

**Management delle informazioni e
gestione della conoscenza
AA 2021-22**

Orange Data Miner

Roberto Boselli
roberto.boselli@unimib.it

Data Mining Tools

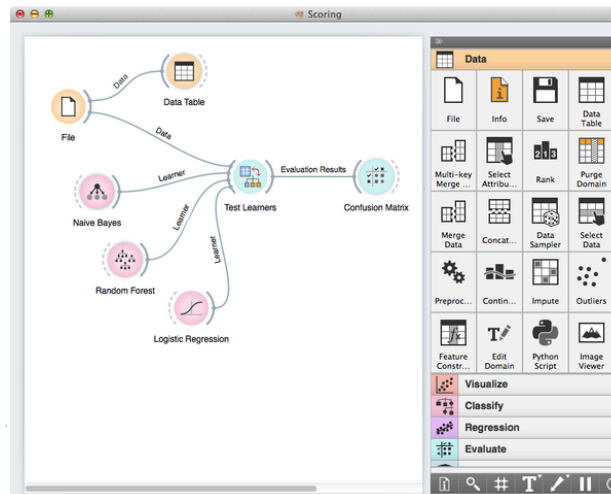
	Caratteristiche	Linguaggio di programmazione	Sistemi operativi	Costi/licenze
RapidMiner	Strumento completo e potente, eccelle soprattutto nell'analisi predittiva	Java	Windows, macOS, Linux	Freeware, diversi versioni a pagamento
WEKA	Numerosi metodi di classificazione	Java	Windows, macOS, Linux	Software libero (GPL)
Orange	Crea visualizzazioni dei dati particolarmente accattivanti e interessanti senza che siano necessarie molte preconcoscenze	Base del software: C++, estensioni e linguaggio per l'accesso ai dati: Python	Windows, macOS, Linux	Software libero (GPL)
KNIME	Il leader del settore tra i tool open source di data mining, che ha reso universalmente accessibile l'analisi predittiva	Java	Windows, macOS, Linux	Software libero (GPL) (a partire dalla versione 2.1)
SAS	Il software di data mining più potente, anche se costoso, adatto per le grandi aziende	SAS Language	Windows, macOS, Linux	Versione limitata freeware per gli istituti di istruzione, prezzo su richiesta, diversi modelli completi



Download

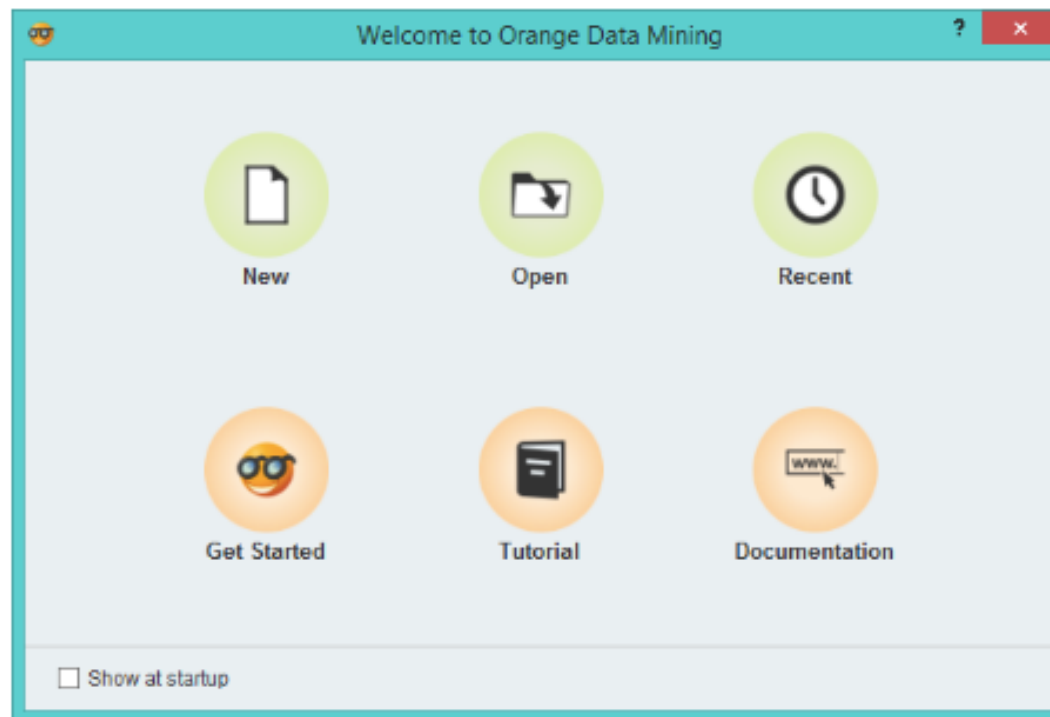


- ▶ Orange 3.30
- ▶ <https://orange.biolab.si/download/>
- ▶ Orange è un software di programmazione visiva basato su componenti per la visualizzazione dei dati, l'apprendimento automatico, il data mining e l'analisi dei dati
- ▶ Tra gli add-on disponibili si utilizzeranno: Text e Image Analytics

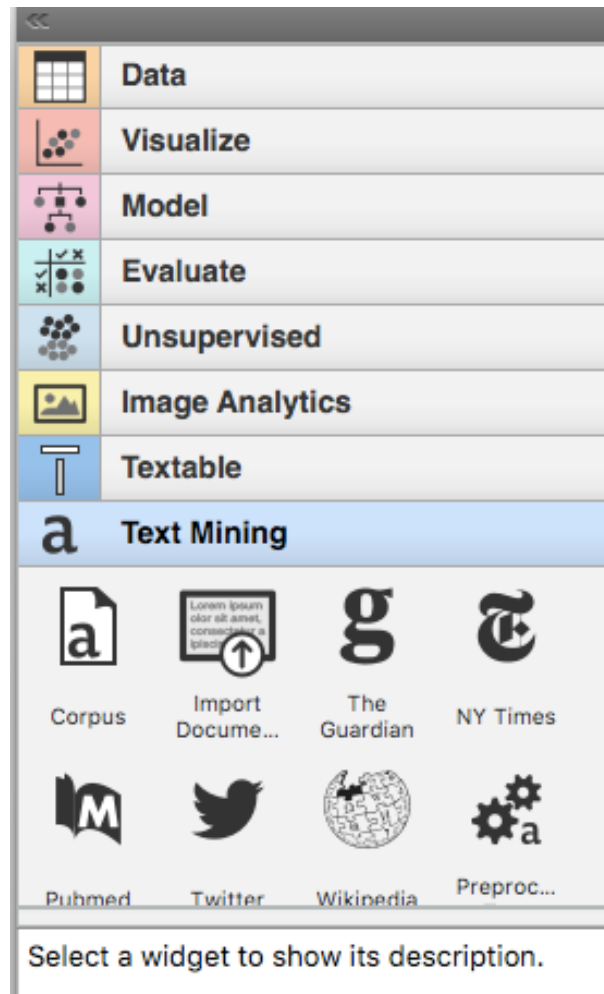


Start

- ▶ All'inizio, Orange apre una schermata di benvenuto. Da qui è possibile creare nuovi flussi di lavoro di data mining o aprire quelli già creati
- ▶ Se si esegue Orange per la prima volta, cliccare su New



Widgets



- ▶ I flussi di lavoro di data mining sono costituiti da componenti computazionali chiamati **widget**
- ▶ I widget eseguono operazioni e scambiano informazioni
- ▶ Si collegano tra loro attraverso connessioni
- ▶ Appena un widget è posto nell'area di lavoro esegue subito il suo compito

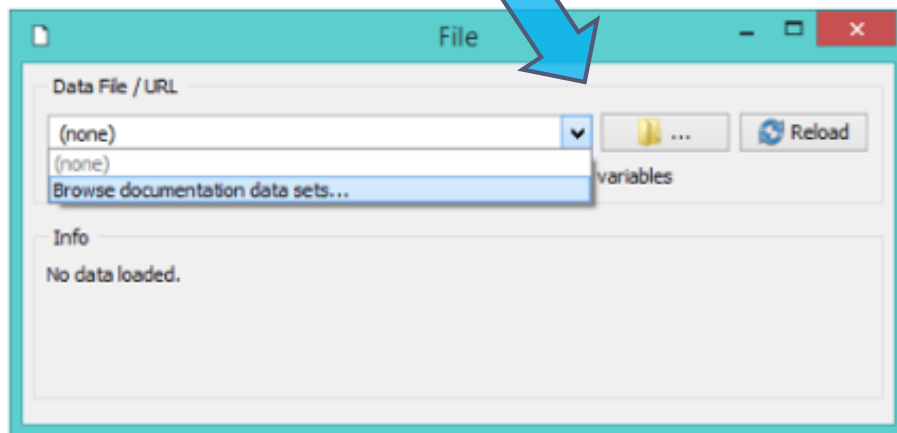
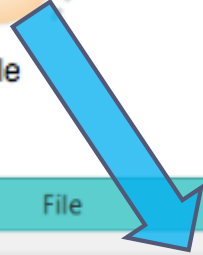
Add-ons

- ▶ Dobbiamo installare alcuni componenti aggiuntivi, fare clic su Options/Add-ons, selezionare i componenti aggiuntivi e cliccare OK
- ▶ Riavviare Orange perché gli add-on appaiano nel menu a sinistra
- ▶ Sugeriamo gli add-ons:
 - ▶ Text
 - ▶ Image Analytics

... (se volete potete sceglierne altri)



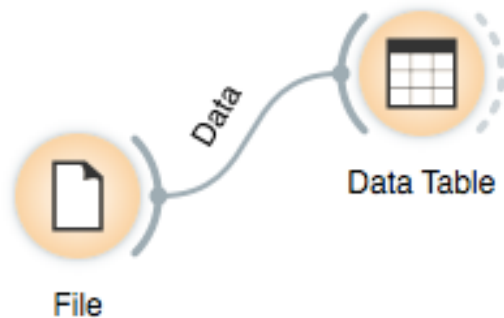
Per usare dati strutturati (quantitativi): File



- ▶ Qualsiasi operazione di data mining inizia con i dati
- ▶ Selezionare il widget **File**
- ▶ Doppio click per aprirlo
- ▶ Selezionare "Browse documentation data sets..." e dall'elenco dei file di dati preinstallati scegliere ***iris.tab***
- ▶ Il widget File ora leggerà il famoso set di dati su 150 fiori Iris



Read the data



The screenshot shows the 'Data Table' widget interface. On the left, there is an 'Info' panel with the following details:

- 150 instances (no missing values)
- 4 features (no missing values)
- Discrete class with 3 values (no missing values)
- No meta attributes

Under 'Variables', the following options are checked:

- Show variable labels (if present)
- Color by instance classes

Under 'Selection', the following option is checked:

- Select full rows

At the bottom, there are two buttons: 'Restore Original Order' and 'Send Automatically' (which is checked).

The main area displays a table with the following columns: 'iris', 'sepal length', 'sepal width', 'petal length', and 'petal width'. The table contains 26 rows of data, with the first row highlighted in grey.

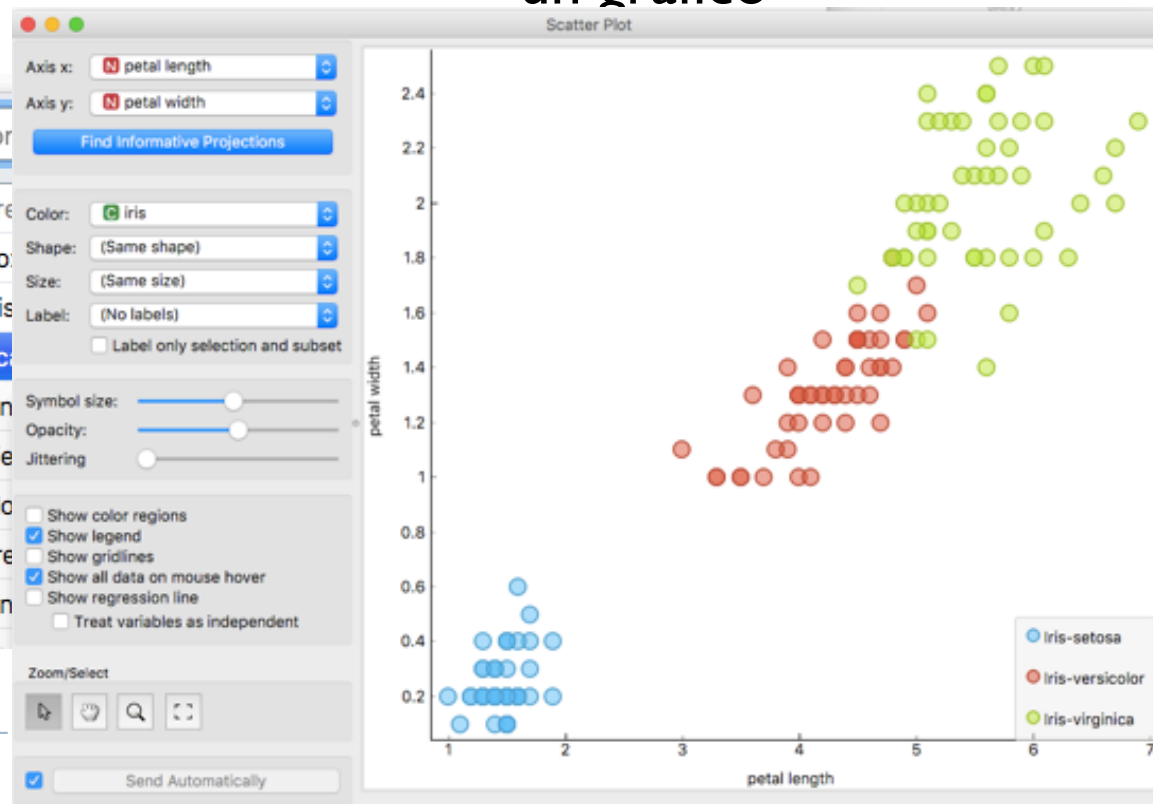
	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	4.4	2.9	1.4	0.2
10	Iris-setosa	4.9	3.1	1.5	0.1
11	Iris-setosa	5.4	3.7	1.5	0.2
12	Iris-setosa	4.8	3.4	1.6	0.2
13	Iris-setosa	4.8	3.0	1.4	0.1
14	Iris-setosa	4.3	3.0	1.1	0.1
15	Iris-setosa	5.8	4.0	1.2	0.2
16	Iris-setosa	5.7	4.4	1.5	0.4
17	Iris-setosa	5.4	3.9	1.3	0.4
18	Iris-setosa	5.1	3.5	1.4	0.3
19	Iris-setosa	5.7	3.8	1.7	0.3
20	Iris-setosa	5.1	3.8	1.5	0.3
21	Iris-setosa	5.4	3.4	1.7	0.2
22	Iris-setosa	5.1	3.7	1.5	0.4
23	Iris-setosa	4.6	3.6	1.0	0.2
24	Iris-setosa	5.1	3.3	1.7	0.5
25	Iris-setosa	4.8	3.4	1.9	0.2
26	Iris-setosa	6.0	3.0	1.6	0.2

- ▶ Vogliamo che il widget File legga i dati e li invii alla **Data Table** per l'ispezione
- ▶ Dobbiamo collegare questi due widget per stabilire una comunicazione tra di loro
- ▶ Fare clic sulla linea tratteggiata del widget File e trascinare la linea verso Data Table
- ▶ Doppio clic su Data Table per aprirlo
- ▶ Questo mostra i dati che abbiamo appena caricato

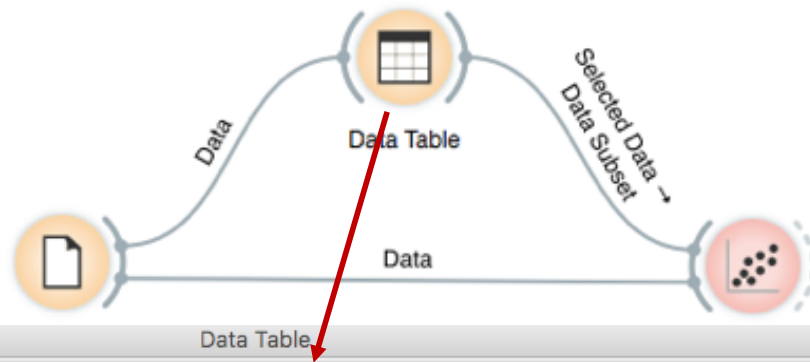
Visualize the data



- ▶ Collegare File con **Scatter plot** per visualizzare i dati in un grafico

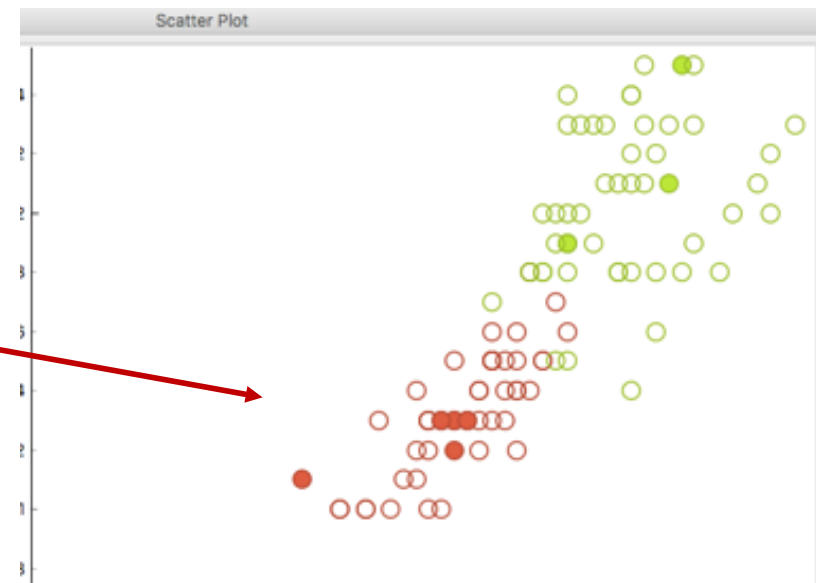


Visualize the data



- ▶ Se si selezionano alcuni valori nella Data Table, si vede immediatamente l'effetto nel grafico e viceversa

Data Table					
	iris	sepal length	sepal width	petal length	petal width
92	Iris-versicolor	5.1	3.0	4.0	1.4
93	Iris-versicolor	5.8	2.6	4.0	1.2
94	Iris-versicolor	5.0	2.3	3.3	1.0
95	Iris-versicolor	5.6	2.7	4.2	1.3
96	Iris-versicolor	5.7	3.0	4.2	1.2
97	Iris-versicolor	5.7	2.9	4.2	1.3
98	Iris-versicolor	6.2	2.9	4.3	1.3
99	Iris-versicolor	5.1	2.5	3.0	1.1
100	Iris-versicolor	5.7	2.8	4.1	1.3
101	Iris-virginica	6.3	3.3	6.0	2.5
102	Iris-virginica	5.8	2.7	5.1	1.9
103	Iris-virginica	7.1	3.0	5.9	2.1
104	Iris-virginica	6.3	2.9	5.6	1.8
105	Iris-virginica	6.5	3.0	5.8	2.2
106	Iris-virginica	7.6	3.0	6.6	2.1



Load your data

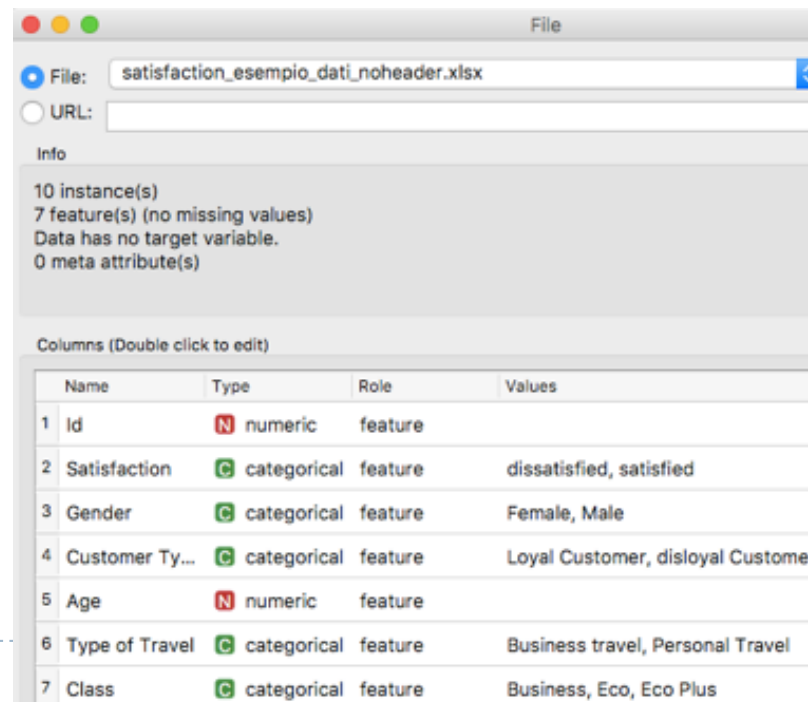
- ▶ Orange lavora con un proprio formato di dati, ma può anche gestire Excel nativi, file di dati delimitati da virgole (csv) o da tabulazioni (tab) o documenti di Fogli Google (via URL)
- ▶ Il set di dati di input è generalmente una tabella, con istanze di dati in riga e attributi di dati in colonna
- ▶ Esempio:

satisfaction_esempio_dati_noheader

Id	Satisfaction	Gender	Customer Type	Age	Type of Travel	Class
11112	satisfied	Female	Loyal Customer	65	Personal Travel	Eco
37147	dissatisfied	Female	disloyal Customer	25	Business travel	Eco
39454	dissatisfied	Male	Loyal Customer	40	Personal Travel	Eco
53026	dissatisfied	Female	disloyal Customer	18	Business travel	Eco
83338	satisfied	Female	Loyal Customer	27	Personal Travel	Eco Plus
90701	satisfied	Male	disloyal Customer	38	Business travel	Business
110278	satisfied	Male	Loyal Customer	47	Personal Travel	Business
113976	satisfied	Female	disloyal Customer	38	Business travel	Business
118245	dissatisfied	Male	Loyal Customer	18	Personal Travel	Business
122126	dissatisfied	Male	Loyal Customer	48	Personal Travel	Eco Plus

Load your data (2)

- ▶ Per caricare i dati, aprire il widget File
- ▶ Cliccare sull'icona del file browser e selezionare il file sul vostro disco
- ▶ Orange riconosce i tipi di attributo anche se non sono definiti nel set di dati

The screenshot shows the 'File' widget interface in Orange. At the top, there are three colored window control buttons (red, yellow, green). Below them, the 'File' field is selected with a blue radio button and contains the text 'satisfaction_esempio_dati_noheader.xlsx'. There is also an 'URL' field with an unselected radio button. Underneath, an 'Info' section displays: '10 instance(s)', '7 feature(s) (no missing values)', 'Data has no target variable.', and '0 meta attribute(s)'. At the bottom, a 'Columns (Double click to edit)' table lists the data features with their names, types, roles, and values.

	Name	Type	Role	Values
1	Id	N numeric	feature	
2	Satisfaction	C categorical	feature	dissatisfied, satisfied
3	Gender	C categorical	feature	Female, Male
4	Customer Ty...	C categorical	feature	Loyal Customer, disloyal Customer
5	Age	N numeric	feature	
6	Type of Travel	C categorical	feature	Business travel, Personal Travel
7	Class	C categorical	feature	Business, Eco, Eco Plus

Load your data (3)

- ▶ Gli attributi possono essere di diverso tipo (**continuous, discrete, time, strings**) e avere diversi ruoli (features, meta attributes, class)

satisfaction_esempio_dati_header

Attribute types
Roles

Id	Satisfaction	Gender	Customer Type	Age	Type of Travel	Class
c	d	d	d	c	d	d
meta	class	class	class		class	class
11112	satisfied	Female	Loyal Customer	65	Personal Travel	Eco
37147	dissatisfied	Female	disloyal Customer	25	Business travel	Eco
39454	dissatisfied	Male	Loyal Customer	40	Personal Travel	Eco
53026	dissatisfied	Female	disloyal Customer	18	Business travel	Eco
83338	satisfied	Female	Loyal Customer	27	Personal Travel	Eco Plus
90701	satisfied	Male	disloyal Customer	38	Business travel	Business
110278	satisfied	Male	Loyal Customer	47	Personal Travel	Business
113976	satisfied	Female	disloyal Customer	38	Business travel	Business
118245	dissatisfied	Male	Loyal Customer	18	Personal Travel	Business
122126	dissatisfied	Male	Loyal Customer	48	Personal Travel	Eco Plus

Formato nativo
di Orange con
tre righe di
intestazione

N.B.: Attribute types (continuous, discrete o categorical, time, string)
Roles: (class, meta, ignore the attribute, instance weights)

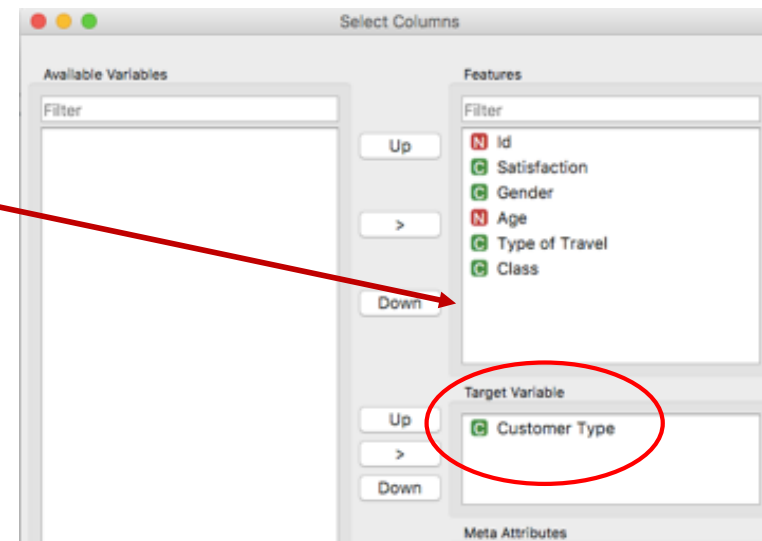
Load your data (4)

- ▶ Se hai definito tipi e ruoli, Orange li riconosce e li evidenzia
- ▶ Tipi e ruoli di attributi possono essere forniti nell'intestazione della Data Table
- ▶ Possono anche essere successivamente modificati nel widget File, mentre anche il ruolo dei dati può essere modificato con il widget **Select Columns**

10 instance(s)
1 feature(s) (no missing values)
Multi-target; 5 target variables (no missing values)
1 meta attribute(s)

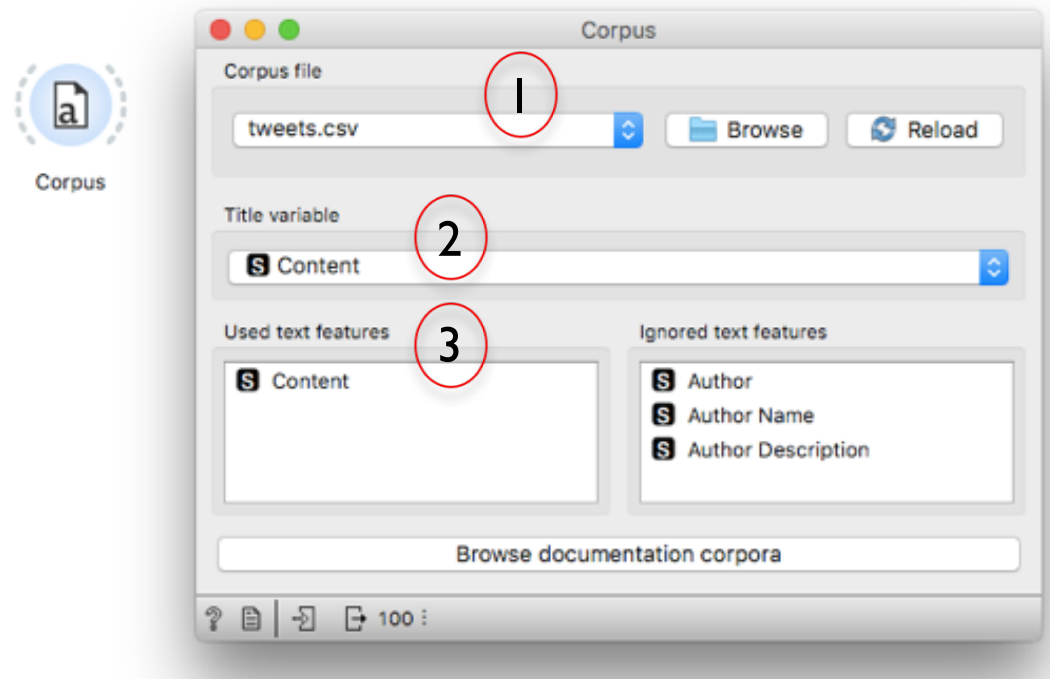
Columns (Double click to edit)

Name	Type	Role	Values
1 Age	N numeric	feature	
2 Satisfaction	C categorical	target	dissatisfied, satisfied
3 Gender	C categorical	target	Female, Male
4 Customer Ty...	C categorical	target	Loyal Customer, disloyal Customer
5 Type of Travel	C categorical	target	Business travel, Personal Travel
6 Class	C categorical	target	Business, Eco, Eco Plus
7 Id	N numeric	meta	



Text Mining with Orange

► 1° widget: Corpus



1 Caricare il file con i testi

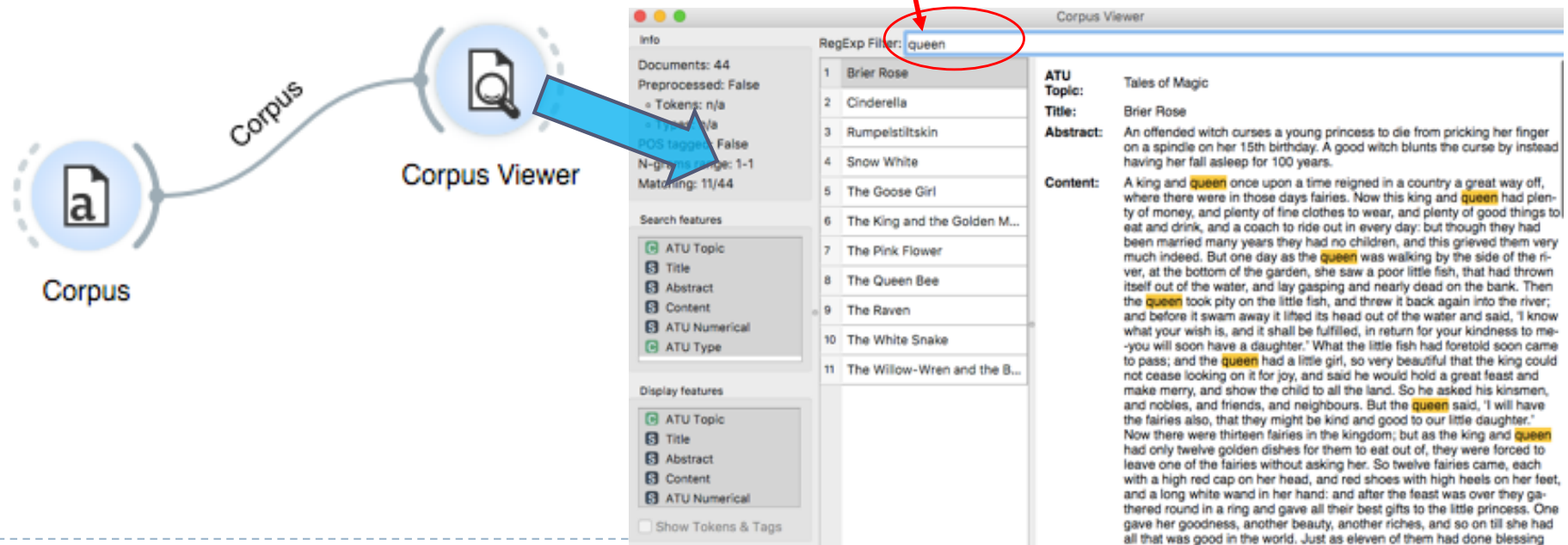
2 Selezionare il nome della colonna che contiene i testi

3 Identificare la variabile testuale su cui applicare le analisi



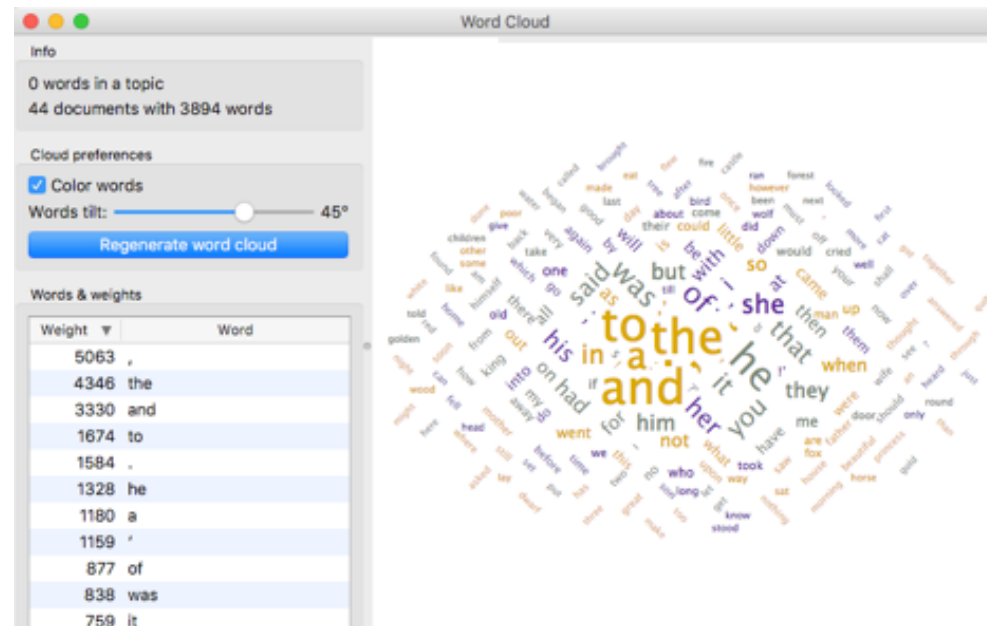
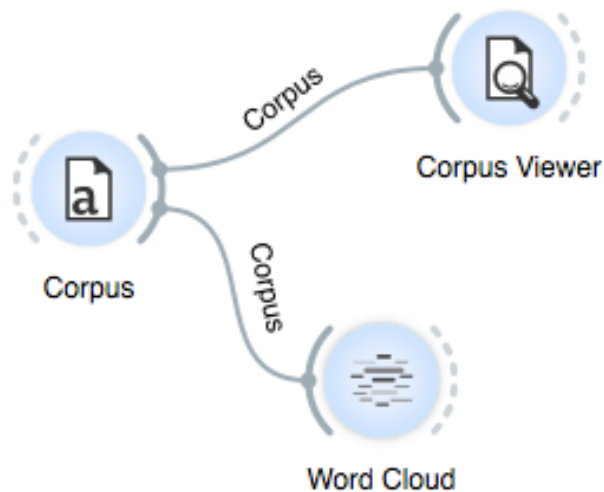
Text processing (1)

- ▶ Cliccare sul widget **Corpus** e aprirlo, per esempio caricare il file Grimm-tales-selected
- ▶ Collegare **Corpus** a **Corpus Viewer** per visualizzare i testi, per esempio cercare alcune espressioni regolari o parole chiave



Text processing (2)

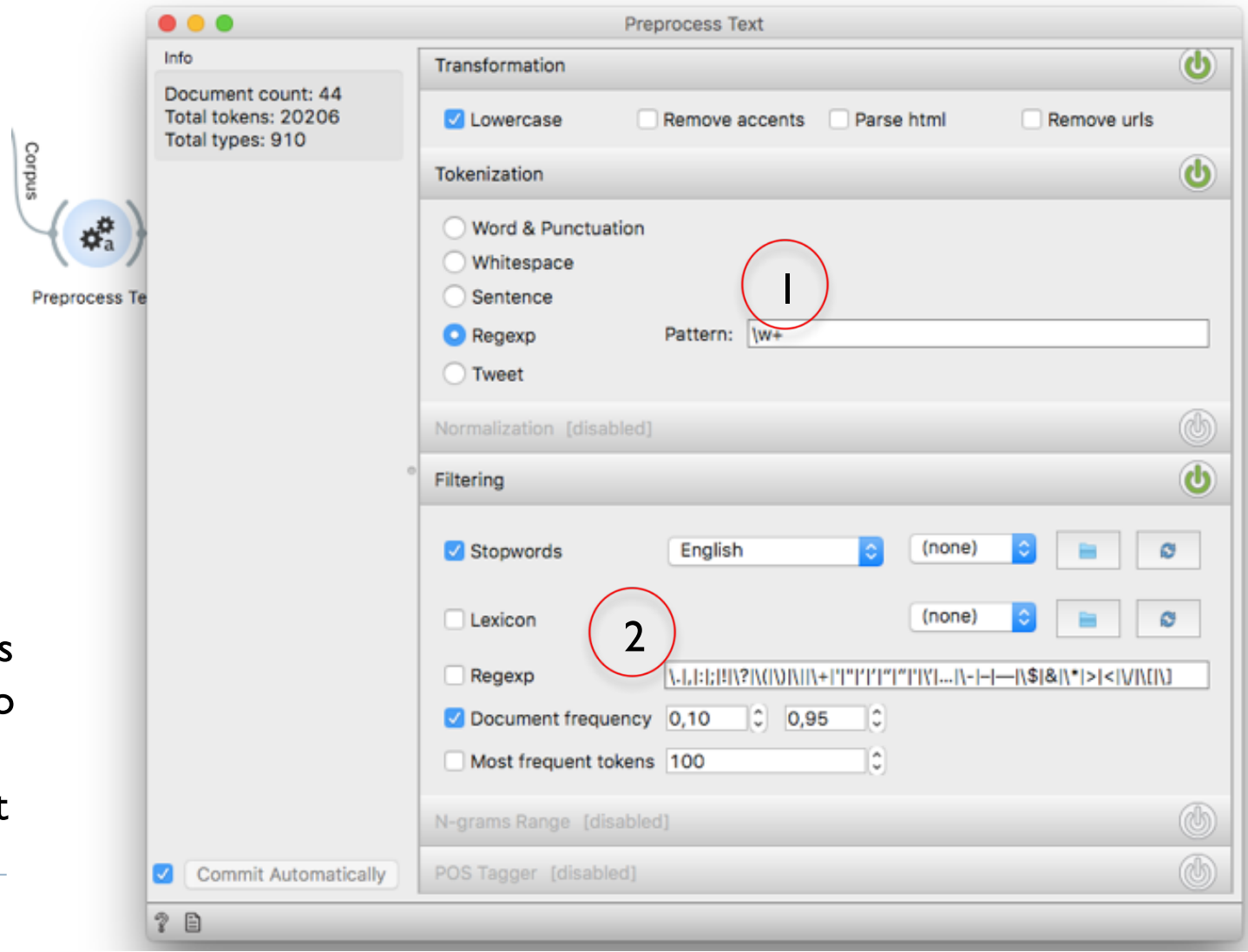
- ▶ Un altro widget per visualizzare il testo è Word Cloud che mostra le frequenze delle parole in forma di nuvola,
- ▶ ...ma quali sono le parole più frequenti?



Text pre-processing

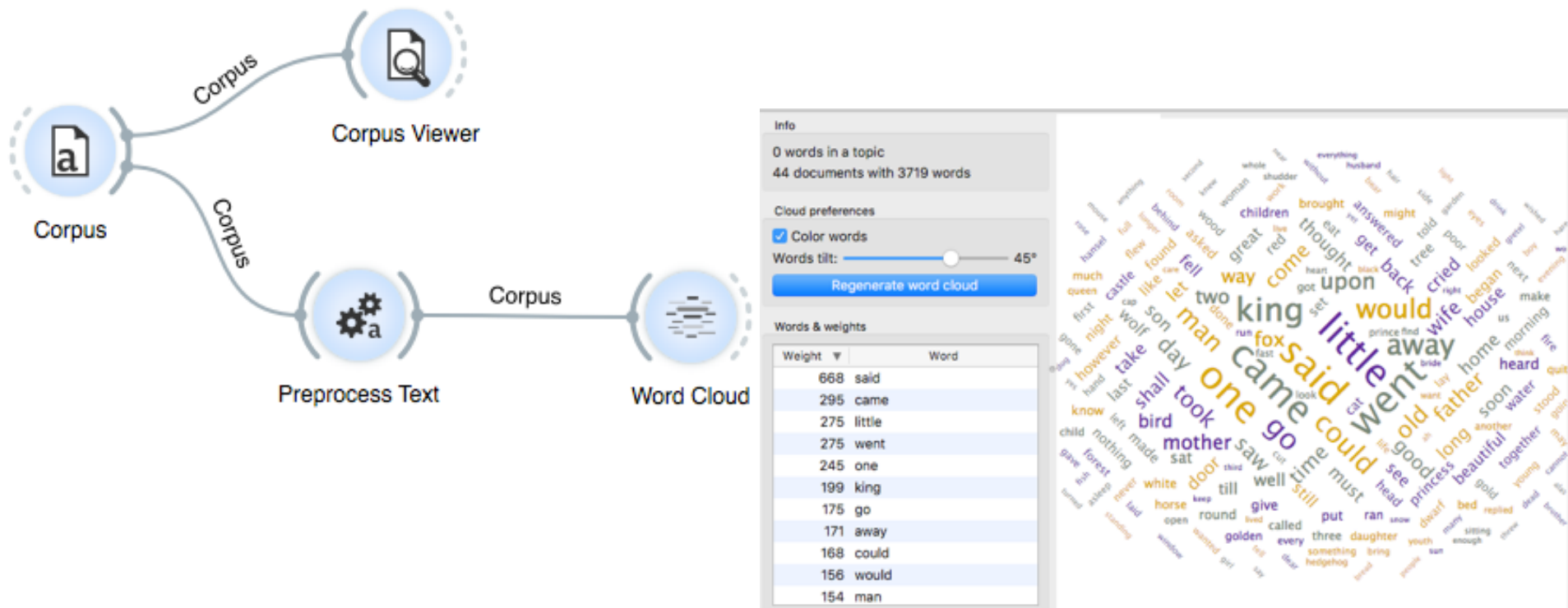
▶ Widget: Preprocess Text

1 Tokenizza
tenendo le
parole e
eliminando la
punteggiatura
2 Filtra i tokens
che rispondono
a espressioni
regolari, default
elimina
punteggiatura



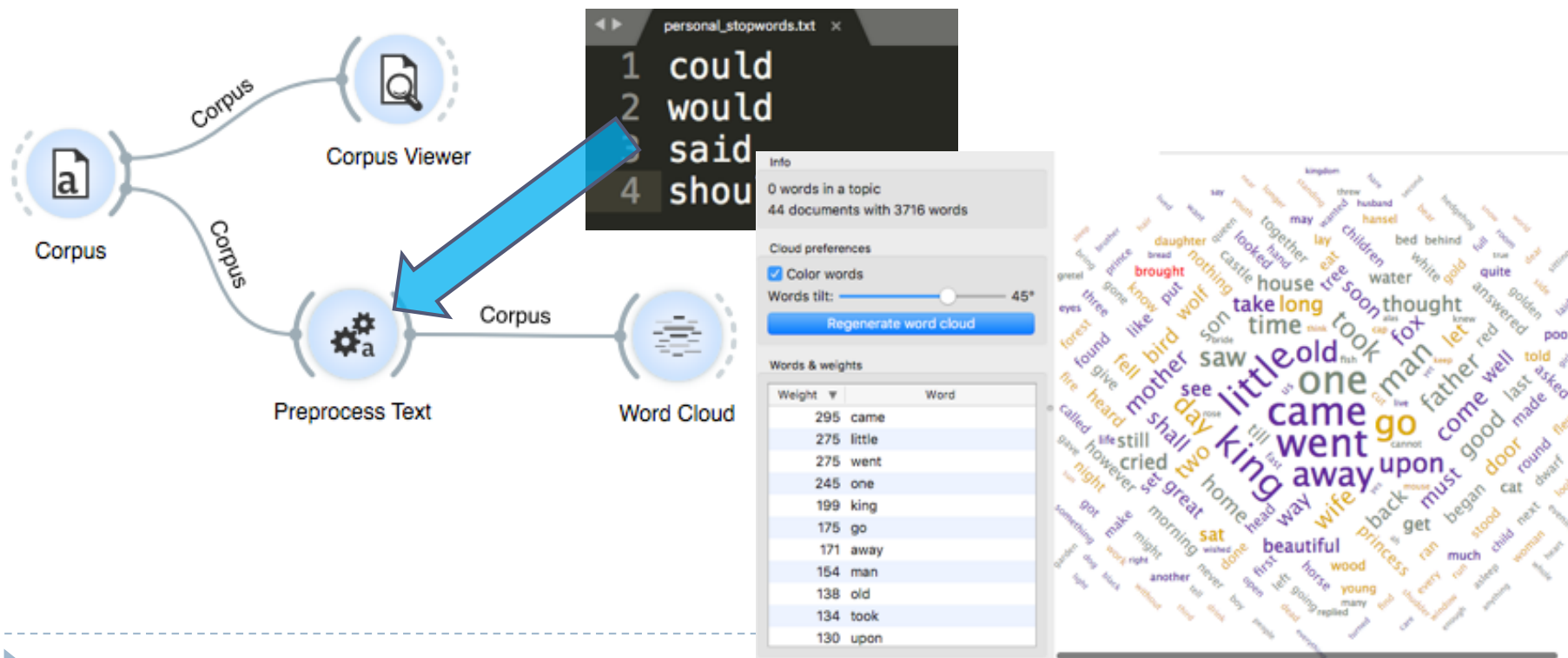
Text processing (3)

- ▶ Usiamo il widget Preprocess Text per tokenizzare, trasformare, eliminare le stop words, fare stemming o generare gli n-grammi



Text processing (4)

- ▶ Possiamo eliminare altre parole, es. “would” “could” ecc., creando una lista di stop words personalizzata, in un file txt da caricare nel widget Preprocess Text (sezione Filtering, 2° menù a dx)
- ▶ I cambiamenti si propagano nel processo, aprire la Word Cloud per vedere le modifiche



Scaricare tweets con Orange (1)

- ▶ Widget Twitter richiede le credenziali della Twitter API, per averle si deve creare una app in <https://apps.twitter.com/> e inserirle nei due campi richiesti

The image shows two overlapping windows from the Orange Data Mining software. On the left is the 'Twitter' widget configuration window, and on the right is a 'Twitter API Credentials' dialog box. A red arrow points from the 'Twitter API Key' field in the widget window to the 'Key' field in the dialog box.

Twitter Widget Configuration:

- Twitter API Key: [Field with circled 1]
- Query: [Field with circled 2]
- Query word list: [Text area]
- Search by: Content
- Language: Any
- Max tweets: 100
- Allow retweets:
- Collect results:
- Text includes: Content (checked), Author Description (unchecked) [Field with circled 3]
- Info: Tweets on output: 0 [Field with circled 4]
- Search: [Field with circled 5]

Twitter API Credentials Dialog:

- Key: [Text input field]
- Secret: [Text input field]
- OK: [Button]

Scaricare tweets con Orange (2)

▶ Inserire i parametri:

- ▶ *Query word list*: una parola per riga, usa l'operatore OR
- ▶ *Search by*: cercare per contenuto (hashtag) o per autore
- ▶ *Language*: scegliere la lingua
- ▶ *Max tweets*: fissare il massimo di tweet da scaricare
- ▶ *Allow retweets*: se è selezionato scarica anche i RT (si consiglia di non selezionarlo)
- ▶ *Collect results*: se è selezionato appenderà nuovi tweets scaricati con nuove queries alle precedenti (si consiglia di non selezionarlo)

▶ Cliccare Search per avviare lo scarico



Esempio scarico tweets

The image shows a workflow for downloading tweets. On the left, a Twitter icon is connected to a Corpus Viewer icon. A red arrow points from the Corpus Viewer icon to the main application window. The application window is titled "Corpus Viewer" and is divided into three main sections:

- Search Panel (Left):** Contains a "Twitter API Key" field, a "Query" section with a text box containing "data mining", "machine learning", and "text mining", a "Search by:" dropdown set to "Content", a "Language:" dropdown set to "English", "Max tweets:" set to 100, and checkboxes for "Allow retweets:" and "Collect results:". Below this is a "Text includes" section with checkboxes for "Content" (checked) and "Author Description". At the bottom, there is an "Info" section showing "Tweets on output: 100" and a "Search" button.
- Tweet List (Center):** A scrollable list of tweets. The selected tweet is highlighted in blue and reads: "Keep up with the best and the latest ... Machine learning competitions offer ... @indigenousNerd @every_penny_ It'... [EN] New machine learning theory ... \$NIHK (Video River Networks Inc) CE... Machine Learning with https://t.co/... Machine Learning with https://t.co/... @wtfprincu Don't know how much ... Machine Learning Photonics - Our ... 'Myanmar court adjourns Suu Kyi ... China has administered 65m vaccine... Brain age prediction in schizophrenia... The convergence of #AI and ... @DrHOSP1 🤖🤖🤖 these guys wor...".
- Tweet Details (Right):** A detailed view of the selected tweet. It includes:
 - Content:** Keep up with the best and the latest #machinelearning #research blogs through reliable sources.
 - Date:** 2021-03-15 08:52:46
 - Language:** en
 - Location:** ?
 - Number of Likes:** 0
 - Number of Retweets:** 0
 - In Reply To:** ?
 - Author Name:** DAC.digital
 - Author Description:** Top IT Services PL 2020 (Clutch), International Chempion 2020 (acc to PwC&Puls Biznesu) #SoftwareDevelopment #DataManagement #Blockchain #HardwareIntegration
 - Author Statuses Count:** 587
 - Author Favourites Count:** 1173
 - Author Friends:** 116

Sentiment Analysis with Orange



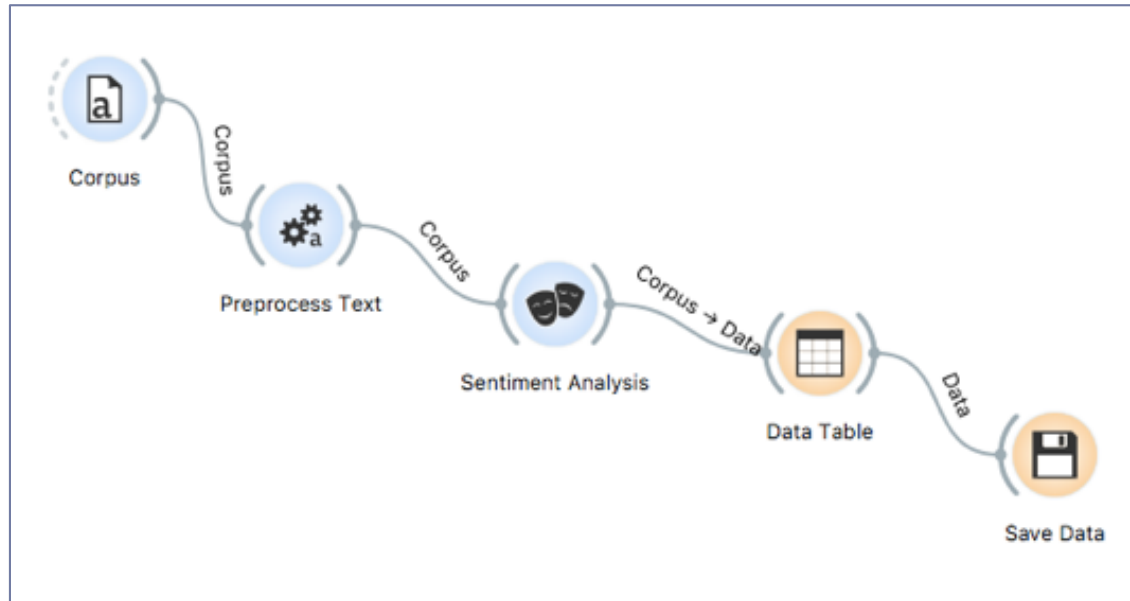
Sentiment Analysis

- ▶ Per eseguire la Sentiment Analysis su testi collegare Corpus con Sentiment Analysis (+ Data Table e Save data)
- ▶ Si ottiene un'analisi migliore se i testi sono già puliti
- ▶ Il widget Sentiment Analysis consente di utilizzare tre diversi metodi basati su lexicon: Liu Hu e Vader (entrambi per la lingua inglese). Dalla versione 3.24 anche multilingua per Italiano
- ▶ N.B. Dalla versione 3.28 è fornita la possibilità di caricare dei lessici personalizzati divisi in parole positive e negative

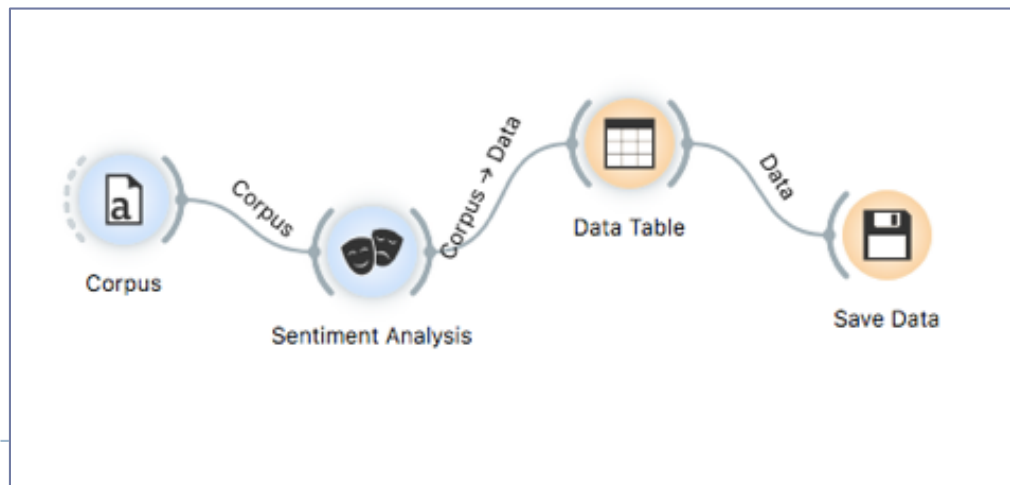


Alternative valide

1)
processo
con
Preprocess
Text



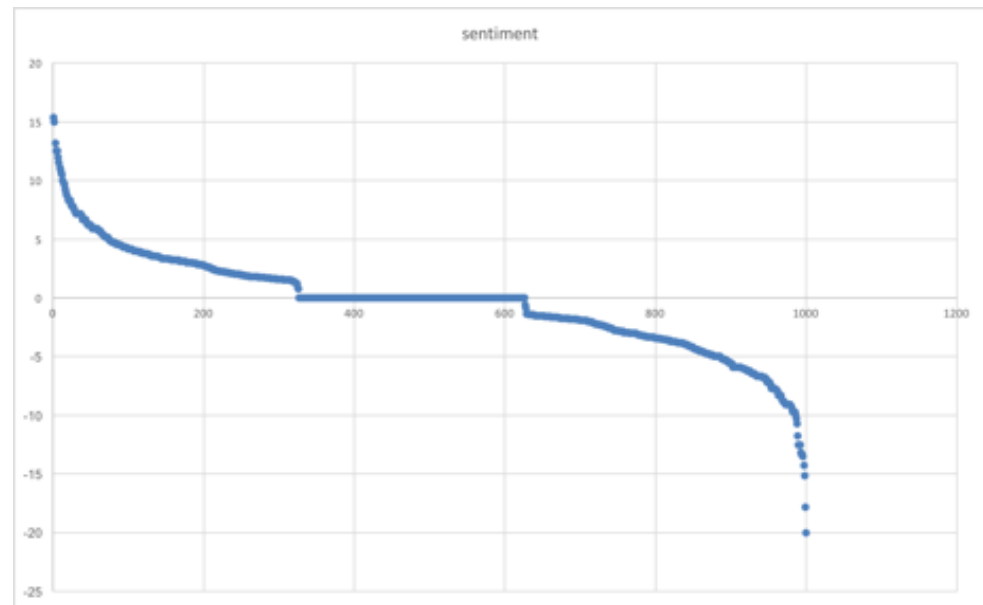
2)
processo
senza
Preprocess
Text



Sentiment Analysis, lettura risultati

- ▶ Il metodo **Liu Hu** (e il **multilingua**) restituisce una colonna con il valore di Sentiment, un numero compreso tra $-n$ e $+n$
- ▶ Ordinando tali valori è possibile creare un grafico (es. dispersione) per interpretare i risultati

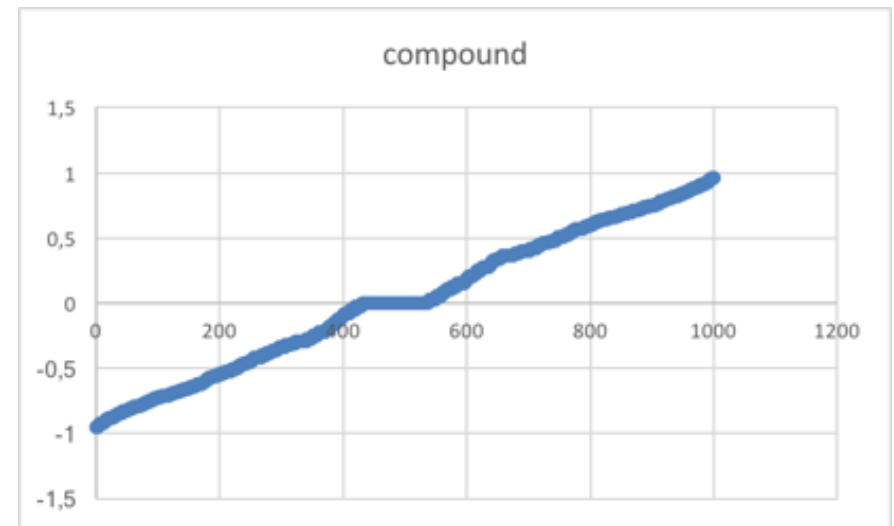
	Text	sentiment
1	What worrie...	-9.09091
2	Until Labour ...	0
3	Labour lost ...	-1.75439
4	@tombstone...	1.96078
5	@davidallen...	-5
6	@LukeWillia...	0
7	@SeemaCha...	-1.78571
8	@CNN The f...	5.71429
9	@THEPFY20...	0
10	@fascinatorf...	3.33333



Sentiment Analysis, lettura risultati (2)

- ▶ Il metodo **Vader** restituisce tre colonne con le polarità pos, neg e neu, e la colonna compound con il valore di sentiment per ciascun testo (compreso tra -1 e +1)
- ▶ Ordinando i valori di compound è possibile creare un grafico (es. dispersione) per interpretare i risultati

pos	neg	neu	compound
0	0,286	0,714	-0,4215
0,144	0,136	0,72	0,34
0,052	0,271	0,677	-0,8832
0,152	0,057	0,791	0,6478
0	0,105	0,895	-0,25
0	0	1	0



Sentiment Analysis, lettura risultati (3)

- ▶ Si può collegare a Sentiment Analysis l'operatore Distributions, nei parametri selezionare la variabile sentiment, e selezionare una delle barre nel grafico e, collegando un Corpus Viewer, leggere i testi con la polarità selezionata

