

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

Text Mining con Rapidminer Studio

Roberto Boselli
roberto.boselli@unimib.it

Rapidminer Studio, introduzione

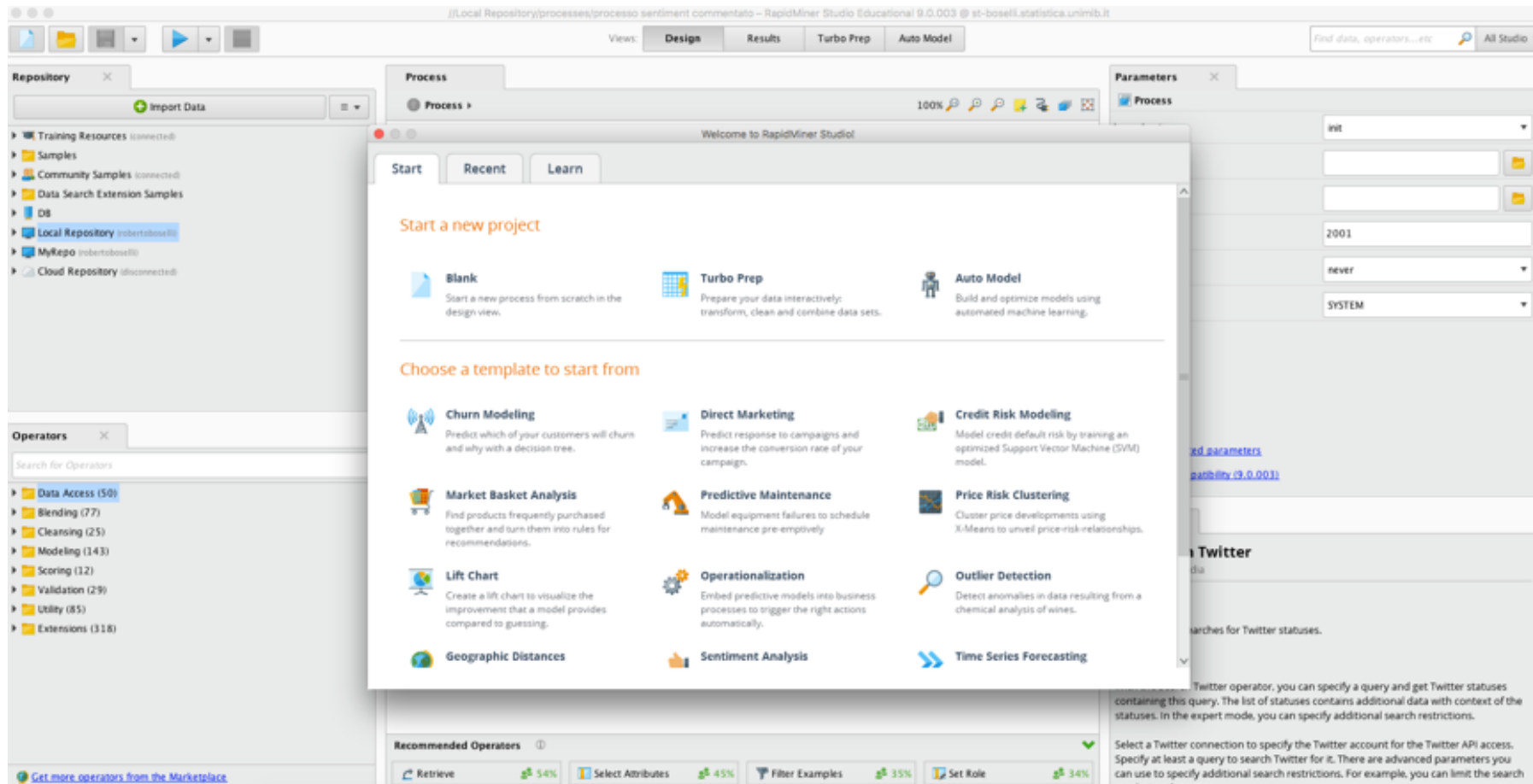


Installazione e licenza

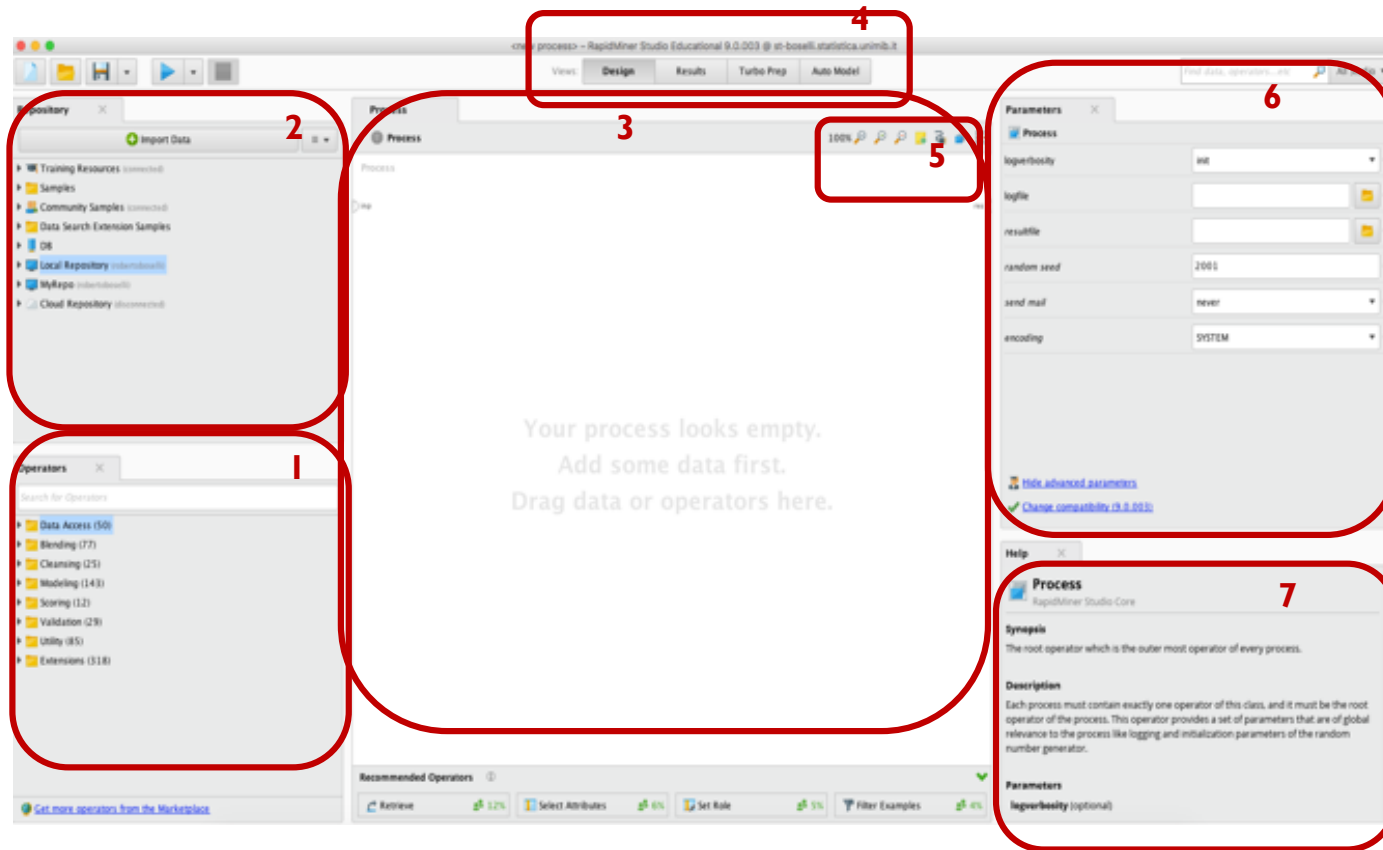
- Scaricare Rapidminer Studio (9.10) dal sito <https://rapidminer.com/get-started/> (è gratuito, versione 30 gg.)
 - Dopo averlo installato nel proprio pc, aprire il programma, **creare un account**
 - Dalla piattaforma scaricare:
 1. Licenza da inserire per usarlo illimitatamente (fino a luglio 2022)
 2. Guida di installazione Rapidminer, estensioni e connessioni
-



Rapidminer Studio 9.10



New process



1 Operators Building blocks used to create RapidMiner processes

2 Repositories Storage for data and processes

3 Process view (Main Process) Working area for building processes




4 Views Buttons for accessing specific functionality

5 Ports Input and output mechanisms for operators and processes

6 Parameters Settings that modify operator behavior

7 Help Context-sensitive help for selected operator.

Estensioni da installare

- Web mining 9.7 
 - Text processing 9.3.1 
 - MeaningCloud 2.1 
-
- Per installarle: menu **Extensions/Marketplace** cercare il nome e installare l'estensione una alla volta, alla fine di tutte le installazioni riavviare il programma



Connessione MeaningCloud API



-
- ▶ Estensione “MeaningCloud text analytics”
 - ▶ Dalla versione 9.3 di Rapidminer non si deve creare una connessione, ma inserire la Key nelle preferenze del software:
 - ▶ Menu Rapidminer Studio/Preferences: cercare la voce MeaningCloud e incollare la key nel campo vuoto

LINK:

<https://www.meaningcloud.com/developer/>

- ▶ N.B. Nelle versioni precedenti di Rapidminer invece la Key andava incollata come parametro degli operatori di MeaningCloud (es. Sentiment Analysis)
-
- ▶

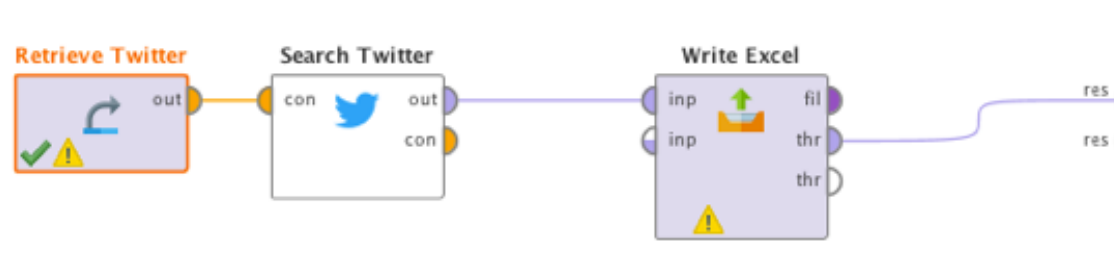
Scaricare tweets via API, connessione

- Per poter scaricare dei tweets e analizzarli in Rapidminer bisogna prima di tutto registrarsi in Twitter (avere un **account Twitter**)
 - In Rapidminer cliccare sul menu Connections/Create Connections e aggiungere una nuova connessione di tipo Twitter
 - Dare un nome alla nuova connessione (si consiglia: Twitter)
 - Access token: cliccare su tasto dx del campo testo
 - Richiedere un access token (un codice numerico)
 - Permettere a Rapidminer di accedere al vostro account Twitter
 - Copiare l'access token (numero) e incollarlo nella finestra della Connection di Rapidminer -> Complete
 - Testare la connessione
-



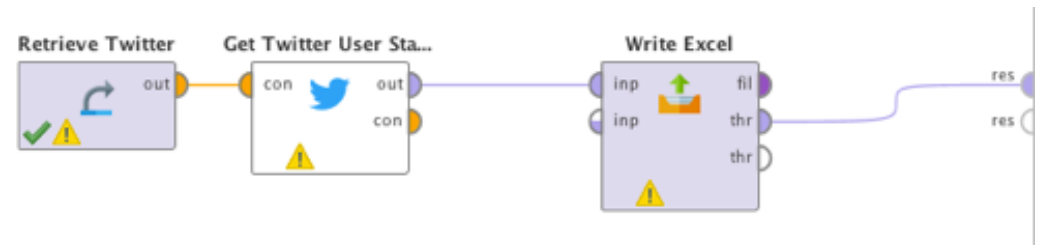
Processo di scarico di tweets (hashtag)

- ▶ Per scaricare tweets partendo da un hashtag:
 1. Trascinare la connection Twitter nell'area di lavoro (da Local Repository/Connections)
 2. Collegare operatore Search Twitter:
 - Parametro query: scrivere una keyword di ricerca (per es. apple) seguita da `-rt -http -https` (per evitare retweet e links)
 - Limit = 1000 (o numeri >)
 - Language = en (it o altro)
 3. Collegare operatore Write Excel
 - ▶ Dare un nome al file (cambiare nome ogni volta che si scarica)
 - ▶ 4. Collegare porta uscita dell'operatore
 - ▶ 5. Run



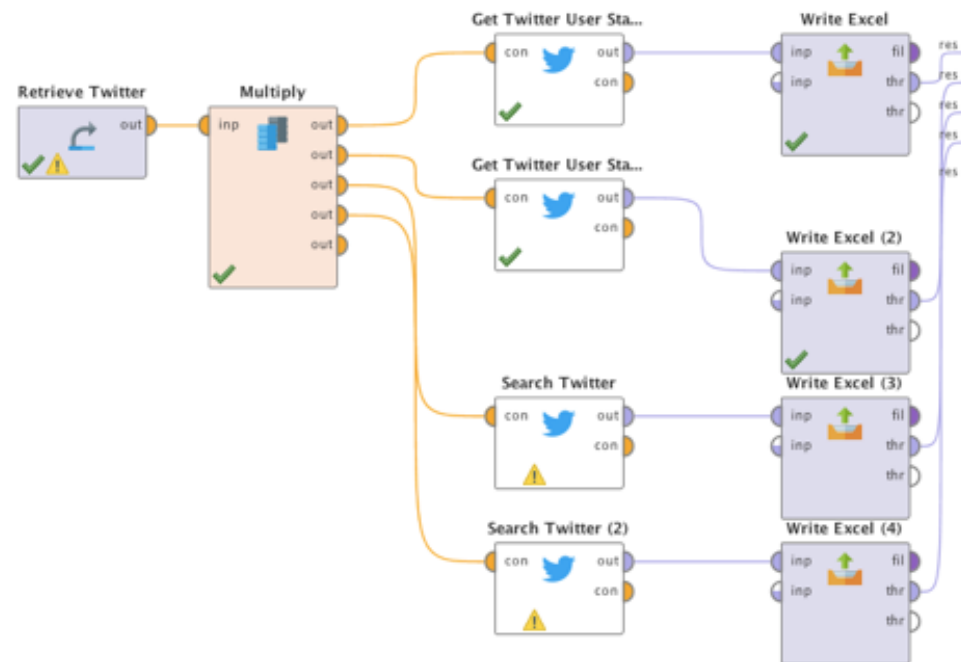
Processo di scarico di tweets (account)

- ▶ Per scaricare tweets partendo da un account:
 1. Trascinare la connection Twitter nell'area di lavoro (da Local Repository/Connections)
 2. Usare operatore Get Twitter User Statuses:
 - ▶ Parametro query type: scrivere il nome dell'account
 - ▶ Limit = 1000 (o numero >, max 3500)
 3. Collegare operatore Write Excel
 - ▶ Dare un nome al file (cambiare nome ogni volta che si scarica)
 4. Collegare porta uscita dell'operatore
 5. Run



Processi paralleli

- ▶ E' possibile scaricare contemporaneamente usando più hashtag e account in un unico processo replicando i flussi e cambiando per ogni flusso la query e il nome del file su cui salvare i dati, con un operatore Multiply che permette a più operatori di sfruttare un'unica connessione a Twitter. Ricordarsi sempre di cambiare nome ai file excel su cui si scrive (altrimenti sovrascrive sul precedente)

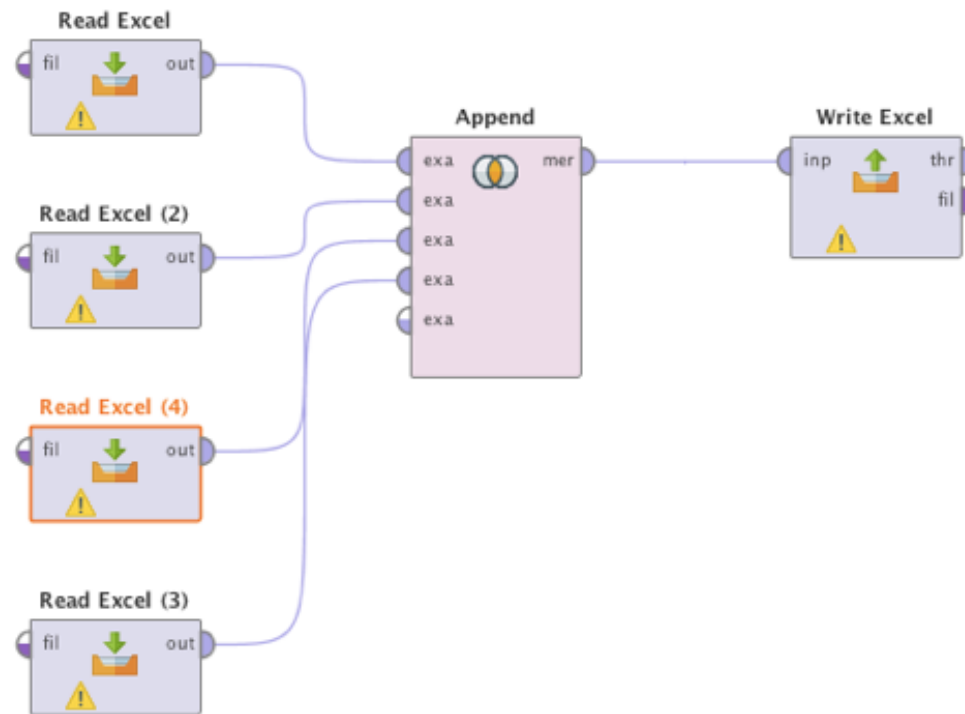


Visualizzare i dati

- Dopo aver cliccato tasto Run si vedono in Results una tabella di dati:
 - ▶ Example = 1 riga della tabella risultati
 - ▶ Example set = intero dataset
 - ▶ (Special attribute = la colonna verde)
- Tasto Statistics
- Tasto Charts = visualizzare dati con diversi grafici (istogrammi, scatter ecc.)



Unire i file excel



Letture dei dati

- Aprire File excel
- Colonna Text contiene i tweet
- Su questo testo applicheremo diverse analisi



Salvare il processo e i dati

- Save process as...
- In Local Repository:
 - ▶ Cartella data contiene i dataset importati
 - ▶ Cartella processes contiene i processi salvati (Store Process Here)
- I processi hanno estensione .rmp, possono essere importati ed esportati (da menu File/Import process)
- I processi sono anche salvati in un file xml che può essere copiato nel pannello XML ed eseguito (utile per passare un processo con i parametri da un pc a altro)

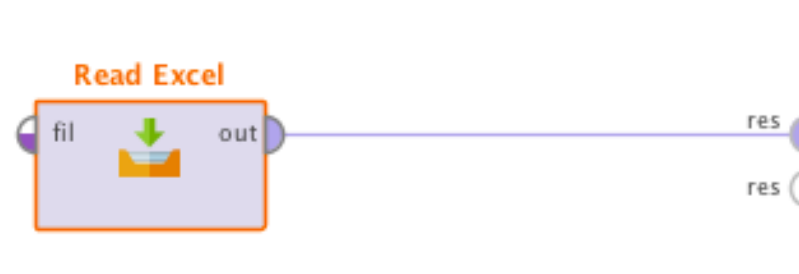


Caricare dati

Usiamo le estensioni Text processing e Web Mining

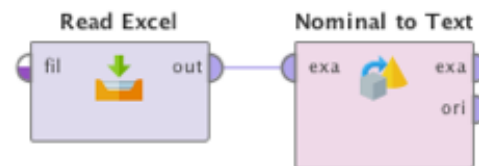
Utilizziamo un file excel con tweet scaricati e salvati (vd. slide precedenti)

- ▶ Per caricare dati:
 - ▶ operatore Read Excel
 - ▶ Settare parametri: Import Configuration Wizard -> selezionare celle da importare
 - ▶ Finish
- ▶ Eseguire processo non prima di aver collegato l'operatore alla porta Res



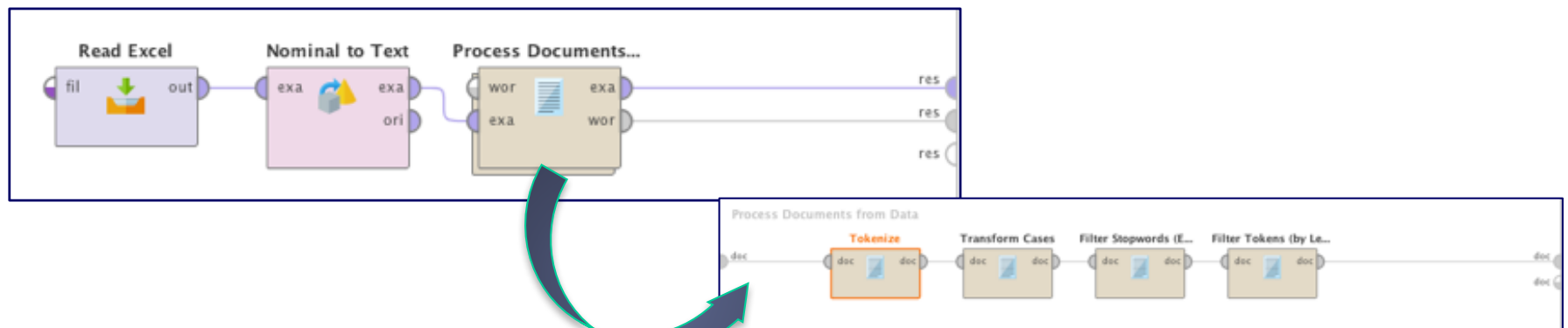
Nominal to Text

- ▶ Collegare alla porta uscita di Read Excel l'operatore Nominal to Text
- ▶ Settare parametri:
 - ▶ Attribute filter type: Single
 - ▶ Attribute: dal menu a tendina individuare il nome della colonna dei testi (es. Text)
- ▶ Fondamentale per poter passare i testi nel formato corretto all'operatore Process Documents from Data



Process Documents from Data

- ▶ Operatore Process Documents from Data (collegare la porta exa di NtT a exa di PDFD)
- ▶ Crea vettore di parole dai dati (lista parole), es. TF-IDF
- ▶ Necessita di sotto-processo con operatori di pre-processing (cliccando 2 volte sull'operatore)
- ▶ Collegare le porte di uscita dell'operatore Process Doc from Data a res



Parametri PDfD

Process Documents from Data

create word vector ⓘ

add meta information ⓘ

keep text ⓘ

prune method ⓘ

prune below absolute ⓘ

prune above absolute ⓘ

datamanagement ⓘ

select attributes and weights ⓘ

create word vector: crea vettore, es. TF-IDF

add meta information: aggiunge metadati come colonne nei risultati

keep text: testo è preso come Special attribute nei risultati

prune below absolute/percentual: Ignora le parole che appaiono in meno di un range di documenti

prune above absolute/percentual: Ignora le parole che appaiono in più di un range di documenti



Pre-processing

- ▶ Tokenize -> Run (prime osservazioni freq parole)
- ▶ Transform Cases: lower cases (trasforma tutto in minuscolo)
- ▶ Filter stopwords (English)
 - ▶ Filter stopwords (Dictionary) per italiano (+ file stopwords-ita.txt da scaricare dalla piattaforma)
- ▶ Filter tokens (by Length):
 - ▶ Param: min 2 max 25
- ▶ Run (eliminate stopwords)



Analisi occorrenze parole e n-grammi

Cerchiamo frasi significative nei testi processati: n-grammi (serie di tokens consecutivi di lunghezza n) composti da 2 o + parole

Aggiungere nel sotto-processo di pre-processing:

- ▶ Operator Generate n-Grams (Terms), posto tra Stopwords e Filter tokens
 - ▶ Param: max lenght 3
- ▶ Run
- ▶ Opzionale: operator Stem (Porter)



Parametri PDfD (2)

Process Documents from Data

create word vector

vector creation Binary Term Occ

add meta information

keep text

prune method absolute

prune below absolute 2

prune above absolute 9999

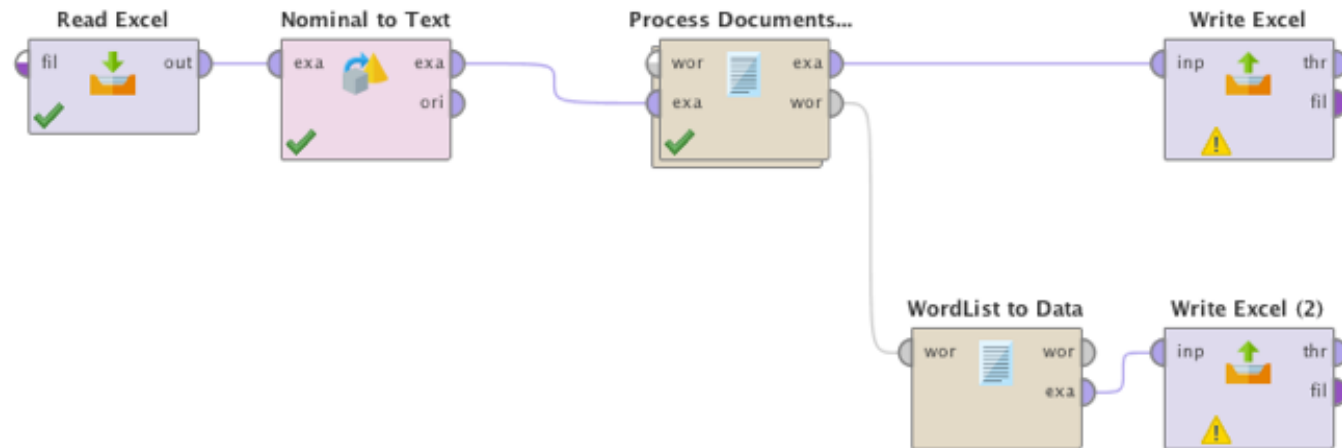
datamanagement double_sparse_a...

Row No.	text	ab	abbotsford	abstract	abilities	ability	ability_acc...	ability_anti...
1	ogplanet og	0	0	0	0	0	0	0
2	company co	0	0	0	0	0	0	0
3	ready ready	0	0	0	0	0	0	0
4	due due_raj	0	0	0	0	1	0	0
5	growing gro	0	0	0	0	1	0	0
6	direct direc	0	0	0	0	0	0	0
7	hour hour_r	0	0	0	0	0	0	0



Salvataggio dei risultati

- ▶ Utilizzando operatore Write Excel possiamo salvare su file excel i risultati di tutti i processi, sia l'ExampleSet sia la Wordlist (con operatore WordList to Data)
- ▶ **N.B. Se l'operatore Write Excel non dovesse funzionare, perché l'excel da scrivere supera il limite di 16000 colonne, sostituirlo con Write CSV**

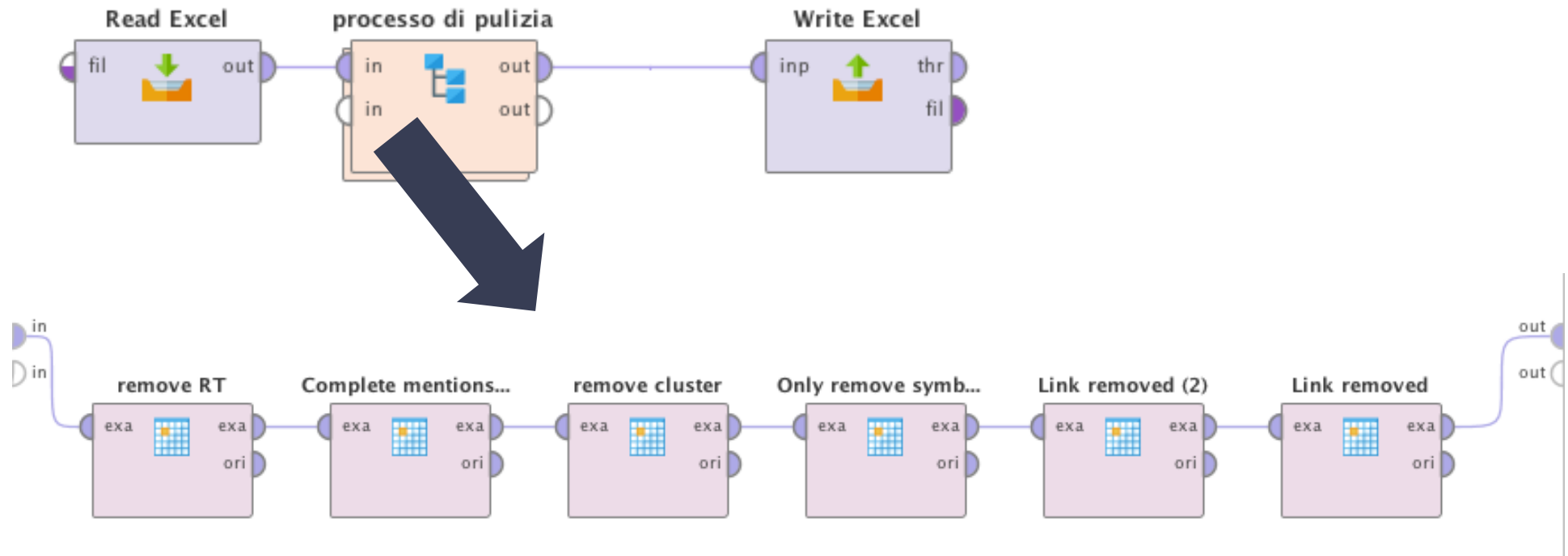


Pulire tweets: problema di partenza

- ▶ Tweets scaricati e salvati in un file excel possono essere caratterizzati da molte mentions (@...) e/o hashtags (#...) e/o links (http://...) che disturbano la lettura del testo
- ▶ Obiettivo è pulire per conservare solo la parte di testo significativa



Processo con operatore Replace



N.B. L'operatore Subprocess è stato rinominato "processo di pulizia".
All'interno del processo di pulizia l'operatore Replace si ripete n volte ed è stato rinominato per evidenziare cosa rimuove ogni operatore.



Rimuovere parti di un tweet

Parametri comuni a tutti gli operatori Replace:

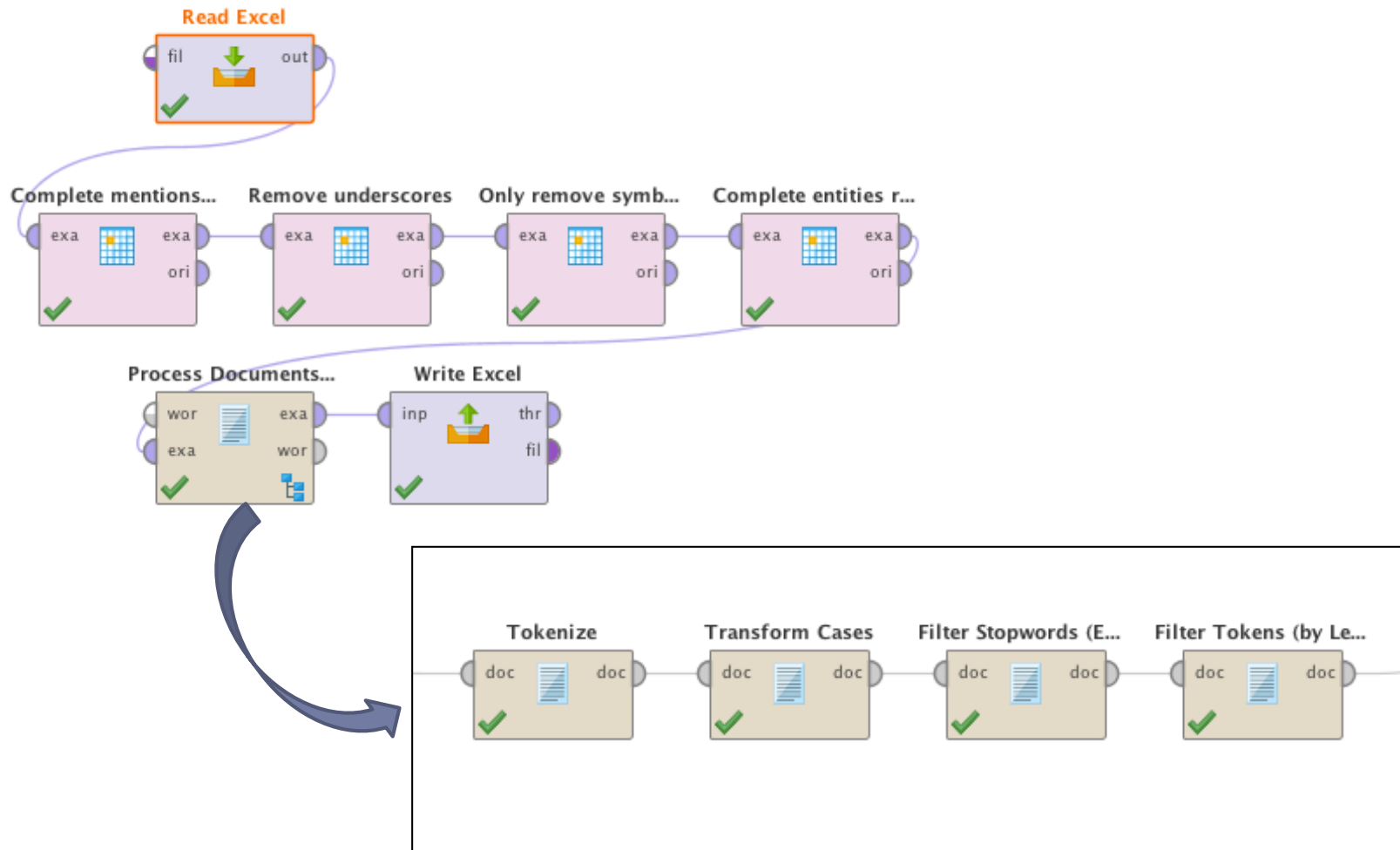
- ▶ Attribute filter type: Single
- ▶ Attribute: Text (il nome della colonna del testo)

Parametri diversi per singolo Replace:

1. Replace what: RT rimuove la stringa all'inizio dei retweet
2. Replace what: @[a-zA-Z0-9/d\-_]* rimuove le mentions
3. Replace what: # rimuove il cancelletto degli hashtags
4. Replace what: http://[a-zA-Z0-9/d\-_].* rimuove link
5. Replace what: https://[a-zA-Z0-9/d\-_].* rimuove link
6. Replace what: [cC]luster rimuove token cluster e lo sostituisce con una stringa "WWWWW" per riconoscerlo nel testo



Aggiunta di pre-processing (opzionale)



Finale

PRIMA:

@musicassetta A ognuno secondo i propri bisogni

<https://t.co/r8aAB8MAqR>

DOPO:

A ognuno secondo i propri bisogni



Alcuni suggerimenti

- ▶ La sentiment è meglio applicarla dopo aver ripulito il tweet ma senza il pre-processing
- ▶ Altre analisi di text mining (word occurrence, TF-IDF, similarità, clustering ecc.) è meglio applicarle dopo la fase di pre-processing

