamo analizzando.

Usiamo gli itemset frequenti oppure le regole di associazione per calcolare le frequenze/occorrenze dei brani musicali e delle loro combinazioni?

Passi per risolvere il problema con RapidMiner:

- Selezionare il componente “FP-Growth”digitando “FP-Growth” nella casella operators e trascinarlo sull’area di lavoro centrale.
- Connettere l’uscita “out” del componente “Read Excel” con il connettore di input “exa” del componente “FP-Growth”.
- Connettere l’uscita “fre” del componente “FP-Growth” con il connettore “res” del processo che si sta realizzando.
- Configurare il componente “FP-Growth”
  - Disabilitare l’opzione “find min number of itemsets”.
  - Impostare il parametro “positive value” al valore "true" (doppi apici inclusi).
  - Impostare il supporto minimo (parametro “min support”) al valore 0.01 (ossia al valore 1%).

Il processo generato è il seguente:



Ora potete eseguire il processo che avete appena creato premendo la freccia blu presente nella barra dei comandi.

Analizzando l’output del processo appena realizzato, provate a rispondere alle seguenti domande:

- Quali sono i due brani musicali più popolari? I parametri “Min. Size” e “Max. Size” permettono di focalizzarsi su itemset con lunghezze specifiche, se necessario.
- In quante playlist è presente il brano ***Never Say Never***? Il parametro “Contains Item” permette di selezionare solo gli itemset che contengono l’item specificato (ossia il brano specificato nel nostro caso). Attenzione che il sistema è *case sensitive* e quindi maiuscole e minuscole sono considerate simboli diversi tra loro.
- Qual è il brano più popolare tra **Never Say Never** e **All My Life**?
- Quali sono i brani che compaiono frequentemente insieme a **All My Life**?
- Quali sono i brani che compaiono frequentemente insieme a **Never Say Never**?
- Quanti e quali brani dei “Foo Fighters” sono popolari nelle playlist che stiamo analizzando?
- Quanti e quali brani di “Justin Bieber” sono popolari nelle playlist che stiamo analizzando?
- In quante playlist è presente il brano “Somebody To Love” di Justin Bieber? Cosa dobbiamo modificare nel processo per cercare di rispondere a questa domanda?

- Un nuovo utente di Spotify vuole creare la sua prima playlist. Quali brani gli suggerireste di inserire supponendo che abbia gusti simili alla maggioranza degli altri utenti? Vi servono altre informazioni per fornire un buon suggerimento?

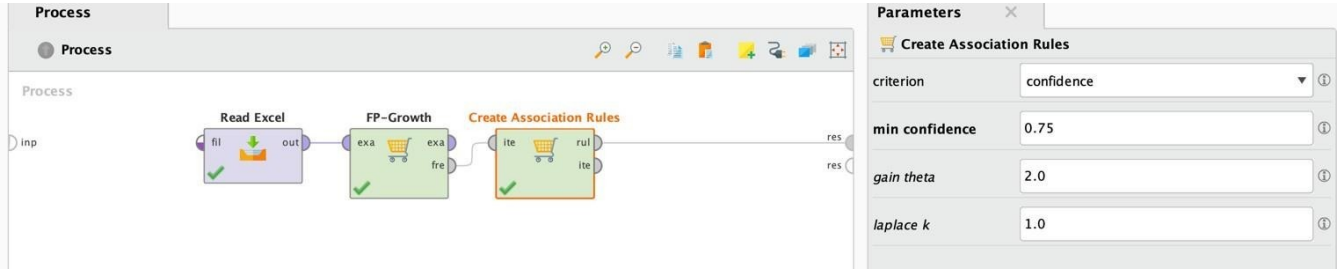
Obiettivo 3 – Semplice sistema di raccomandazioni di brani musicali

Il terzo obiettivo dell'esercitazione consiste nel cercare di suggerire nuovi bravi musicali a utenti che hanno già definito delle loro playlist. In particolare, data una playlist si vuole suggerire all'utente che ha definito tale playlist come estenderla con ulteriori brani musicali che potrebbero essere di suo interesse basandosi sulle playlist definite dagli altri utenti.

Usiamo gli itemset frequenti oppure le regole di associazione per risolvere questo problema?

Passi per risolvere il problema con RapidMiner:

- Estendere il processo precedente aggiungendo alla fine del processo il componente "Create Association Rules".
    - Connettere l'uscita "fre" del componente "FP-Growth" con l'input "ite" del componente "Create Association Rules".
    - Connettere l'uscita "rul" del componente "Create Association Rules" con il connettore "res" del processo che si sta realizzando.
    - Configurare il componente "Create Association Rules" ○ Impostare il parametro "min confidence" al valore 0.75 (ossia al valore 75%). Il processo generato è il seguente



Ora potete eseguire il processo che avete appena creato premendo la freccia blu presente nella barra dei comandi.

Analizzando l'output del processo appena realizzato, provate a rispondere alle seguenti domande:

- A quali playlist pensate sarebbe potenzialmente interessante aggiungere la canzone "Sweet Dreams" di Beyonce perché potenzialmente affine agli altri brani già presenti nella playlist?
- A quali playlist aggiungereste "All My Life" dei Foo Fighters?
- A chi suggerireste "Never Say Never"?
- Provate a rieseguire il processo impostando "min confidence" del componente "Create Association Rules" al valore 0.15 (ossia 15%) e provate a rispondere nuovamente alla precedente domanda.

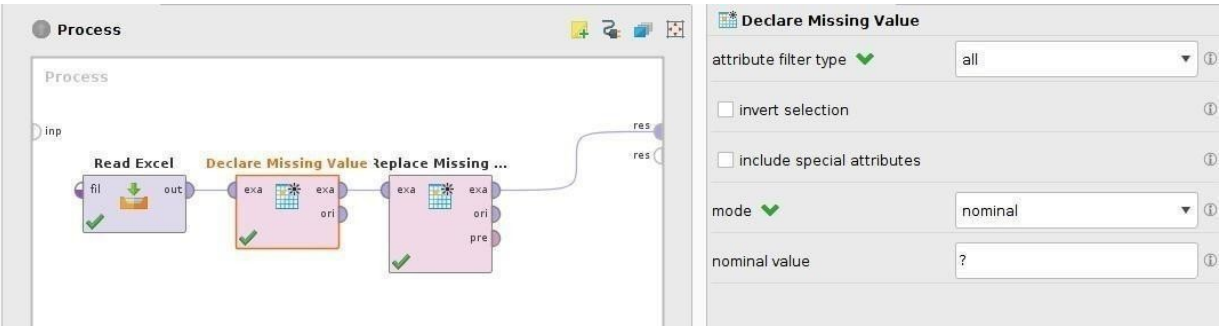
Estrazione di associazioni da dati strutturati

L'obiettivo dell'analisi dei dati è sfruttare i dati relativi alle persone contattate per capire quali loro caratteristiche (anagrafiche, lavorative o una combinazione delle precedenti) è maggiormente correlata con la risposta (campo *Response*). Si vuole rispondere alle seguenti richieste:

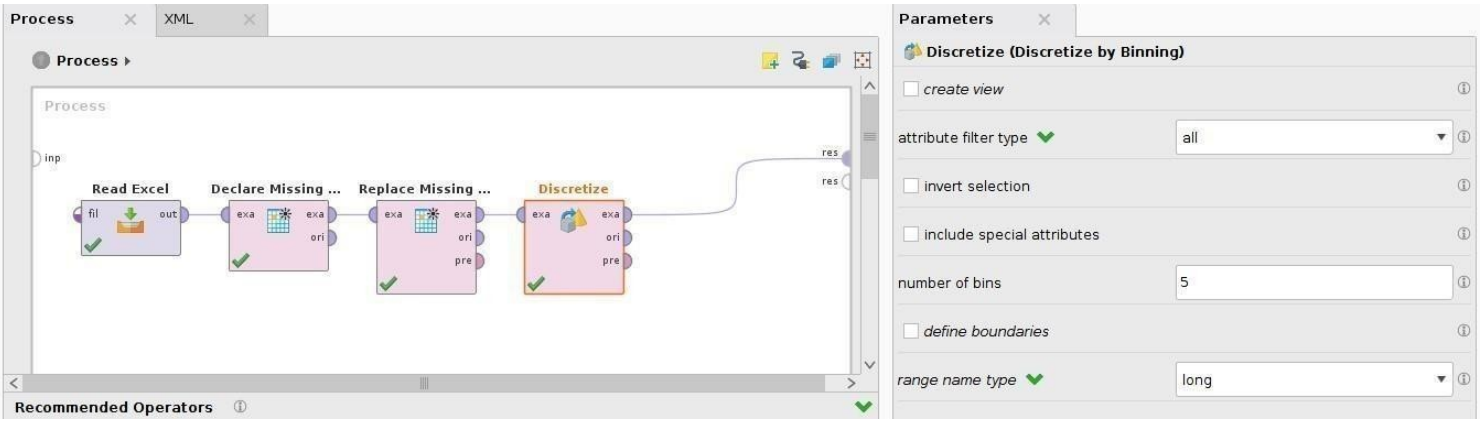
1. *Visualizzare le combinazioni di nazionalità e fascia d'età maggiormente correlate con la risposta Negative. Per esempio, un'analista potrebbe scoprire che le persone di età compresa tra i 30 e i 40 anni di nazionalità brasiliana sono fortemente correlate con la risposta Negative.*

Passi da svolgere con RapidMiner:

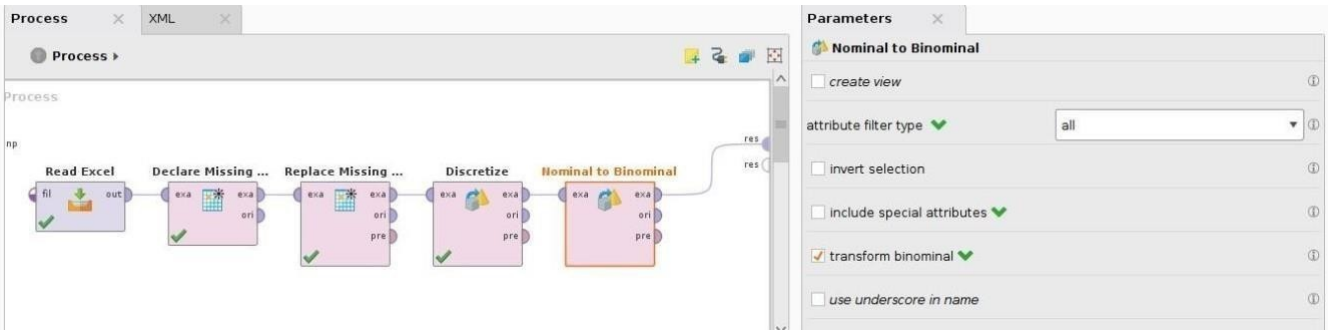
- Nel pannello **Operators** cercare l'operatore **Read Excel** e trascinarlo nello spazio di lavoro.
- Importare il dataset ***UsersSmall.xls*** utilizzando la procedura guidata ***Import Configuration Wizard***.
- Dichiarare per tutti gli attributi il '**?**' come valore NULL attraverso l'operatore **Declare Missing Value**.
- Sostituire i valori nulli dichiarati al punto precedente con il valore più frequente usando l'operatore **Replace Missing Values**.



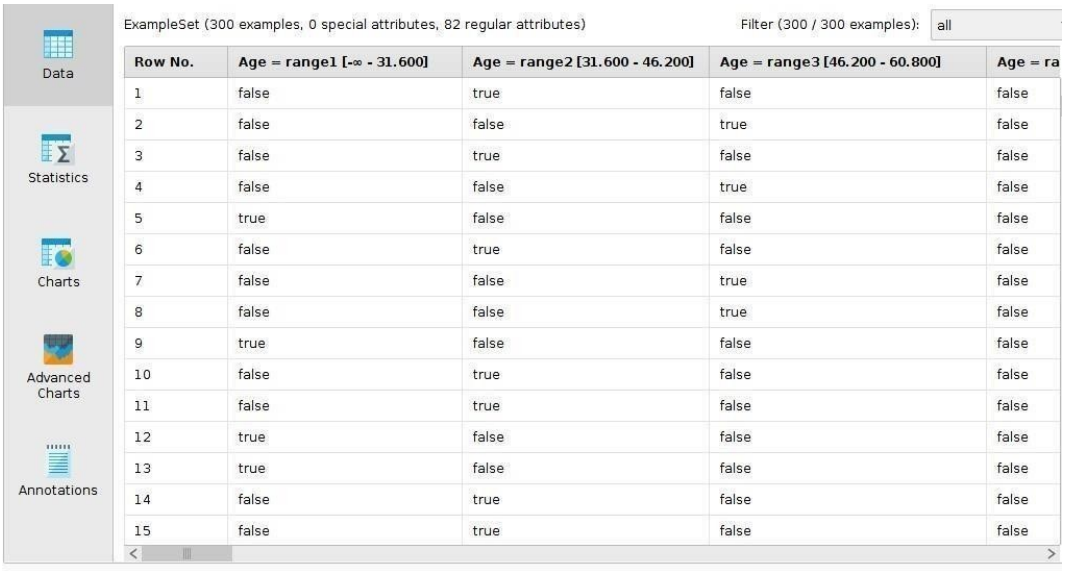
- **Discretizzare** opportunamente l'attributo *Age*. (**N.B.** In RapidMiner esistono diversi metodi per discretizzare un attributo numerico. Quale useresti in questo caso?)



- Trasformare i valori nominali in binomiali (operatori ***Nominal to Binomial***)

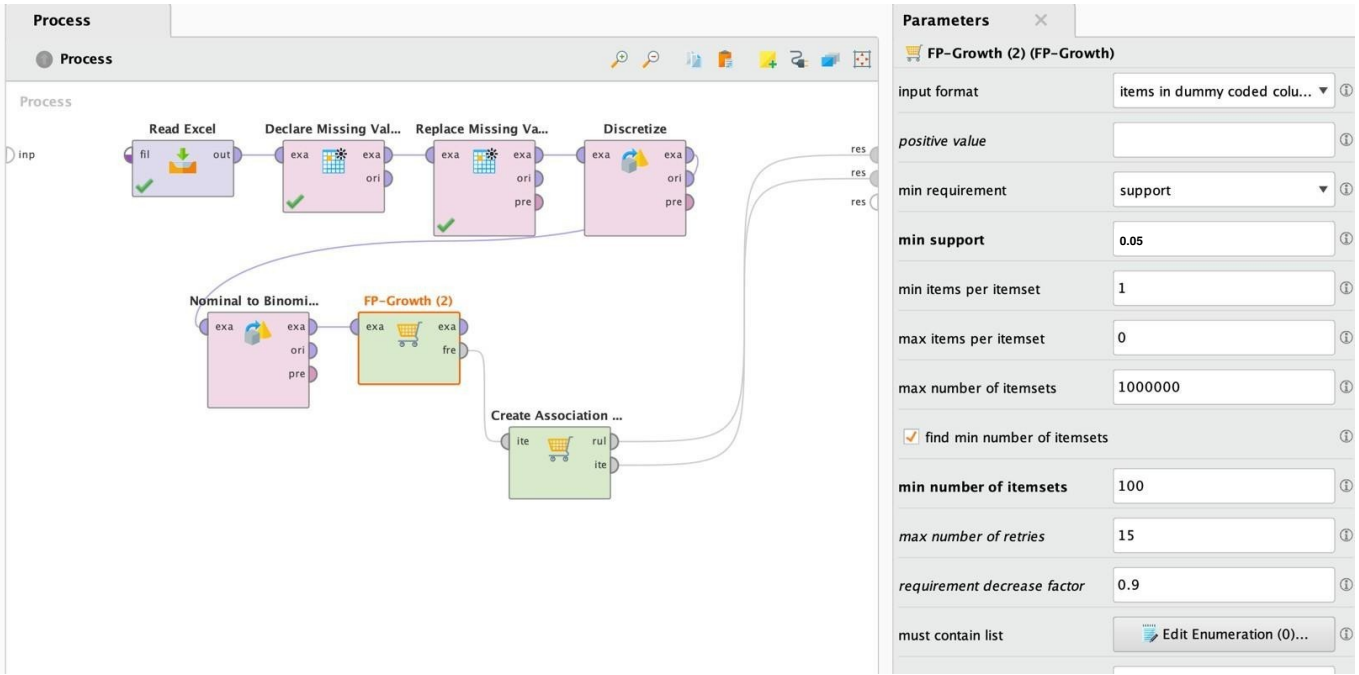


In questo modo trasformerete ogni record del file in coppie “Attributo” = ”Valore”. Esegui il processo e osserva la matrice dei dati. Sarà l’input da dare all’algoritmo per estrarre le regole di associazione.



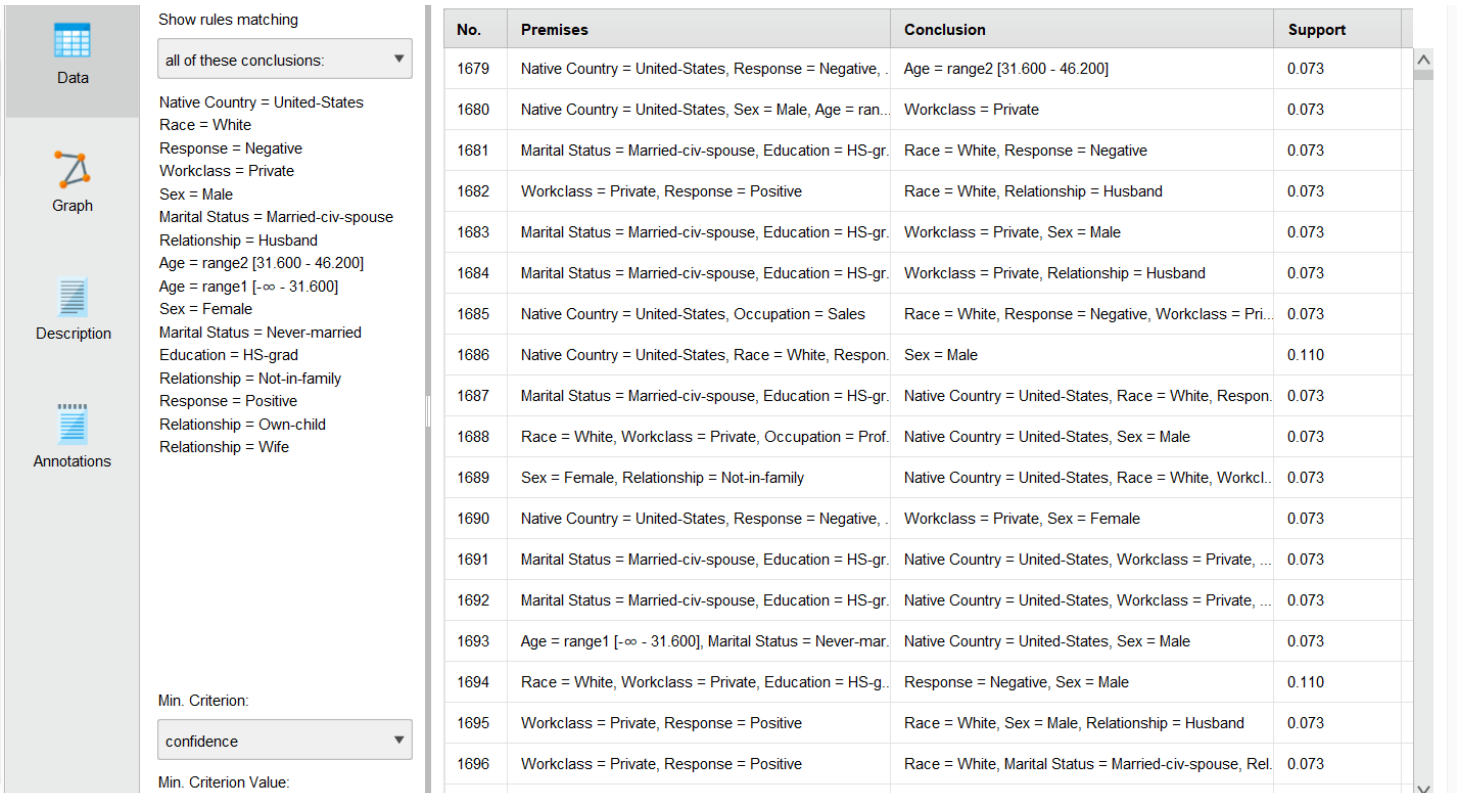
Row No.	Age = range1 [-∞ - 31.600]	Age = range2 [31.600 - 46.200]	Age = range3 [46.200 - 60.800]	Age = ra
1	false	true	false	false
2	false	false	true	false
3	false	true	false	false
4	false	false	true	false
5	true	false	false	false
6	false	true	false	false
7	false	false	true	false
8	false	false	true	false
9	true	false	false	false
10	false	true	false	false
11	false	true	false	false
12	true	false	false	false
13	true	false	false	false
14	false	true	false	false
15	false	true	false	false

- Usare l’algoritmo FP-Growth per l’estrazione gli itemset frequenti (operatore *FP-Growth*) e analizzare i risultati con **differenti valori di supporto minimo**.



- Generare le regole di associazione dagli itemset frequenti e analizzare i risultati con differenti valori di **confidenza**.

Analizzare le diverse regole che vengono estratte dall'algoritmo. Come deve essere il lift per considerare una regola interessante?



No.	Premises	Conclusion	Support
1679	Native Country = United-States, Response = Negative, .	Age = range2 [31.600 - 46.200]	0.073
1680	Native Country = United-States, Sex = Male, Age = ran...	Workclass = Private	0.073
1681	Marital Status = Married-civ-spouse, Education = HS-gr.	Race = White, Response = Negative	0.073
1682	Workclass = Private, Response = Positive	Race = White, Relationship = Husband	0.073
1683	Marital Status = Married-civ-spouse, Education = HS-gr.	Workclass = Private, Sex = Male	0.073
1684	Marital Status = Married-civ-spouse, Education = HS-gr.	Workclass = Private, Relationship = Husband	0.073
1685	Native Country = United-States, Occupation = Sales	Race = White, Response = Negative, Workclass = Pri...	0.073
1686	Native Country = United-States, Race = White, Respon.	Sex = Male	0.110
1687	Marital Status = Married-civ-spouse, Education = HS-gr.	Native Country = United-States, Race = White, Respon.	0.073
1688	Race = White, Workclass = Private, Occupation = Prof.	Native Country = United-States, Sex = Male	0.073
1689	Sex = Female, Relationship = Not-in-family	Native Country = United-States, Race = White, Workcl..	0.073
1690	Native Country = United-States, Response = Negative, .	Workclass = Private, Sex = Female	0.073
1691	Marital Status = Married-civ-spouse, Education = HS-gr.	Native Country = United-States, Workclass = Private, ...	0.073
1692	Marital Status = Married-civ-spouse, Education = HS-gr.	Native Country = United-States, Workclass = Private, ...	0.073
1693	Age = range1 [-∞ - 31.600], Marital Status = Never-mar.	Native Country = United-States, Sex = Male	0.073
1694	Race = White, Workclass = Private, Education = HS-g..	Response = Negative, Sex = Male	0.110
1695	Workclass = Private, Response = Positive	Race = White, Sex = Male, Relationship = Husband	0.073
1696	Workclass = Private, Response = Positive	Race = White, Marital Status = Married-civ-spouse, Rel.	0.073

Se siete interessati a una particolare conclusione, potete selezionare nel riquadro "all of these conclusions" solo quella che volete analizzare.

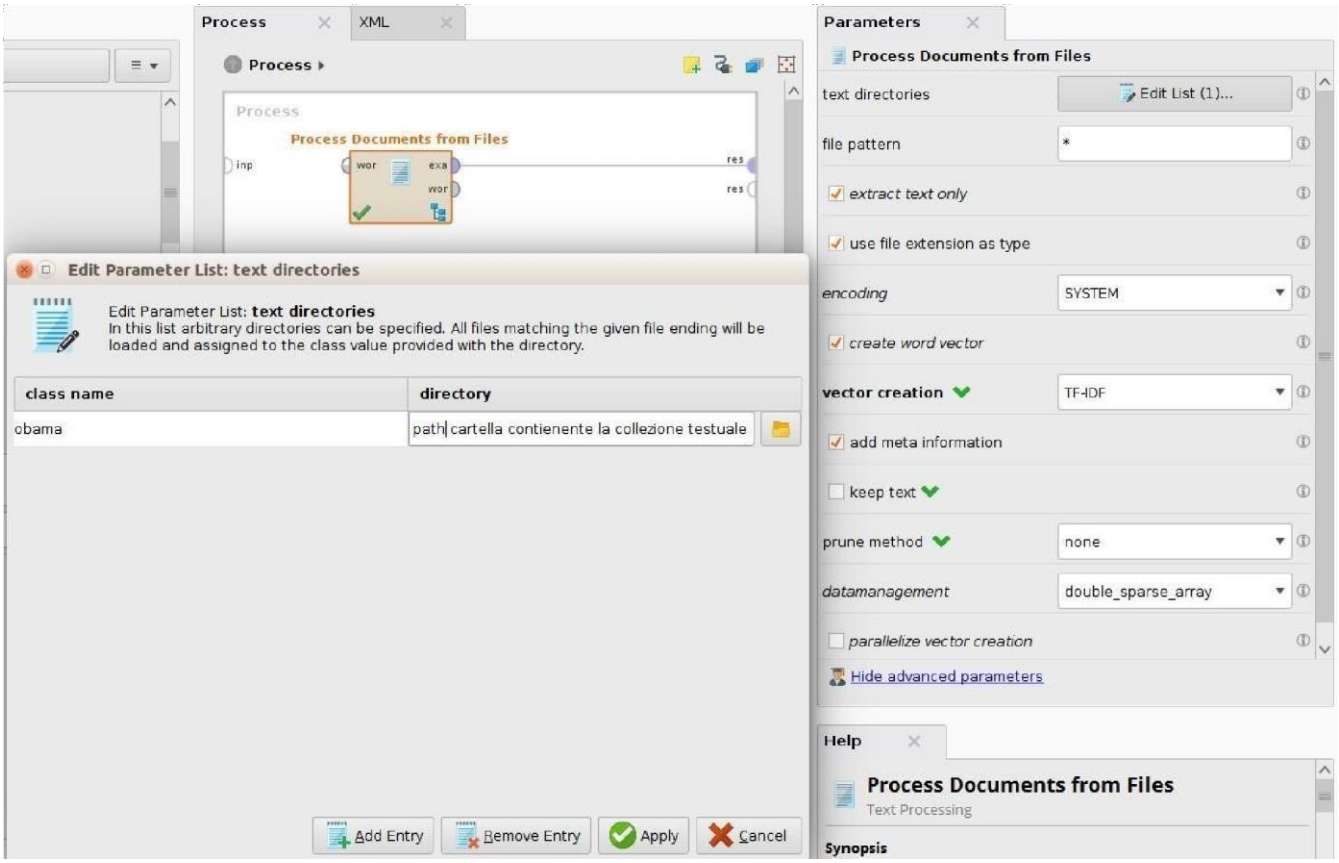
- Quale titolo di studio è maggiormente correlato con "Response=Positive"?
- Quale range di età e quale paese di origine sono maggiormente correlati con "Response=Negative"?

Estrazione di associazioni da dati testuali

L'obiettivo dell'analisi dei dati è scoprire le correlazioni tra termini nascoste all'interno della collezione ObamaNews. Si vuole rispondere alle seguenti richieste:

1. *Identificare i termini e le coppie di termini più ricorrenti.*
2. *Identificare i termini maggiormente correlati tra di loro.*

- Trasforma la collezione di documenti nella matrice **document*term**. Per fare questo passaggio utilizza l'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.

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

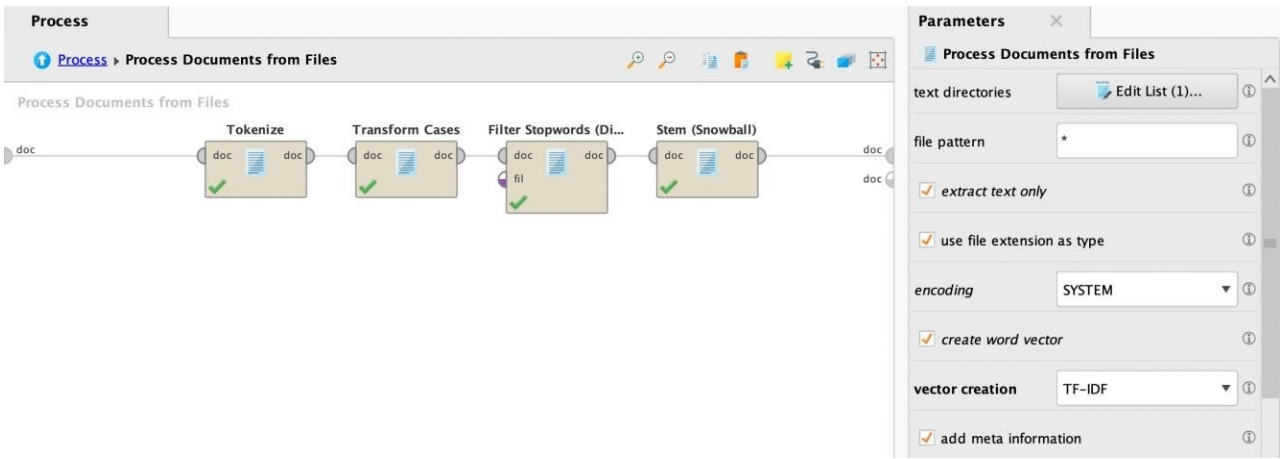
$\mathrm{tf}_{i,j}$ = total number of occurences of i in j

df_i = total number of documents (speeches) containing i

N = total number of documents (speeches)

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**.

- 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:



L'operatore **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).

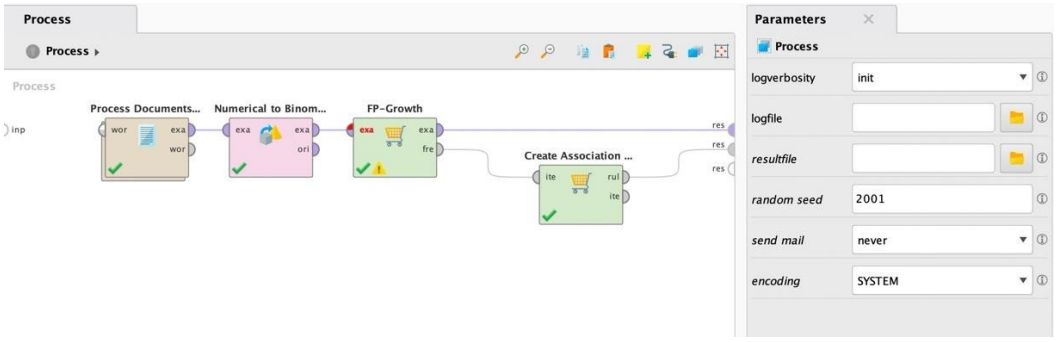
L'operatore **Transform Cases**: Trasforma il testo in maiuscolo o minuscolo.

L'operatore **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. (Utilizzarela lingua italiana).

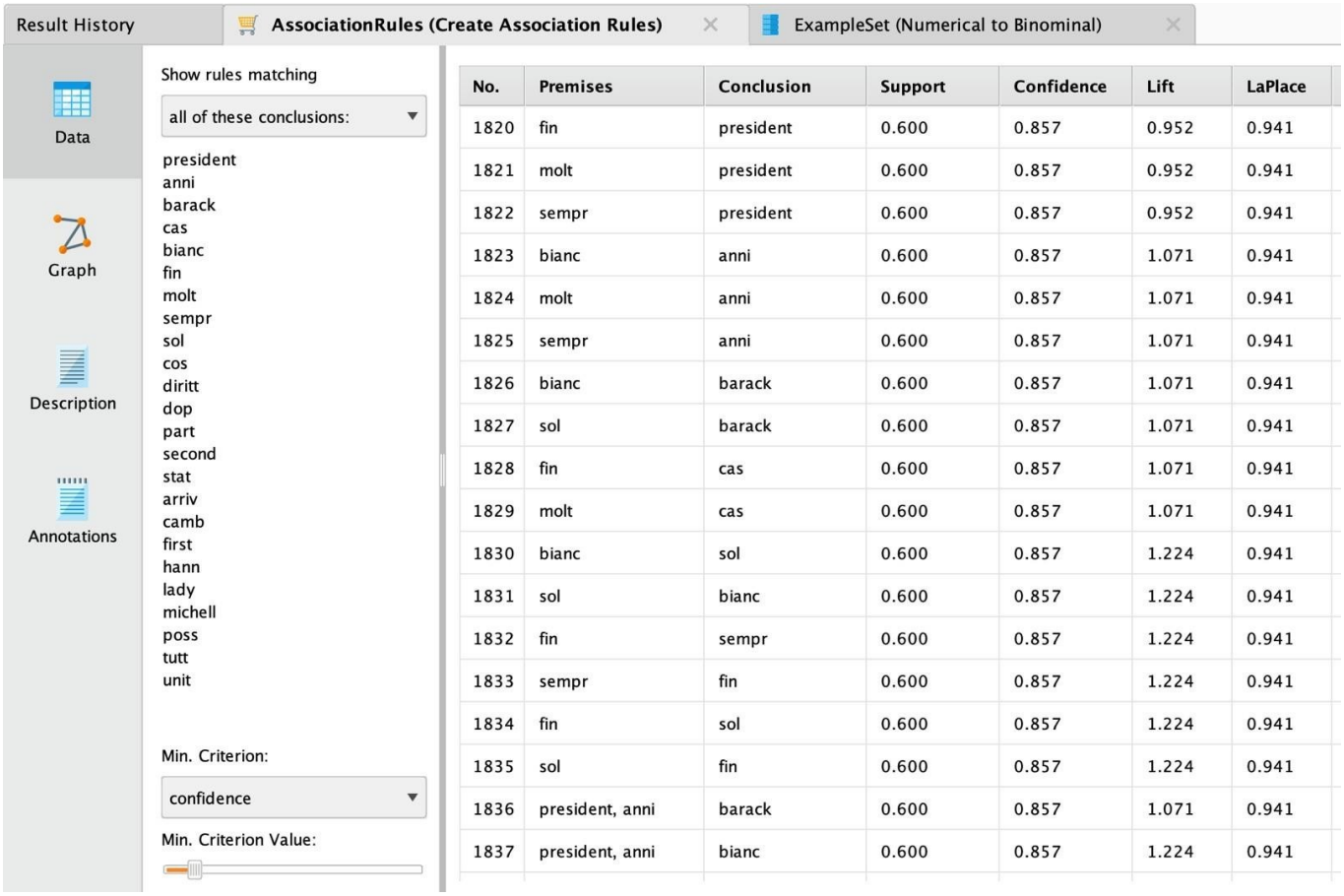
L'operatore **Filter Stopwords (Dictionary)**: Permette di eliminare le parole definite Stopword, parole che non hanno un particolare significato se isolate dal testo e quindi vengono ignorate dai programmi. Sono parole poco significative perché possono essere usate spesso all' interno delle frasi. Ad 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 stopword.

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

- Trasformare i valori ottenuti con il TF-IDF in valori binomiali (operatore ***Numerical to Binomial***)



- *Utilizzare l'**FP-Growth** per l'estrazione dei pattern frequenti e l'operatore **Create Association Rules** per estrarre le regole. Come per l'esempio precedente analizza le correlazioni più interessanti.*



No.	Premises	Conclusion	Support	Confidence	Lift	LaPlace
1820	fin	president	0.600	0.857	0.952	0.941
1821	molt	president	0.600	0.857	0.952	0.941
1822	sempr	president	0.600	0.857	0.952	0.941
1823	bianc	anni	0.600	0.857	1.071	0.941
1824	molt	anni	0.600	0.857	1.071	0.941
1825	sempr	anni	0.600	0.857	1.071	0.941
1826	bianc	barack	0.600	0.857	1.071	0.941
1827	sol	barack	0.600	0.857	1.071	0.941
1828	fin	cas	0.600	0.857	1.071	0.941
1829	molt	cas	0.600	0.857	1.071	0.941
1830	bianc	sol	0.600	0.857	1.224	0.941
1831	sol	bianc	0.600	0.857	1.224	0.941
1832	fin	sempr	0.600	0.857	1.224	0.941
1833	sempr	fin	0.600	0.857	1.224	0.941
1834	fin	sol	0.600	0.857	1.224	0.941
1835	sol	fin	0.600	0.857	1.224	0.941
1836	president, anni	barack	0.600	0.857	1.071	0.941
1837	president, anni	bianc	0.600	0.857	1.224	0.941