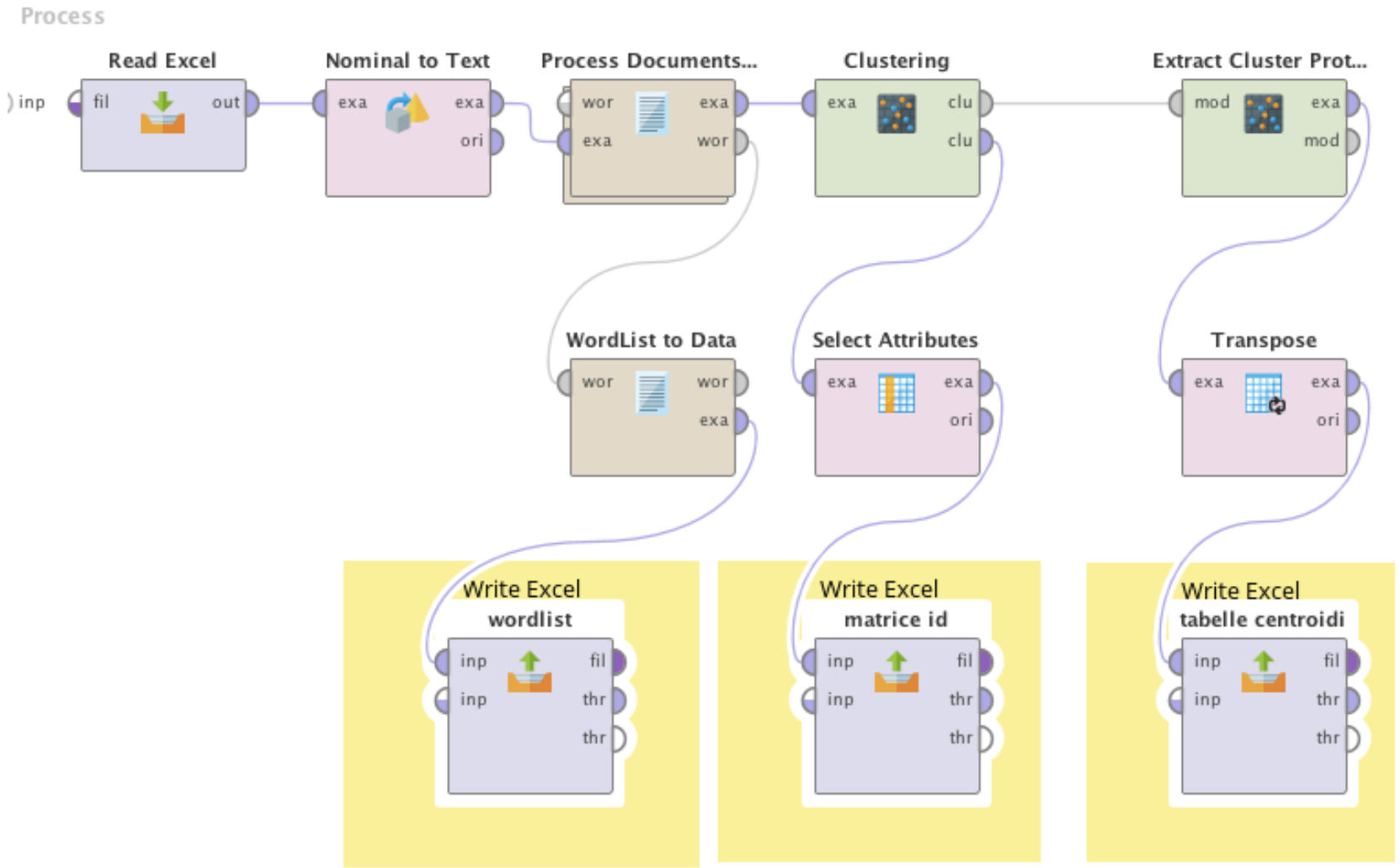
Cluster Analysis con Rapidminer

Clustering, k-means



Operatori e parametri k-means

- Process documents from data:
  - Add meta information = no flag
- Clustering, usa algoritmo k-means, parametri:
  - K = numero >0 (default 5)
  - Measure types =Numerical Measure
  - Numerical Measure=Cosine
- Select Attribute:
  - Attribute filter type = single
  - Attribute = cluster
- Extract Cluster Prototypes + Transpose per salvare su Excel i valori dei centroidi
- Write Excel “matrice id” salva su file la colonna “cluster” con id di ogni cluster per ogni testo

Letture dei cluster

- Le tabelle centroide sono salvate in un excel
- Ordinare ogni colonna corrispondente a un cluster dall'alto al basso (dal più grande al più piccolo)
- I valori numerici più alti nelle prime righe corrispondono a delle parole che indicano il tema del cluster

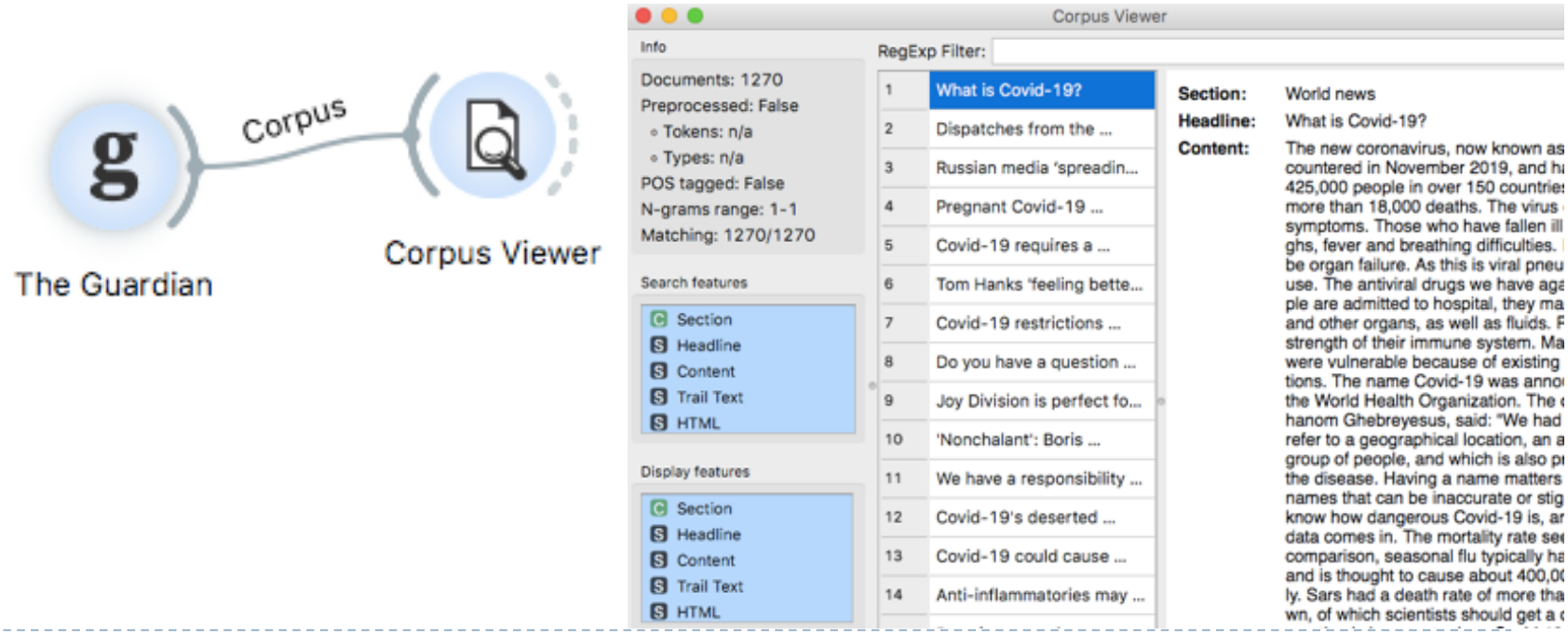
ATTENZIONE che il valore numerico inizi sempre con 0,…

cluster_0	word	cluster_1	word
0,0283	privacy	0,0658	facebook
0,0273	dream	0,0515	zuckerberg
0,0228	deleted	0,0498	think
0,0227	sorry	0,0412	win
0,0222	trust	0,0400	data

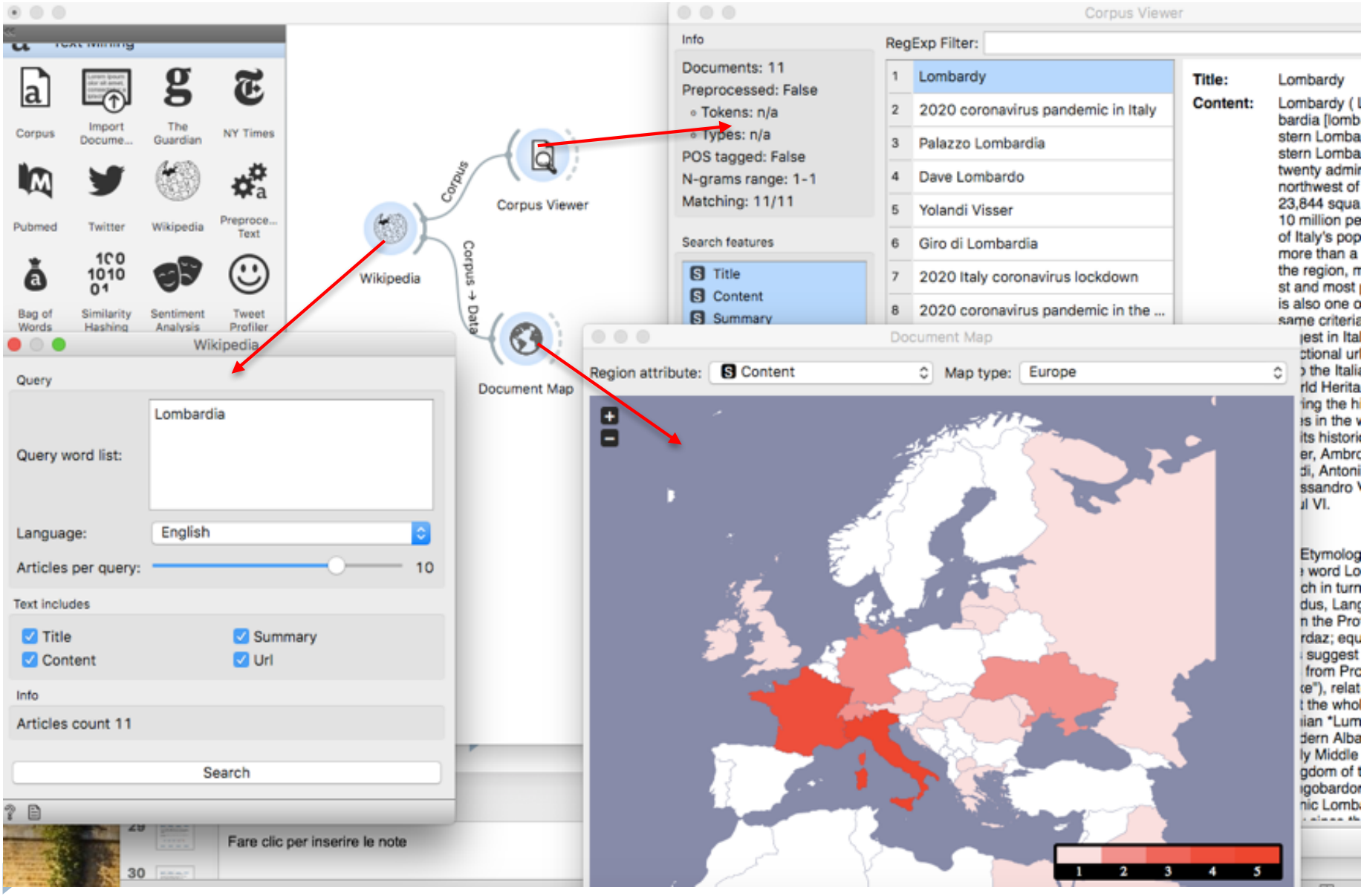
Processi avanzati TM Orange

Scaricare articoli

- Alcuni widget permettono di scaricare articoli via API: Guardian, NY Times, Twitter, Wikipedia, Pubmed
- Necessitano delle chiavi delle API (previa registrazione, eccetto Wikipedia, Pubmed richiede email)

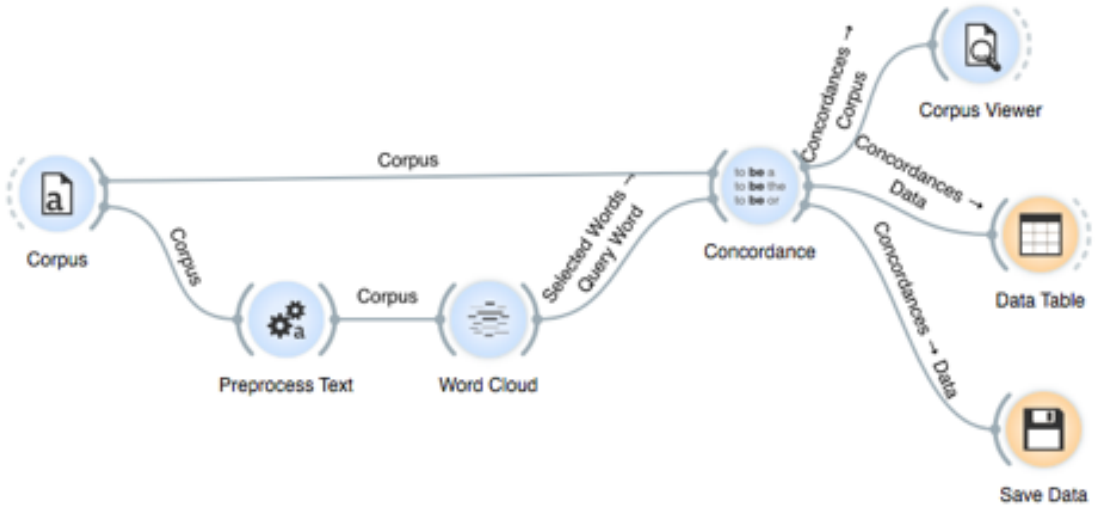


Esempio da Wikipedia con Document Map



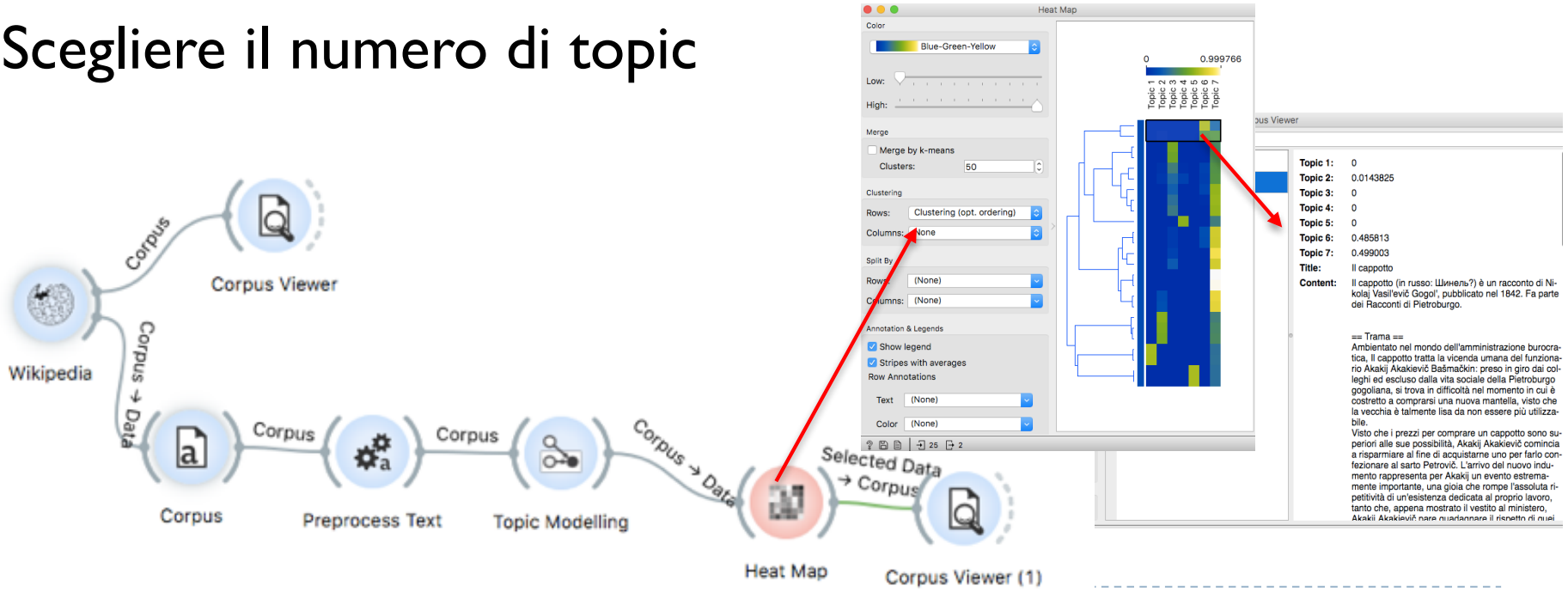
Concordance

- **Concordance** trova la parola interrogata in un testo e visualizza il contesto in cui viene utilizzata la parola
- Seleziona una parola in Word Cloud,
- La parola selezionata verrà inviata a Concordance come parola di ricerca
- Concordance trova tutte le occorrenze della parola da Word Cloud nel corpus
- Seleziona documenti interessanti e osservali in Corpus Viewer o in una Data Table
- Salva i dati in un file .tab per ulteriori analisi



Topic Modelling

- LDA (Latent Dirichlet Allocation) è un metodo matematico basato sull'occorrenza delle parole, utile per trovare un set di parole associato a ciascun argomento e determinare anche la combinazione di topic che descrive ogni documento (alternativo a clustering)
- Scegliere il numero di topic



Tweet Profiler, emozioni rilevate