

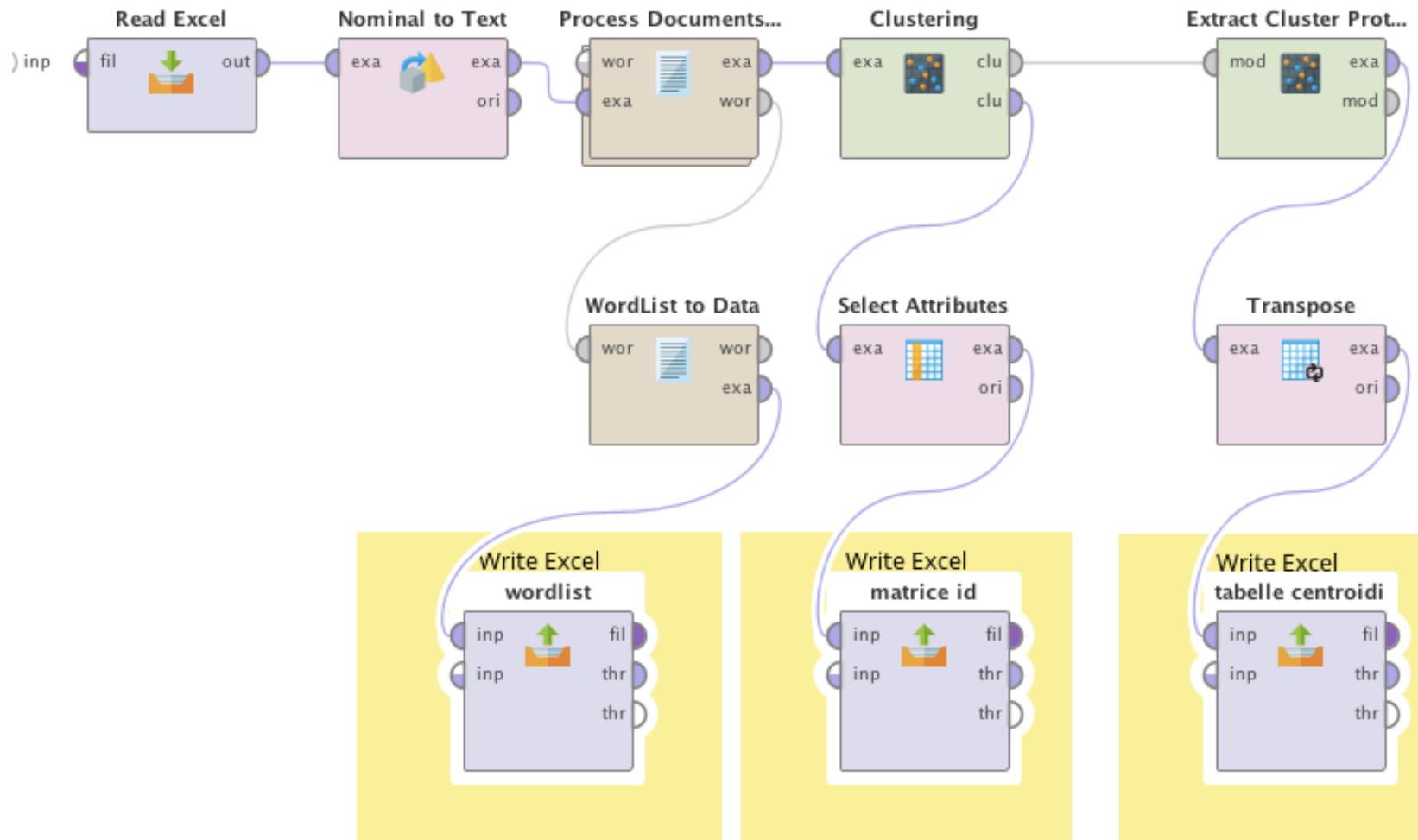
Cluster Analysis con Rapidminer



Clustering, k-means



Process



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

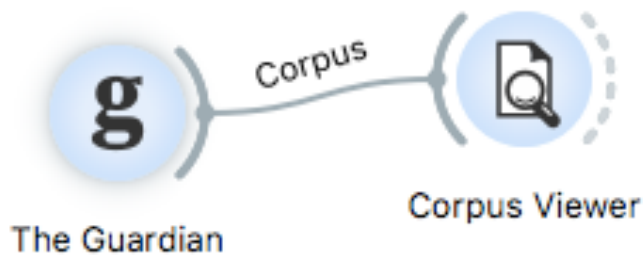
ATTENZIONE che il valore numerico inizi sempre con 0,...

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



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)



The screenshot shows the 'Corpus Viewer' application window. The title bar reads 'Corpus Viewer'. The interface is divided into several sections:

- Info:** Documents: 1270, Preprocessed: False, Tokens: n/a, Types: n/a, POS tagged: False, N-grams range: 1-1, Matching: 1270/1270.
- Search features:** A list of search criteria with checkboxes: Section (checked), Headline, Content, Trail Text, HTML.
- Display features:** A list of display options with checkboxes: Section (checked), Headline, Content, Trail Text, HTML.
- RegExp Filter:** A text input field for filtering results.
- Results Table:** A table with 14 rows. The first row is highlighted in blue and contains the text 'What is Covid-19?'. The other rows contain truncated titles of articles.
- Article Preview:** To the right of the table, the details for the selected article are shown, including the **Section:** World news, **Headline:** What is Covid-19?, and **Content:** The new coronavirus, now known as countered in November 2019, and hi 425,000 people in over 150 countrie more than 18,000 deaths. The virus symptoms. Those who have fallen ill ghs, fever and breathing difficulties. l be organ failure. As this is viral pneu use. The antiviral drugs we have age ple are admitted to hospital, they ma and other organs, as well as fluids. F strength of their immune system. Ma were vulnerable because of existing tions. The name Covid-19 was anno the World Health Organization. The c hanom Ghebreyesus, said: "We had refer to a geographical location, an a group of people, and which is also p the disease. Having a name matters names that can be inaccurate or stig know how dangerous Covid-19 is, ar data comes in. The mortality rate set comparison, seasonal flu typically ha and is thought to cause about 400,0t ly. Sars had a death rate of more tha wn, of which scientists should get a c

Esempio da Wikipedia con Document Map

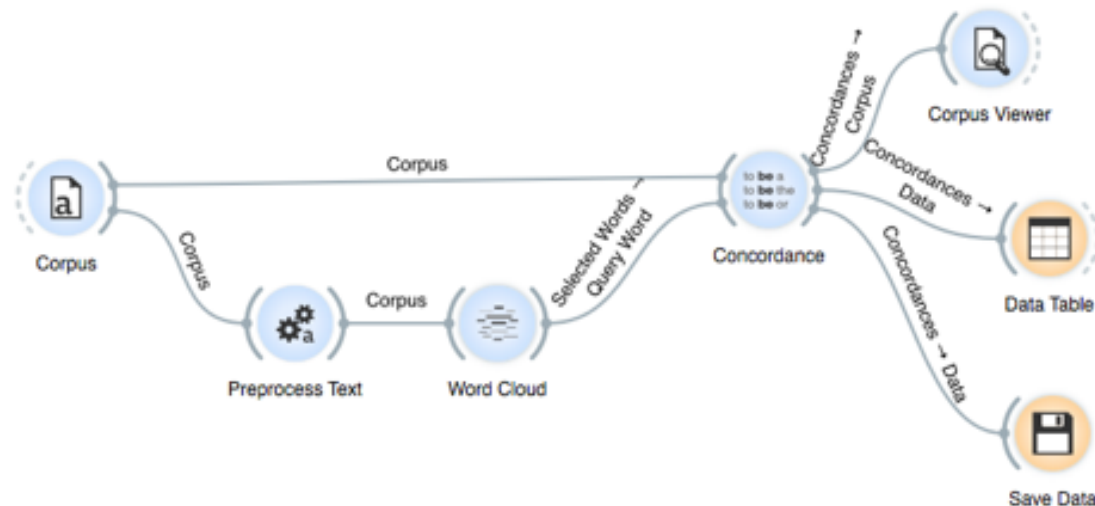
The screenshot illustrates a workflow for generating a Document Map from Wikipedia data. The interface is divided into several windows:

- Wikipedia:** A search window where the query "Lombardia" is entered. The language is set to "English" and the number of articles per query is set to 10. The "Text includes" section has checkboxes for Title, Content, Summary, and Url, all of which are checked. The "Info" section shows "Articles count 11".
- Corpus Viewer:** A window showing the search results. The "Info" section displays: Documents: 11, Preprocessed: False, Tokens: n/a, Types: n/a, POS tagged: False, N-grams range: 1-1, Matching: 11/11. The "Search features" section has checkboxes for Title, Content, and Summary, all checked. The "RegExp Filter" section is empty. The "Title" and "Content" columns are visible, with the first result being "Lombardy".
- Document Map:** A window showing a map of Europe. The "Region attribute" is set to "Content" and the "Map type" is set to "Europe". The map shows Italy highlighted in red, indicating the location of the documents. A legend at the bottom right shows a color scale from 1 to 5.

Red arrows indicate the flow of data: from the Wikipedia search results to the Corpus Viewer, and from the Corpus Viewer to the Document Map.

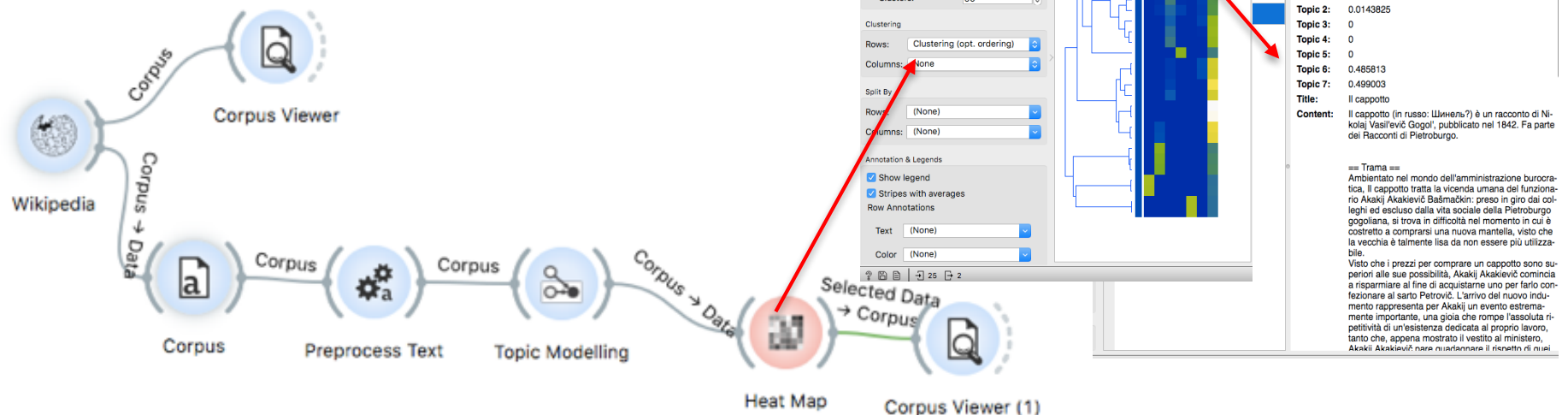
Concordance

- ▶ **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

