



Business Analytics

Metodi statistici per il Decision Making

Text Analytics





Copyright© 2025 Alfredo Roccatò. Tutti i diritti riservati.

I testi, le immagini e la grafica qui presenti sono protetti ai sensi delle normative vigenti sul diritto d'autore, sui brevetti e sulla proprietà intellettuale. È vietata la riproduzione anche parziale e con qualsiasi mezzo senza l'autorizzazione scritta dell'autore.

Per informazioni sui permessi per riprodurre parti del presente lavoro, inviare un messaggio e-mail ad Alfredo Roccatò all'indirizzo alfredo.roccato@fastwebnet.it. Si prega di indicare quali pagine si desidera utilizzare e per quale scopo.

Questo libro è stato aggiornato per il software KNIME® Analytics Platform (Versione 4.5.2 e superiori), R (Versione 4.2.0 e superiori).



- **Concetti base**
- Raccolta dati
- Preparazione
- Trasformazione
- Analisi



■ Scopi

L'obiettivo del Text Analytics è quello di ottenere informazioni rilevanti nel **testo** trasformandolo in **dati** che possono essere usati per ulteriori analisi.

Tale obiettivo viene raggiunto attraverso l'utilizzo di diverse metodologie di analisi¹ per

- ✓ Analizzare fonti testuali
- ✓ Strutturare e classificare automaticamente il contenuto
- ✓ Ottenere informazioni rilevanti con l'utilizzo di tecniche di analisi dati

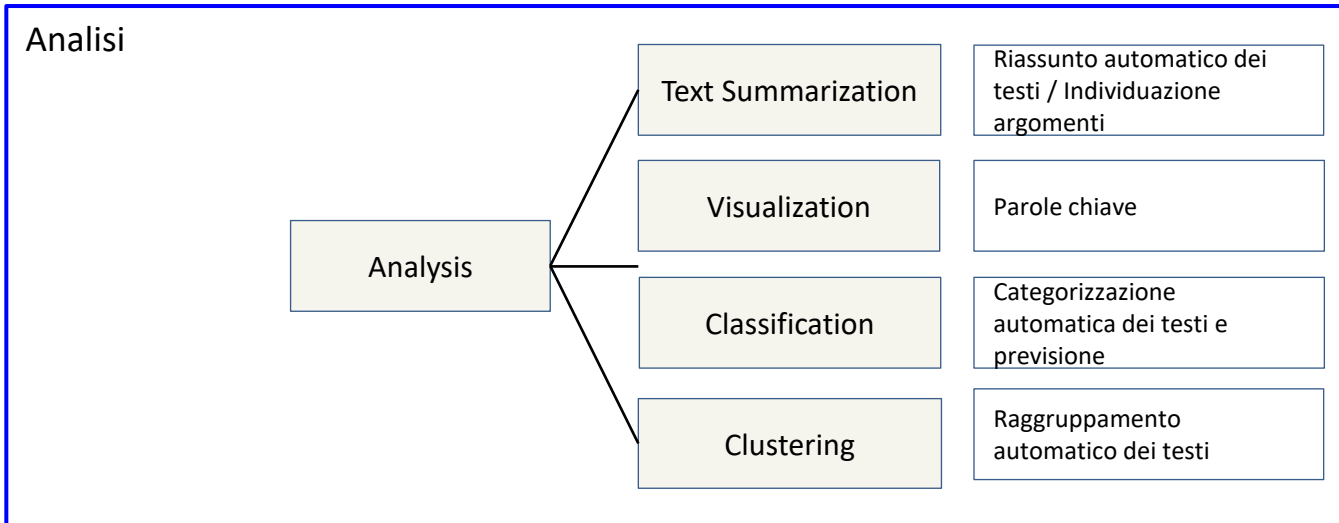
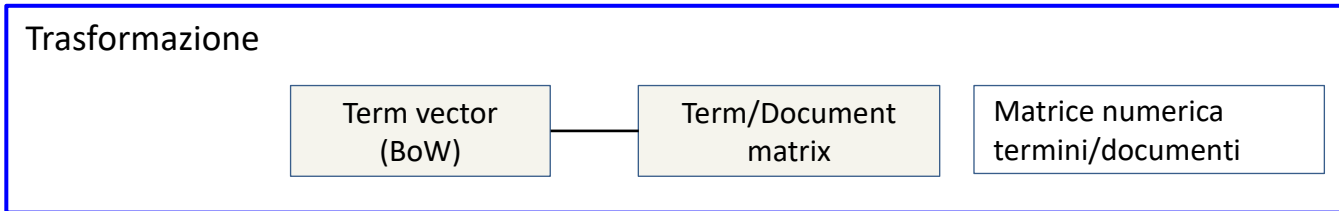
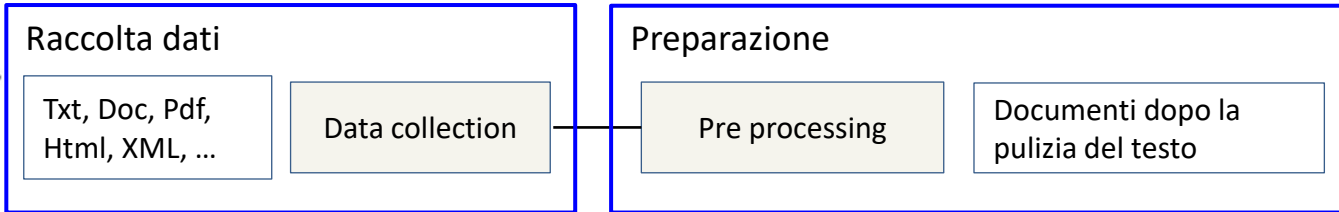
¹ Una di queste è il **Natural Language Processing (NLP)** che è il processo di trattamento automatico delle informazioni scritte in lingua naturale attraverso diverse fasi di **analisi (lessicale, grammaticale, sintattica e semantica): quest'ultima non viene trattata in queste note.**



■ Applicazioni

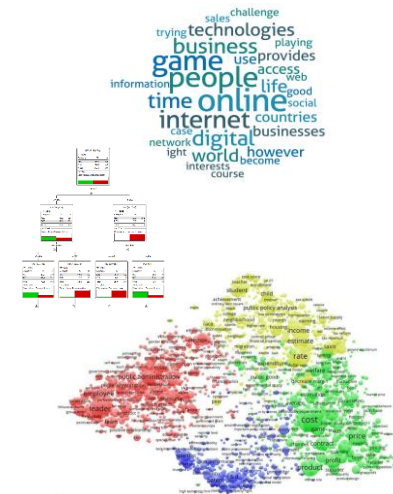
- Customer Care/ Retention
 - ✓ Ottimizzare la qualità e l'efficienza nel risolvere i problemi dalle richieste di assistenza raccolte tramite **e-mail, sms e on-line.**
 - ✓ Individuare i clienti in procinto di **abbandono**
- Marketing e e-commerce
 - ✓ **Ricerche di marketing**
 - ✓ Misurazione dell'efficacia delle azioni di **comunicazione**
- Gestione documentale
 - ✓ Organizzazione e **classificazione automatica di documenti**
- Analisi dei Social Media
 - ✓ Identificazione automatica dei **temi emergenti**
 - ✓ Brand Reputation (**Sentiment Analysis**)

Introduzione



Doc	Text			
c1	Human machine interface for Lab ABC computer applications			
c2	A survey of user opinion of computer system response time			
c3	The EPS user interface management system			
c4	System and human system engineering testing of EPS			
c5	Relation of user-perceived response time to error measurement			
m1	The generation of random, binary, unordered trees			
m2	The intersection graph of paths in trees			
m3	Graph minors IV: Widths of trees and well-quasi-ordering			
m4	Graph minors: A survey			

Term	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	1	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1





- **Concetti base**
- Raccolta dati
- Preparazione
- Trasformazione
- Analisi



- **Corpus**

È l'insieme dei testi che si vogliono analizzare. È composto da diversi documenti (**document**).

Corpus

Non è bello ciò che è bello, ma è bello ciò che piace

Chi va piano va sano e va lontano

- **Document**

È l'unità di testo che appartiene al Corpus, corrisponde a una riga di una tabella. Al suo interno sono contenuti dei blocchi di testo chiamati **token**.

Document

Row1	Non è bello ciò che è bello, ma è bello ciò che piace
Row2	Chi va piano va sano e va lontano



■ Token

È l'unità di base che costituisce il lessico di una lingua (*lessema*). È formato da una stringa di caratteri separati tra loro da uno spazio o da un segno di punteggiatura.

Un token può corrispondere a una parola o a un gruppo di parole (es. "*Banca d'Italia*"), a un numero, a una data, a un simbolo.

Il processo di **pulizia e normalizzazione**¹ dei token porta alla creazione dei termini (**term**).

Document		
Row1	non bello ciò che bello ma bello ciò che piace	10 token
Row2	chi va piano va sano va lontano	7 token

¹ Di solito dopo la conversione in maiuscolo/minuscolo, la rimozione della punteggiatura e le parole di poche lettere (in questo esempio, 1).



■ Term

È un elemento della lista dei token unici normalizzati presenti in un documento.

Nell'esempio utilizzato, nel primo documento si hanno 10 token e 6 term (con le rispettive occorrenze).

	Document	TERM	
Row1	non bello ciò che bello ma bello ciò che piace	non	1
	non bello ciò che bello ma bello ciò che piace	bello	3
	non bello ciò che bello ma bello ciò che piace	ciò	2
	non bello ciò che bello ma bello ciò che piace	che	2
	non bello ciò che bello ma bello ciò che piace	ma	1
	non bello ciò che bello ma bello ciò che piace	piace	1

A ogni termine possono essere associate delle etichette (**Tag**).



■ Tag (1/2)

È la categoria lessicale che viene attribuita ogni termine per classificarlo. La più nota è quella di tipo grammaticale, chiamata POS (*Part-of-Speech*, parte del discorso).

Sono disponibili diversi modelli di **POS Tagger**. Questa, ad esempio, è una lista parziale dei tag del *Penn Treebank tag set*¹:

NN	Sostantivo	RB	Averbio	IN	Prepos./Cong. Subord.
JJ	Aggettivo	NNP	Nome proprio	PRP	Pronome
VB*	Verbo (forma base)	CC	Congiunz. Coord.	DT	Articolo

Ad esempio:

ai termini della frase *the black dog barks at the moon*

vengono associati i tag *the[DT] black[JJ] dog[NN] barks[VBZ] at[IN] the[DT] moon[NN]*

¹ Per approfondimenti: <https://www.cs.cmu.edu/afs/cs/Web/People/dgovinda/pdf/semantics/tagguide.pdf/>





■ Tag (2/2)

A ogni termine si possono contrassegnare altre categorie di Tag che rappresentano entità (**NE, Named Entities**¹) come persone, luoghi, società, date, ... oppure opinioni (**Sentiment**²).

TOKEN	TAG		
	POS	NE	SENTIMENT
Great	[JJ]		[POSITIVE]
Rome	[NNP]	[LOCATION]	
Julius Caesar		[PERSON]	

¹ Apache OpenNLP library. Per approfondimenti: <https://opennlp.apache.org/>

² MPQA Opinion Corpus. Per approfondimenti: http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/



■ Stop Words

Sono parole che non sono portatrici di significato autonomo in quanto sono elementi necessari alla costruzione della frase e quindi possono essere rimosse. Esempi tipici sono gli **articoli**, gli **avverbi**, le **proposizioni**, gli aggettivi indefiniti, i **verbi ausiliari** (essere, avere, andare, venire) e **verbi servili** (dovere, potere, sapere, sembrare, volere).

La rimozione avviene mediante un filtraggio basato su un'apposita lista.



👉 L'utilizzo delle stop word può far perdere informazioni rilevanti. Bisogna perciò valutare se ricorrere eventualmente a normalizzazioni preliminari (parole composte).

Ad esempio: Il grande freddo è un film del 1983 → *freddo film 1983*





■ Lemmatizzazione

• Lemmatization

La lemmatizzazione è il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta lemma. Ad esempio, la forma canonica del verbo **camminare** è il **lemma** delle parole **cammina**, **camminò**, **camminando**, ...

• Stemming

È il processo di riduzione della forma flessa di una parola alla sua forma radice, detta tema¹. La radice si ottiene rimuovendo affissi e desinenze. È utile in quanto riduce le varianti di una stessa parola a un concetto comune. Ad esempio, con **rid-ere**, **rid-endo**, **rid-eva** si ottiene **rid**; con **mand-are**, **mandar-gli**, **mandar-gli(e)-lo** si ottiene **mand**.



Lo stemming è soggetto a errore:

- Over-stemming

quando due termini vengono uniti insieme con perdita di specificità
"università", "universo" → "univers"

- Under-stemming due termini simili non vengono uniti con perdita di generalità

"falso", "falsificato" → "fals", "falsific"

¹ Per approfondimenti <http://snowball.tartarus.org/algorithms/italian/stemmer.html>



- Concetti base
- **Raccolta dati**
- Preparazione
- Trasformazione
- Analisi



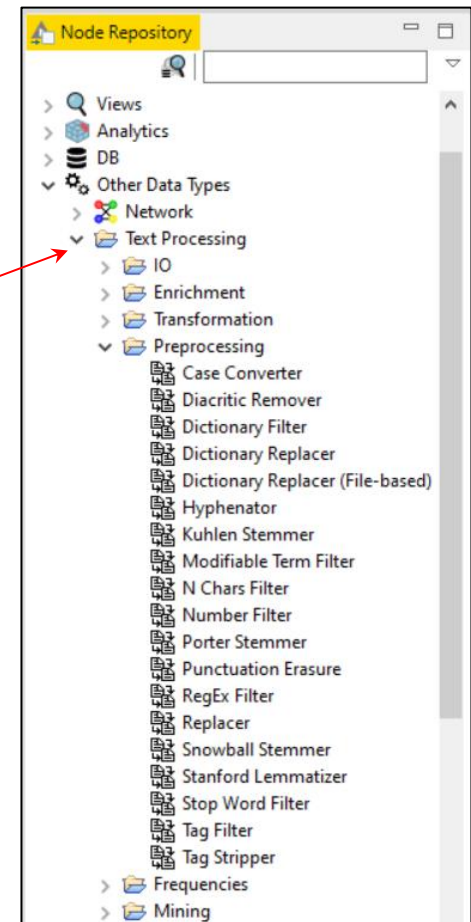
■ Software

Queste note si basano sull'utilizzo del software open-source KNIME Analytics Platform



Analytics Platform

attraverso i suoi nodi specifici per il text processing e l'installazione del toolkit java-based **Palladian**¹



¹ Necessario per eseguire le tipiche attività di recupero delle informazioni su Internet, vedi [Appendice](#)



■ Accesso ai dati

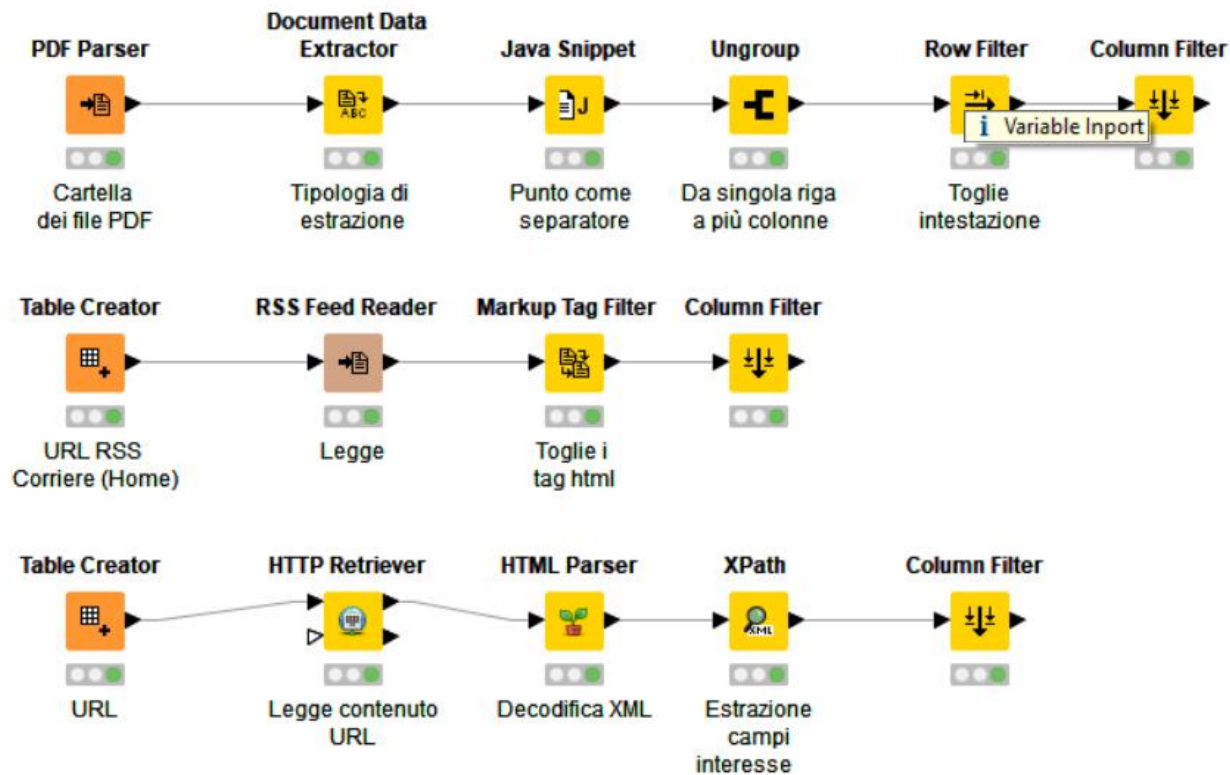
In KNIME sono presenti diversi nodi per accedere al testo da analizzare creando una tabella dove ogni riga è un documento.

	<i>Funzione</i>	<i>Sorgente</i>
✓	Nodo Table Reader	KNIME
✓	Nodo File Reader	Txt, Csv
✓	Nodo CSV Reader	Csv
✓	Nodo Excel Reader (XLS)	Xls, Xlsx
✓	Nodo Word Parser	Doc, Docx
✓	Nodo PDF Parser	Pdf
✓	Nodo XML	Xml
✓	Nodi Palladian	Html

Creazione del corpus



■ Utilizzo del software KNIME – Modulo3_Esempi_Raccolta_Dati





- **Concetti base**
- **Raccolta dati**
- **Preparazione**
 - Segmentazione dei termini
 - Pulizia
 - Assegnazione etichette
 - Troncamento
- **Trasformazione**
- **Analisi**



■ Segmentazione

- Conversione della colonna stringa contenente il testo in una colonna di tipo “documento”
- ✓ **Nodo Strings To Document**
Si indicano, se presenti, le colonne contenenti
 - il titolo¹
 - il testo
 - le fonti dei documenti
 - gli autori
 - la categoria del documento
 - il tipo del documento
 - la data di pubblicazione
 - modalità di suddivisione delle parole (Tokenization)



¹ Se presente, vengono elaborate anche le parole del titolo.



■ Segmentazione

- Conversione della colonna stringa contenente il testo in una colonna di tipo “documento”

✓ Nodo Strings To Document

- Modalità di suddivisione delle parole (**Tokenization**)

KNIME utilizza i tokenizer delle librerie open-source Apache **OpenNLP**¹ e **StanfordNLP**.

Strings To Document

Converti testo in documento (tokenizzatore Simple)

Text

Use title from column Title column **S** column2

Use authors from column Authors column **S** column2

Author names separator ,

Default author first name - Default author last name -

Full text **S** column2

Source and Category

Document source

Use sources from column Document source column **S** column2

Document category

Use categories from column Document category column **S** column2

Type and Date

Document type UNKNOWN

Publication date (dd-mm-yyyy) 27-02-2017

Use publication date from column Publication date column **S** column2

Processes

Number of maximal parallel processes 2

Tokenization

Word tokenizer **OpenNLP English WordTokenizer**

OpenNLP English WordTokenizer

OpenNLP English WordTokenizer

OpenNLP SimpleTokenizer

OpenNLP WhitespaceTokenizer

StanfordNLP PTBTokenizer

¹ <https://opennlp.apache.org/docs/1.5.3/manual/opennlp.html#tools.tokenizer.introduction>



■ Segmentazione



- Conversione della colonna stringa contenente il testo in una colonna di tipo “*documento*”

✓ Nodo **Strings To Document**

- Modalità di suddivisione delle parole (**Tokenization**)

Il numero di token varia a seconda del tipo di tokenizer utilizzato:

La | Banca | d'Italia | ha | abbassato | il | tasso | d'interesse | di | due | punti, | portandolo | dall'8 | al | 6%

OpenNLP **Whitespace**Tokenizer: 15 token

La | Banca | d | ' | Italia | ha | abbassato | il | tasso | d | ' | interesse | di | due | punti | , | portandolo | dall | ' | 8 | al | 6 | %

OpenNLP **Simple**Tokenizer¹: 21 token

¹ Negli esempi che seguono verrà utilizzato, dove non indicato, questo tipo di Tokenizer.



■ Pulizia

- Conversione parole Minuscole/Maiuscole

- ✓ **Nodo Case Converter**

Converte tutti i termini contenuti nel documento in maiuscolo o minuscolo



La | Banca | d | ' | Italia | ha | abbassato | il | tasso | d | ' | interesse | di | due | punti | , | portandolo | dall | ' | 8 | al | 6 | %

la | banca | d | ' | italia | ha | abbassato | il | tasso | d | ' | interesse | di | due | punti | , | portandolo | dall | ' | 8 | al | 6 | %



■ Pulizia

- Rimozione punteggiatura

- ✓ **Nodo Punctuation Erasure**

Toglie tutti i caratteri di punteggiatura



la | banca | d | ' | italia | ha | abbassato | il | tasso | d | ' | interesse | di | due | punti | , | portandolo | dall | ' | 8 | al | 6 | %
↳ *la | banca | d | italia | ha | abbassato | il | tasso | d | interesse | di | due | punti | portandolo | dall | 8 | al | 6*

- ✓ **Nodo Replacer**

Sostituisce tutti i caratteri che soddisfano un'espressione regolare (p.e. quella usata nel nodo Punctuation Erasure: `[!#$%&'()*+,-./:;<=>?@^_`{|}~\|/]+`) con quello specificato. In questo esempio, uno spazio:

la | banca | d | ' | italia | ha | abbassato | il | tasso | d | ' | interesse | di | due | punti | , | portandolo | dall | ' | 8 | al | 6 | %
↳ *la | banca | d | _ | italia | ha | abbassato | il | tasso | d | _ | interesse | di | due | punti | _ | portandolo | dall | _ | 8 | al | 6 | _*



■ Pulizia

- Rimozione numeri

- ✓ **Nodo Number Filter**

- Toglie i caratteri numerici (inclusi separatori e i segni ±)



la | banca | d | italia | ha | abbassato | il | tasso | d | interesse | di | due | punti | portandolo | dall | 8 | al | 6

la | banca | d | italia | ha | abbassato | il | tasso | d | interesse | di | due | punti | portandolo | dall | al

- ✓ **Nodo RegEx Filter**

- Toglie i caratteri che soddisfano un'espressione regolare (p.e. quella usata all'interno del nodo Number Filter: `*\d+.*`).

la | banca | d | italia | ha | abbassato | il | tasso | d | interesse | di | due | punti | portandolo | dall | 8 | al | 6

la | banca | d | italia | ha | abbassato | il | tasso | d | interesse | di | due | punti | portandolo | dall | al



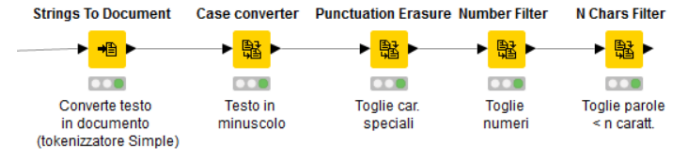
■ Pulizia

- Rimozione termini di breve lunghezza

✓ Nodo **N Chars Filter**

Toglie tutti i termini con meno di N caratteri (articoli, preposizioni e altro testo non informativo). In questo esempio meno di 3 caratteri:


la | banca | d | italia | ha | abbassato | il | tasso | d | interesse | di | due | punti | portandolo | dall | al
↳ *banca | italia | abbassato | tasso | interesse | due | punti | portandolo | dall*



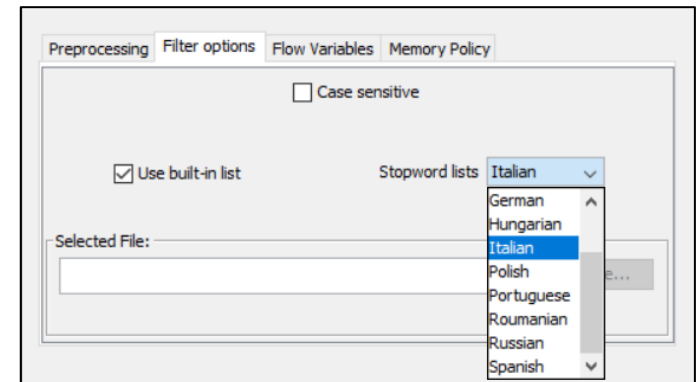
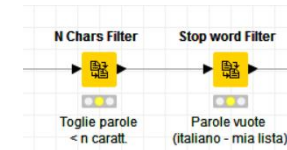


■ Pulizia

- Rimozione delle "parole non significative" (*stop word*)

✓ Nodo **Stop word Filter** 
Toglie tutte le stop word

- Si può usare la lista fornita da KNIME
- Si può usare una lista propria¹



banca | italia | abbassato | tasso | interesse | **due** | punti | **portandolo** | **dall**
↙
banca / italia / abbassato / tasso / interesse / punti

¹ Una lista di stop-word per la **lingua italiana** viene fornita dall'autore delle presenti note.



■ Pulizia

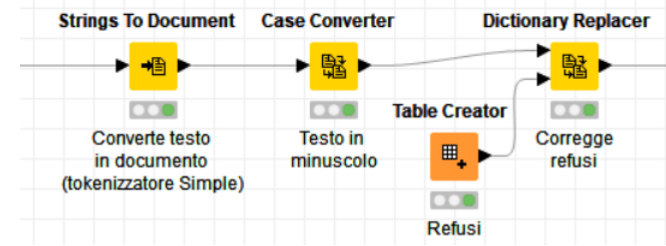
- Sostituzione dei termini

- ✓ **Nodo Dict Replacer**

Sostituisce i termini presenti nel documento che corrispondono ai termini contenuti in un file esterno¹ con il valore ivi specificato.



Utile, ad esempio, per gli errori grammaticali.



Row ID	S Col0	S Col1
Row40	purche	purché
Row41	purchè	purché
Row42	quà	qua
Row43	qui	qui
Row44	qual'è	qual è
Row45	qualè	qual è
Row46	sà	sa
Row47	sè	sé
Row48	sicche	sicché
Row49	sicchè	sicché
Row50	sò	so
Row51	soprattutto	soprattutto

Row ID	S Testo	S Argomento	Document	Preprocessed Document
Row11	UN'altro qual'è purchè qualè	Errori Ortografia	"UN'altro qual'è purchè qualè"	"un altro qual è purché qual è"
Row4	soprattutto	Errori Ortografia	"soprattutto"	"soprattutto"
Row15	un pò deludente	Errori Ortografia	"un pò deludente"	"un po' deludente"

¹ Una lista di **errori grammaticali** più comuni viene fornita dall'autore delle presenti note.



■ Assegnazione dei "TAG"

- di tipo **POS** (Part of Speech)

✓ Nodo [POS Tagger](#)

✓ Nodo **Stanford Tagger**

✓ *Metanodo* **Italian POS Tagger**



Inglese

Inglese, tedesco, spagnolo e francese

Italiano¹

- di tipo **NE** (Named Entity)

✓ Nodo **OpenNLP NE Tagger**

Inglese

✓ Nodo **StanfordNLP NE Tagger**

Inglese, tedesco e spagnolo

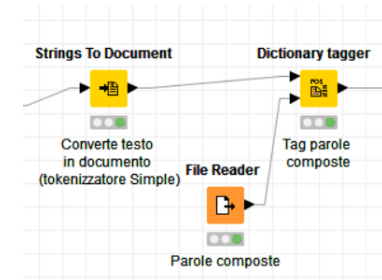
¹ Questo **metanodo** viene fornito dall'autore delle presenti note.



■ Assegnazione dei "TAG"

- di tipo **MWT** (Multi-word term)

Parole composte



General options | Tagger options | Flow Variables | Memory Policy

Dictionary column: \$ Col0

Set named entities unmodifiable

Case sensitive | Exact match

Tag type: MWT | Tag value: MULTIWORDTERM

Row ID	\$ Col0
Row0	Text Analytics
Row1	Banca d'Italia

La | Banca | d | ' | Italia | ha | abbassato | il | tasso | d | ' | interesse | di | due | punti | , | portandolo | dall | ' | 8 | al | 6 | %

↳ **Banca d'Italia [MULTIWORDTERM(MWT)] / abbassato / tasso / interesse / punti / portandolo**

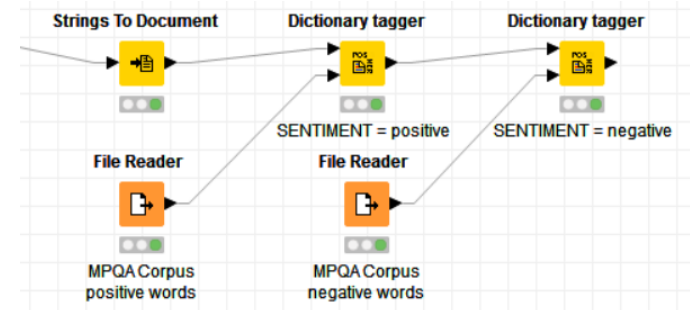


■ Assegnazione dei "TAG"

- di tipo **SENTIMENT**

- ✓ Nodo **Dictionary tagger**

Opinioni, ...



In questo esempio vengono creati i Tag dei termini secondo gli standard SENTIMENT¹ (POSITIVE, NEGATIVE, NEUTRAL, IRONY, ...):

General options
Tagger options

Dictionary column S Col0

Set named entities unmodifiable

Case sensitive Exact match

Tag type SENTIMENT Tag value POSITIVE

VERY_POSITIVE
POSITIVE
 NEUTRAL
 NEGATIVE
 VERY_NEGATIVE
 UNDERSTATEMENT
 EXAGGERATION
 IRONY

Row ID	T Term	Document
Row 118	wrong[NEGATIVE(SENTIMENT)]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 119	cheaper[POSITIVE(SENTIMENT)]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 120	saleperson[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 121	not[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 122	very[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 123	friendly[POSITIVE(SENTIMENT)]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 124	instead[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 125	otter[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 126	salesperson[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 127	leo[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 128	who[]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly
Row 129	helped[POSITIVE(SENTIMENT)]	"the pricetag the silver highheels was wrong cheaperthe saleperson was not very friendly

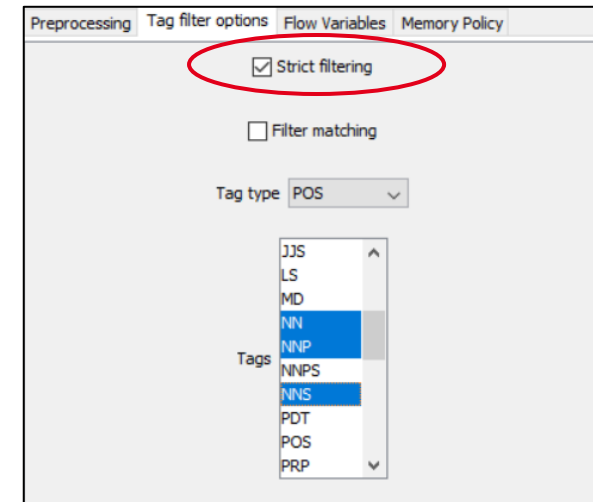
¹ Un vocabolario con la positività/negatività dei termini in italiano viene fornito dall'autore delle presenti note.



■ Assegnazione dei "TAG"

- Filtraggio dei Tag
- ✓ Nodo **Tag Filter**

Serve a filtrare i Tag di interesse (possono essercene o solo loro). In questo esempio, solo loro (Strict filtering).



Document	T Term
"la viola e il violino sono archi"	la[DT(POS)]
"la viola e il violino sono archi"	viola[VB(POS) NN(POS) NNP(POS)]
"la viola e il violino sono archi"	e[CC(POS)]
"la viola e il violino sono archi"	il[DT(POS)]
"la viola e il violino sono archi"	violino[VB(POS) NN(POS)]
"la viola e il violino sono archi"	sono[VB(POS)]
"la viola e il violino sono archi"	archi[NNS(POS)]

Document	T Term	Prepr...
"la viola e il violino sono archi"	archi[NNS(POS)]	"archi"
"la viola e il violino sono archi"	sono[VB(POS)]	"archi"
"la viola e il violino sono archi"	violino[VB(POS) NN(POS)]	"archi"
"la viola e il violino sono archi"	il[DT(POS)]	"archi"
"la viola e il violino sono archi"	e[CC(POS)]	"archi"
"la viola e il violino sono archi"	viola[VB(POS) NN(POS) NNP(POS)]	"archi"
"la viola e il violino sono archi"	la[DT(POS)]	"archi"



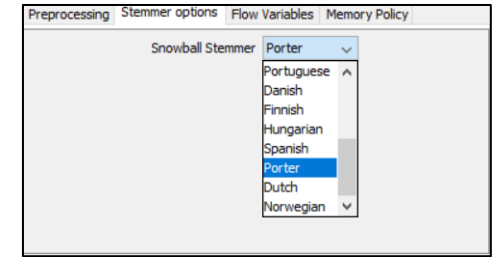
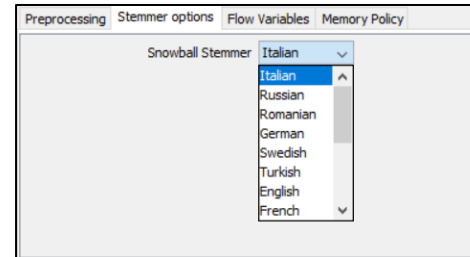
■ Troncamento



- Riduzione dei termini alla loro radice grammaticale ([Stemming](#))

✓ Nodo **Snowball Stemmer**

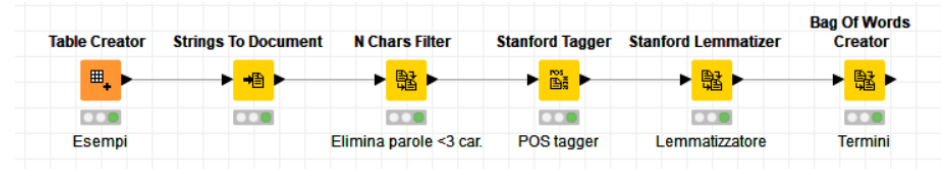
- Dipendente dalla lingua
- Porter (solo inglese)



Il |12 | GENNAIO | 2002 | l'Euro | diventa | moneta | corrente | in | 12 | paesi | dell'Unione | Europea
↙
gennai | 2002 | eur | divent | monet | corrent | paes | union | europe



■ Troncamento



- Riduzione dei termini alla loro radice grammaticale ([Lemmatization](#))
 - ✓ Nodo **Stanford Lemmatizer**
 - Disponibile solo per la lingua inglese

S Text	T ▼ Term
A mouse lives in my house.	mouse [NN(POS)]
A mouse lives in my house.	live [VBZ(POS)]
A mouse lives in my house.	house [NN(POS)]
There are mice living in my houses.	there [EX(POS)]
There are mice living in my houses.	be [VBP(POS)]
There are mice living in my houses.	mouse [NNS(POS)]
There are mice living in my houses.	live [VBG(POS)]
There are mice living in my houses.	house [NNS(POS)]



■ Utilizzo del software KNIME – Modulo3_Esercizio1 (parte 1)

- Importare con il nodo **Excel Reader (XLS)** dalla cartella Dati nella chiavetta Usb la tabella ***SmartCard.xlsx***;
- convertire le righe in documenti con il nodo **Strings to Document** usando la colonna **TESTO** nei campi Title e Text; selezionare come Word Tokenizer ***OpenNLP Simple Tokenizer***;
- convertire il testo in caratteri minuscoli con il **Case Converter**;
- togliere tutti i caratteri della punteggiatura con il nodo **Punctuation Erasure**;
- togliere i termini che hanno meno di 5 caratteri con il nodo **N Chars Filter**;
- togliere i termini "inutili" con il nodo **Stop Word Filter**, dalla lista contenuta nel file ***Stop_words_it.txt*** letto con il nodo **File Reader**;



- Concetti base
- Raccolta dati
- Preparazione
- **Trasformazione**
 - Costruzione matrice termini/documenti
- Analisi



■ Costruzione della matrice termini-documenti

- Costruzione del vettore dei termini (1/3)

Il corpus dei documenti finora ottenuto è una matrice dove le righe sono documenti e le colonne sono i termini presenti nel corpus.

Per poter utilizzare gli strumenti di analisi è necessario trasformare questo insieme in un vettore numerico. Questo processo viene chiamato **vectorization** (o vettorizzazione).

Prima il corpus viene trasformato in una lista di termini (**Bag of words** o **BoW**) che costituisce il "*vocabolario*" dei termini usati in base al quale viene calcolata la frequenza con la quale è presente ogni token in ciascun documento (**Term Frequency** o **TF**).



■ Costruzione della matrice termini-documenti

- Costruzione del vettore dei termini (2/3)

Ad esempio, da questi due documenti

Doc1: A Giovanni piacciono i film. Anche a Maria piacciono i film.

Doc2: A Giovanni piacciono anche le partite.

viene creata la lista dei termini unici (Bag of Words) che compaiono in ogni documento alla quale verranno associate le relative frequenze.

	Termine	Frequenza
<i>Doc1:</i>	a	2
<i>Doc1:</i>	giovanni	1
<i>Doc1:</i>	piacciono	2
<i>Doc1:</i>	i	2
<i>Doc1:</i>	film	2
<i>Doc1:</i>	anche	1
<i>Doc1:</i>	maria	1
<i>Doc2:</i>	a	1
<i>Doc2:</i>	giovanni	1
<i>Doc2:</i>	piacciono	1
<i>Doc2:</i>	anche	1
<i>Doc2:</i>	le	1
<i>Doc2:</i>	partite	1

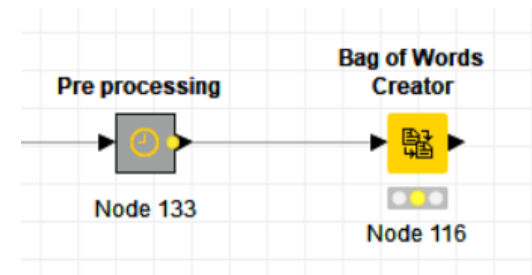


■ Costruzione della matrice termini-documenti

- Costruzione del vettore dei termini (3/3)

✓ Nodo **Bag of Words Creator**

Trasforma il corpus in un vocabolario di termini (**BoW**)



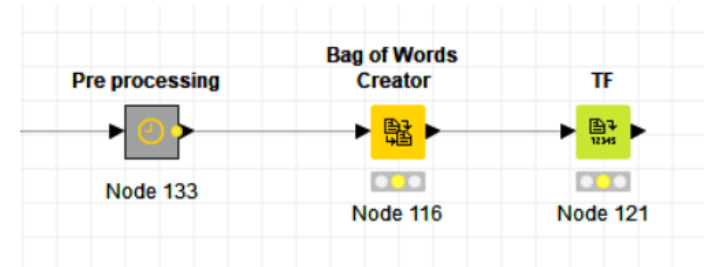
Row ID	S Text
Row0	A Giovanni piacciono i film. Anche a Maria piacciono i film.
Row1	A Giovanni piacciono anche le partite.

Row ID	S Text	T Term
Row0	A Giovanni piacciono i film. Anche a Maria piacciono i film.	a[]
Row1	A Giovanni piacciono i film. Anche a Maria piacciono i film.	giovanni[]
Row2	A Giovanni piacciono i film. Anche a Maria piacciono i film.	piacciono[]
Row3	A Giovanni piacciono i film. Anche a Maria piacciono i film.	i[]
Row4	A Giovanni piacciono i film. Anche a Maria piacciono i film.	film[]
Row5	A Giovanni piacciono i film. Anche a Maria piacciono i film.	anche[]
Row6	A Giovanni piacciono i film. Anche a Maria piacciono i film.	maria[]
Row7	A Giovanni piacciono anche le partite.	a[]
Row8	A Giovanni piacciono anche le partite.	giovanni[]
Row9	A Giovanni piacciono anche le partite.	piacciono[]
Row10	A Giovanni piacciono anche le partite.	anche[]
Row11	A Giovanni piacciono anche le partite.	le[]
Row12	A Giovanni piacciono anche le partite.	partite[]



■ Costruzione della matrice termini-documenti

- Costruzione del vettore dei termini
 - ✓ Nodo **TF (Term Frequency, senza l'opzione *Relative*)**
Calcola la frequenza assoluta di ogni termine in ciascun documento.



Row ID	S Text
Row0	A Giovanni piacciono i film. Anche a Maria piacciono i film.
Row1	A Giovanni piacciono anche le partite.

Row ID	S Text	T Term	TF abs
Row0	A Giovanni piacciono i film. Anche a Maria piacciono i film.	a[]	2
Row1	A Giovanni piacciono i film. Anche a Maria piacciono i film.	giovanni[]	1
Row2	A Giovanni piacciono i film. Anche a Maria piacciono i film.	piacciono[]	2
Row3	A Giovanni piacciono i film. Anche a Maria piacciono i film.	i[]	2
Row4	A Giovanni piacciono i film. Anche a Maria piacciono i film.	film[]	2
Row5	A Giovanni piacciono i film. Anche a Maria piacciono i film.	anche[]	1
Row6	A Giovanni piacciono i film. Anche a Maria piacciono i film.	maria[]	1
Row7	A Giovanni piacciono anche le partite.	a[]	1
Row8	A Giovanni piacciono anche le partite.	giovanni[]	1
Row9	A Giovanni piacciono anche le partite.	piacciono[]	1
Row10	A Giovanni piacciono anche le partite.	anche[]	1
Row11	A Giovanni piacciono anche le partite.	le[]	1
Row12	A Giovanni piacciono anche le partite.	partite[]	1



■ Costruzione della matrice termini-documenti

- N-grammi (1/2)

Sono una sotto sequenza dei termini consecutivi che si trovano in un documento che consentono di catturare parti di frasi (nomi, città, titoli, ...) o espressioni comuni ("basta e avanza", "copia e incolla", ...).

Ad esempio, nella seguente frase:

Dichiarazione universale dei diritti umani

ci sono

5 "unigrammi"

(Dichiarazione; universale; dei; diritti; umani)

4 "digrammi"

(Dichiarazione universale; universale dei; dei diritti; diritti umani)

3 "trigrammi"

(Dichiarazione universale dei; universale dei diritti; dei diritti umani)

2 "4-grammi"

(Dichiarazione universale dei diritti; universale dei diritti umani)

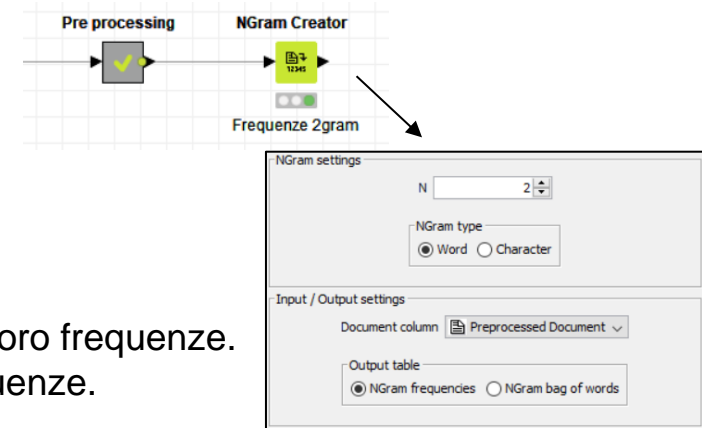


■ Costruzione della matrice termini-documenti

• N-grammi (2/2)

✓ Nodo **NGram Creator**

Crea una lista degli n -grammi del documento con le loro frequenze. L'uscita può essere una tabella di tipo BoW o di frequenze.



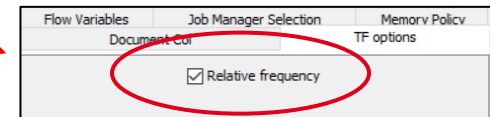
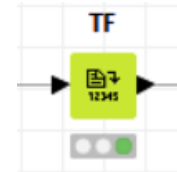
Row ID	S Text
Row0	A Giovanni piacciono i film. Anche a Maria piacciono i film.
Row1	A Giovanni piacciono anche le partite.

Row ID	S Ngram	Corpus frequency	Document frequency	Sentence frequency
0	a giovanni	2	2	2
1	giovanni piacciono	2	2	2
2	piacciono i	2	1	2
3	i film	2	1	2
4	anche a	1	1	1
5	a maria	1	1	1
6	maria piacciono	1	1	1
7	piacciono anche	1	1	1
8	anche le	1	1	1
9	le partite	1	1	1



■ Costruzione della matrice termini-documenti

- Pesatura dei termini (TF Relative)



- ✓ Nodo **TF (Term Frequency)**, con l'opzione **Relative**

Calcola la frequenza relativa di un termine t all'interno di un documento d per tutte le frequenze dei k termini in esso presenti.

$$TF_{t,d} = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

Ad esempio¹, considerando solo questi termini all'interno di 2 documenti (tra parentesi è indicato il numero di volte che il termine compare nel documento):

Doc1 *This (1) | is (1) | a (2) | sample (1)*
Doc2: *This (1) | is (1) | another (2) | example (3)*

TF "This",Doc1 = 1/5 = 0,2
TF "This",Doc2 = 1/7 ≈ 0,143
TF "example",Doc2 = 3/7 ≈ 0,429

T Term	D TF rel
this[]	0.2
is[]	0.2
a[]	0.4
sample[]	0.2
this[]	0.143
is[]	0.143
another[]	0.286
example[]	0.429

¹ Preso da Wikipedia <https://en.wikipedia.org/wiki/Tf-idf>



■ Costruzione della matrice termini-documenti

- Pesatura dei termini (IDF)

✓ Nodo **IDF (Inverse Document Frequency)**

È una misura di quanta informazione il termine fornisce, cioè se il termine t è comune o raro tra tutti i D documenti del corpus.

$$IDF_{t,D} = \log \frac{D}{n_t} \quad \text{dove } n_t \text{ è il numero di documenti dove appare il termine } t.$$

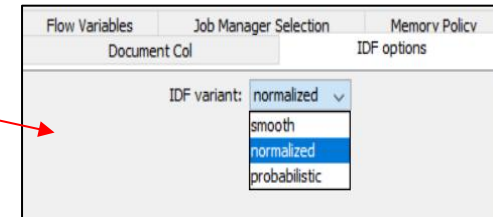
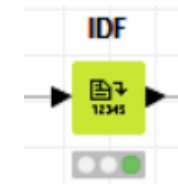
Dall'esempio precedente:

Doc1 *This (1) | is (1) | a (2) | sample (1)*

Doc2: *This (1) | is (1) | another (2) | example (3)*

$$IDF \text{ "This", } D = \log (2/2) = 0$$

$$IDF \text{ "example", } D = \log (2/1) \approx 0,301$$



T Term	D TF rel	D IDF
this[]	0.2	0
is[]	0.2	0
a[]	0.4	0.301
sample[]	0.2	0.301
this[]	0.143	0
is[]	0.143	0
another[]	0.286	0.301
example[]	0.429	0.301



■ Costruzione della matrice termini-documenti

- Pesatura dei termini (TF-IDF) – 1/2

- ✓ Nodo **Math Formula** (o **Java Snippet**)

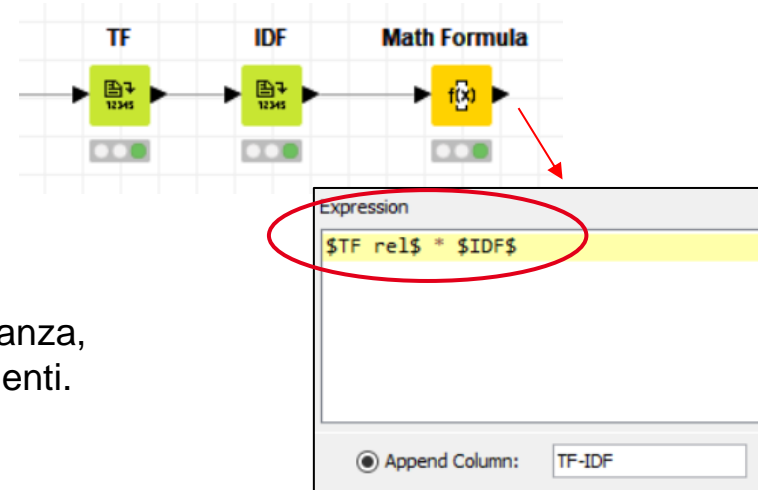
È una statistica che valuta il peso¹, ovvero l'importanza, di un termine all'interno di una collezione di documenti.

$$TF-IDF_{t,d,D} = TF_{i,d} * IDF_{t,D}$$

Un alto valore del peso TF-IDF si ottiene da un alto valore di frequenza del termine t nel documento d e un basso valore di frequenza del termine in tutti i documenti del corpus D .

Segue che:

- Più un termine è raro, più assume "importanza" nel corpus
- Se un termine appare in più documenti, il suo valore si avvicina a zero.



¹ Viene spesso usata in quanto cresce proporzionalmente con il numero di volte che il termine appare nel documento ma viene compensato dal numero di volte che il termine appare nel corpus, evitando così che alcune parole vengano pesate di più solo per il fatto che appaiono più frequentemente di altre.



■ Costruzione della matrice termini-documenti

- Pesatura dei termini (TF-IDF) - 2/2

- ✓ **Nodo Math Formula** (o **Java Snippet**)

Dall'esempio precedente:

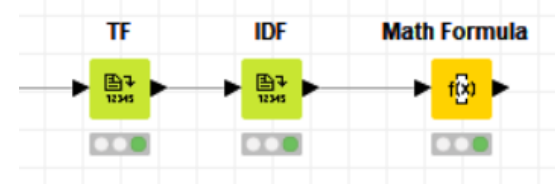
Doc1 *This (1) | is (1) | a (2) | sample (1)*

Doc2: *This (1) | is (1) | another (2) | example (3)*

$$TF-IDF \text{ "example",doc1} = TF \text{ "example",doc1} * IDF \text{ "example",D} = 0/5 * \log(2/1) = 0$$

$$TF-IDF \text{ "example",doc2} = TF \text{ "example",doc2} * IDF \text{ "example",D} = 3/7 * \log(2/1) \approx 0,129$$

T Term	D TF rel	D IDF	D TF-IDF
this[]	0.2	0	0
is[]	0.2	0	0
a[]	0.4	0.301	0.12
sample[]	0.2	0.301	0.06
this[]	0.143	0	0
is[]	0.143	0	0
another[]	0.286	0.301	0.086
example[]	0.429	0.301	0.129





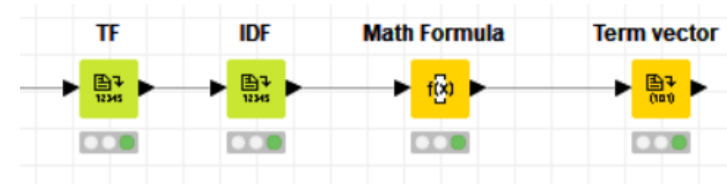
■ Costruzione della matrice termini-documenti

- Matrice Termini/Documenti (TDM)

✓ Nodo **Term Vector**

Ogni riga è un termine, le colonne sono i documenti dove compare.

I valori delle celle possono essere frequenze (assolute o relative) o pesi (TF-IDF)



	Documento 1	Documento 2	...	Documento m
Termine 1	1	1		0
Termine 2	0	2		1
Termine 3	3	0		0
...				
Termine n	0	1		0

Row ID	T Term	D This is a sampl...	D This is another example...
Row7	a[]	0.12	0
Row8	another[]	0	0.086
Row9	example[]	0	0.129
Row10	is[]	0	0
Row11	sample[]	0.06	0
Row12	this[]	0	0



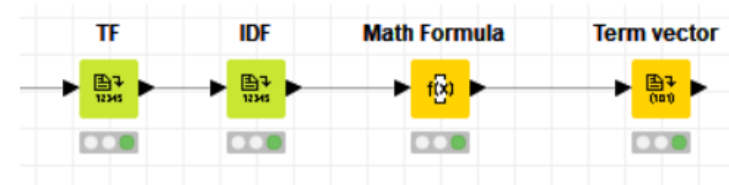
■ Costruzione della matrice termini-documenti

- Matrice Documenti/Termini (DTM)

✓ Nodo **Document Vector**

Ogni riga è un documento, le colonne sono i termini.

I valori delle celle possono essere frequenze (assolute o relative), IDF, o TF-IDF



	Termine 1	Termine 2	...	Termine m
Documento 1	1	1		0
Documento 2	0	2		1
Documento 3	3	0		0
...				
Documento n	0	1		0

Row ID	Document	D this	D is	D a	D sample	D another	D example
Row4	This is a sampl...	0	0	0.12	0.06	0	0
Row5	This is anothe...	0	0	0	0	0.086	0.129



■ Utilizzo del software KNIME – Modulo3_Esercizio1 (parte 2)

- creare la lista di termini con il nodo **Bag of Words**;
- calcolare la frequenza dei termini assoluta e relativa con due nodi **TF**;
- calcolare la metrica IDF con il nodo **IDF**;
- calcolare TF-IDF con il nodo **Math Formula**;
- assegnare con il nodo **Color Manager**, per la colonna *TF rel* il colore nero al valore massimo e al valore minimo il colore grigio chiaro;



- Concetti base
- Raccolta dati
- Preparazione
- Trasformazione
- **Analisi**
 - Visualizzazione; Tag Clouds
 - Raggruppamento; Cluster Analysis
 - Estrazione argomenti
 - Classificazione



■ Metodi applicabili per il Text Mining¹

Metodi	Area
Tag Clouds	Visualizzazione di parole-chiave
Latent Dirichlet allocation (LDA)	Estrazione argomenti
k-Means Clustering	Raggruppamento documenti
Hierarchical Clustering	Raggruppamento documenti
Principal Component Analysis (PCA)	Riduzione dimensionalità
Singular Value Decomposition (SVD)	Riduzione dimensionalità
Naïve Bayes	Classificazione documenti
Logistic regression	Classificazione documenti
Decision trees	Classificazione documenti
Neural network	Classificazione documenti
Support vector machines (SVM)	Classificazione documenti

¹ Per approfondimenti sul raggruppamento e la classificazione si veda "**Business Analytics, Modelli Statistici per il Decision Making – Modelli Esplorativi**" e "**Business Analytics, Modelli Statistici per il Decision Making – Modelli Predittivi**" dell'autore delle presenti note.



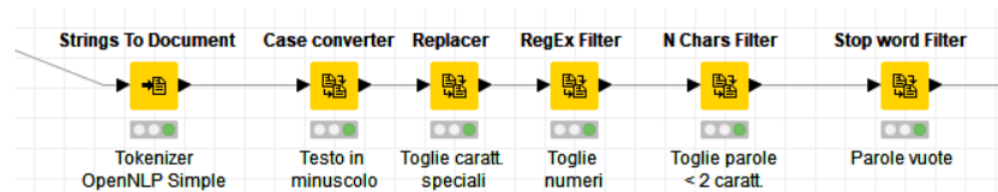
Utilizzo del software KNIME – Modulo3_Esempio1 (Visualizzazione)

I documenti di questo corpus riguardano 9 commenti fatti da alcuni pazienti di un ospedale in merito a una indagine sulla **Customer Satisfaction**. Le analisi che seguono servono per valutare la **qualità dei servizi e delle prestazioni** offerte.

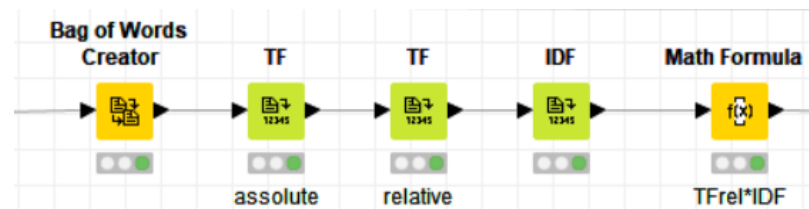
- Corpus

Row ID	S Text
Row0	Cordialità del medico e del personale.
Row1	Il servizio presso la clinica oculistica è stato veloce.
Row2	Il medico e le altre persone sono stati molto, molto cordiali.
Row3	Il tempo di attesa è stato eccellente e il personale è stato molto disponibile.
Row4	Il modo in cui è stata fatta la terapia.
Row5	Nessun problema nel prenotare una visita.
Row6	Velocità nel servizio.
Row7	Il modo in cui sono stato trattato e i miei referti.
Row8	Nessun tempo di attesa, i referti sono stati consegnati velocemente, ottima terapia.

- Preparazione

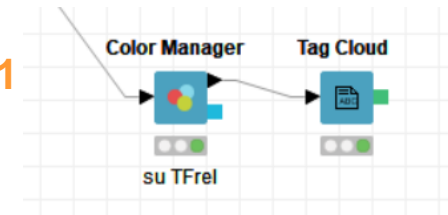


- Trasformazione





■ Utilizzo del software KNIME – Modulo3_Esempio1



- Nuvole di parole-chiave (tag clouds)

✓ Nodo **Tag Cloud**

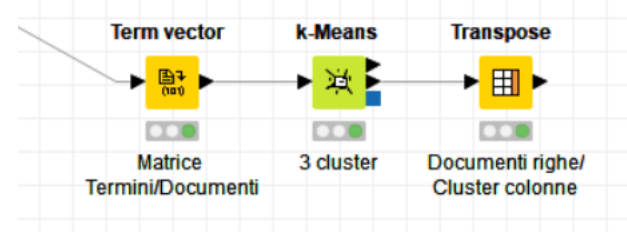
Permette di rappresentare graficamente delle nuvole di parole-chiave” (o "tag") dove la dimensione e la densità del colore dei caratteri è proporzionale alla frequenza (o al peso) con cui il termine appare all'interno di un documento per percepire velocemente le parole più rilevanti.



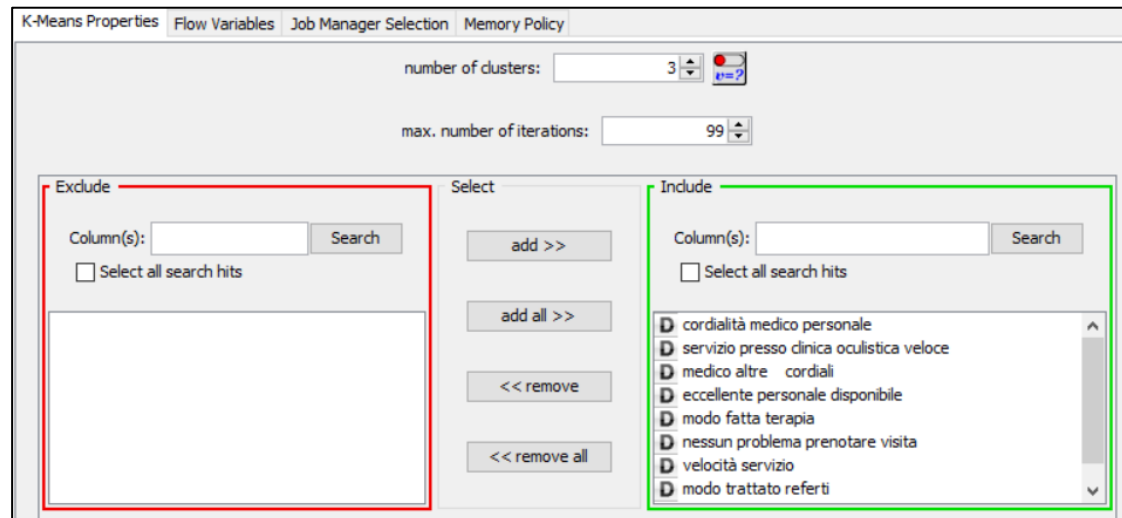


■ Utilizzo del software KNIME – Modulo3_Esempio2 (Raggruppamento)

- Clustering
- ✓ Nodo **K-Means**



Questo metodo di clustering non gerarchico prevede la decisione a priori del numero di cluster che si vogliono ottenere e una procedura iterativa che assegna ogni soggetto a un gruppo. Sono stati scelti a priori 3 cluster.





■ Utilizzo del software KNIME – Modulo3_Esempio2 (Raggruppamento)

- Clustering

- ✓ Nodo **K-Means**

Dalla trasposizione si notano le caratteristiche dei 3 cluster:

Row ID	D ▲ cluster_0	D ▲ cluster_1	D cluster_2
nessun referti consegnati velocemente ottima terapia	0	0	0.05
eccellente personale disponibile	0	0	0.053
modo fatta terapia	0	0	0.047
modo trattato referti	0	0	0.047
nessun problema prenotare visita	0	0	0.055
cordialità medico personale	0.073	0	0.033
medico altre cordiali	0.285	0	0
velocità servizio	0	0.134	0
servizio presso clinica oculistica veloce	0	0.149	0

Cluster 0 "*Gentilezza*": cordialità medici e personale

Cluster 1 "*Efficienza*": servizio veloce

Cluster 2 "*Efficacia*": disponibilità e trattamento



■ Utilizzo del software KNIME – Modulo3_Esempio3 (Estrazione argomenti)

- Estrazione automatica degli argomenti attraverso l'Allocazione Latente di Dirichlet (LDA)

✓ Nodo **Topic Extraction (Parallel LDA)**

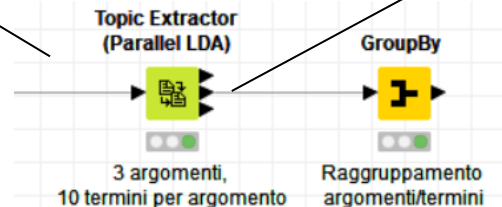
La LDA è un modello statistico che trova un gruppo di termini rilevanti (Keyword) che ricorrono insieme con un'alta probabilità di appartenervi, portando così all'identificazione di un "argomento" (Topic).

Document column: Seed:

No of topics: No of words per topic:

Alpha: Beta:

No of iterations: No of threads:



Row ID	S Topic id	S Term	D Weight
Row0	topic_0	servizio	4
Row1	topic_0	ottima	2
Row2	topic_0	velocemente	2
Row3	topic_0	velocità	2
Row4	topic_0	attesa	2
Row5	topic_0	tempo	2
Row6	topic_0	veloce	2
Row7	topic_0	oculistica	2
Row8	topic_0	clinica	2
Row9	topic_1	personale	4
Row10	topic_1	medico	4
Row11	topic_1	disponibile	2
Row12	topic_1	eccellente	2
Row13	topic_1	attesa	2
Row14	topic_1	tempo	2
Row15	topic_1	cordiali	2
Row16	topic_1	persone	2
Row17	topic_1	cordialità	2
Row18	topic_2	referti	4
Row19	topic_2	nessun	4
Row20	topic_2	terapia	4
Row21	topic_2	modo	4
Row22	topic_2	consegnati	2
Row23	topic_2	trattato	2
Row24	topic_2	visita	2
Row25	topic_2	prenotare	2
Row26	topic_2	problema	2
Row27	topic_2	fatta	2



■ Utilizzo del software KNIME – Modulo3_Esempio3 (Estrazione argomenti)

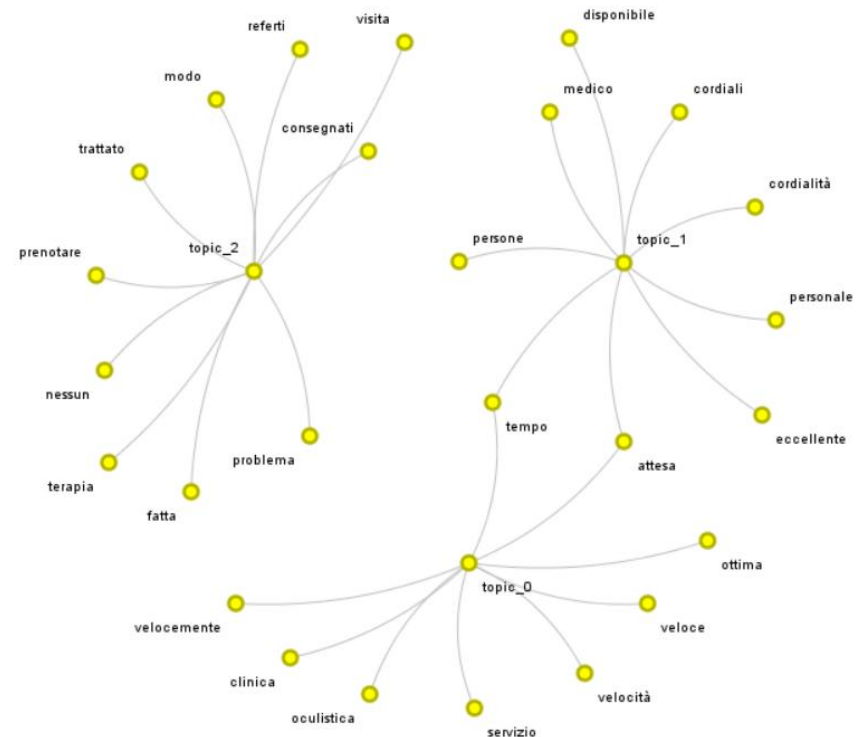
- Estrazione automatica degli argomenti attraverso l'Allocazione Latente di Dirichlet (LDA)

✓ Nodo **Topic Extraction (Parallel LDA)**

Il primo argomento ha come keyword "**servizio**" (peso 4) e "velocità/veloce/velocemente" (tutti peso 2) insieme a "tempo/attesa" (tutti peso 2) e "clinica/oculistica" (tutti peso 2).

Il secondo "**personale/medico**" (tutti peso 4) con "disponibile/eccellente" e "cordiali/cordialità" (tutti peso 2).

Il terzo "**referti/consegnati**" (pesi 4,2) con "**nessun/problema/prenotare**" (pesi 4,2,2) e "**terapia/modo/fatta/visita/trattato**" (pesi 4,4,2,2,2).





■ Utilizzo del software KNIME – Modulo3_Esempio5 (Classificazione)

Una recente tendenza nell'analisi dei testi va oltre il rilevamento di argomenti e cerca di identificare l'emozione dietro un testo. Questa tendenza viene chiamata **Sentiment Analysis**. Estrarre il sentimento da un testo può essere fatto usando tecniche come il NLP, la linguistica computazionale e il text mining.

Per far questo è necessario disporre prima di una collezione di documenti dove ognuno di essi esprime un'opinione (positiva, negativa, neutra, ...) attraverso i Tag di tipo Sentiment (POSITIVE, NEGATIVE, NEUTRAL, IRONY, ...) e poi utilizzare gli algoritmi di Machine Learning per il riconoscimento del sentimento in ciascun testo.

Per esempio, la frase: "Amo viaggiare" è un'affermazione molto positiva, mentre "Odio il tofu" non lo è. Qui i verbi "amare" e "detestare" identificano chiaramente la polarità della frase. Più difficile identificare il sentimento nelle frasi "Non mi dispiace viaggiare" e "Non gradisco il tofu" le affermazioni sono fatte adoperando la negazione di una espressione di senso contrario o nelle frasi più sottili tipo "Non penso che questa dolce sia veramente buono".



■ Utilizzo del software KNIME – Modulo3_Esempio5 (Classificazione)

Un semplice procedimento per l'identificazione del sentimento predominante in un documento, senza ricorrere all'analisi semantica per la disambiguazione, una volta assegnati i tag di tipo Sentiment, può essere il seguente: soglia = media(Sentiment score)

1. Si calcola un punteggio ("Sentiment score") per ciascun documento:

$$\textit{Sentiment score} = (\# \textit{termini positivi} - \# \textit{termini negativi}) / (\# \textit{termini nel documento})$$

2. Si definisce un valore di soglia: *soglia* = *media(Sentiment score)*

3. Si classificano i documenti come *positivi* se *Sentiment score* > *soglia*; altrimenti *negativi*

Se si vuole essere più cauti si possono definire soglie per positivi e i negativi

$$\textit{soglia positivi} = \textit{media(Sentiment score)} + \textit{dev.standard (sentiment score)}$$

$$\textit{soglia negativi} = \textit{media(Sentiment score)} - \textit{dev.standard (sentiment score)}$$

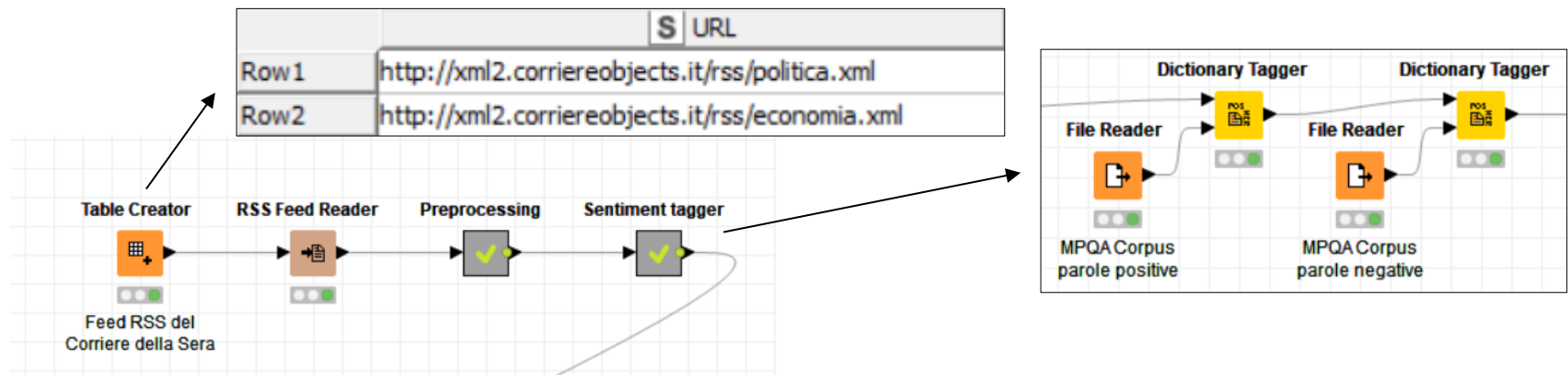
In questo modo tutti i documenti con il sentiment score compreso tra le 2 soglie possono essere classificati come neutri.



■ Utilizzo del software KNIME – Modulo3_Esempio5 (Classificazione)

Il corpus viene creato leggendo i **Feed RSS** (politica ed economia) di un giornale online (*www.corriere.it*). L'intento è quello di classificare i documenti di **Sentiment Positivo, Negativo** secondo il procedimento descritto prima.

- Raccolta dati e preparazione



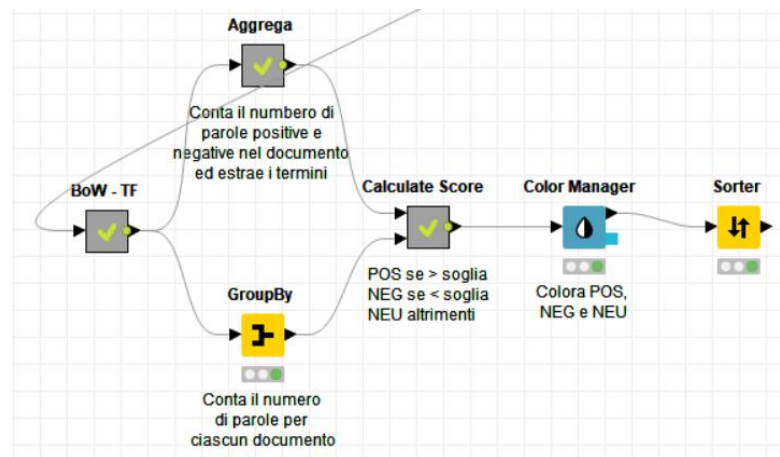


■ Utilizzo del software KNIME – Modulo3_Esempio5 (Classificazione)

Vengono prima calcolati il numero delle parole positive e negative e la loro differenza si divide per il numero totale di parole presenti in ogni singolo documento per ottenere Sentiment score

Poi si calcola la sua media e la deviazione standard per avere il valore di soglia.

Se il Sentiment Score + maggiore della media + la deviazione standard allora il documento viene classificato come positivo





■ Utilizzo del software KNIME – Modulo3_Esempio6 (Classificazione)

I documenti di questo corpus sono le **recensioni di un libro** effettuate online dai lettori. L'analisi che segue **classifica il gradimento o meno del libro in base alle parole contenute nel testo** calcolando una probabilità dell'opinione espressa dalle valutazioni dei recensori con un voto da 1 a 5 portato a un valore dicotomico (*Positivo/Negativo*).

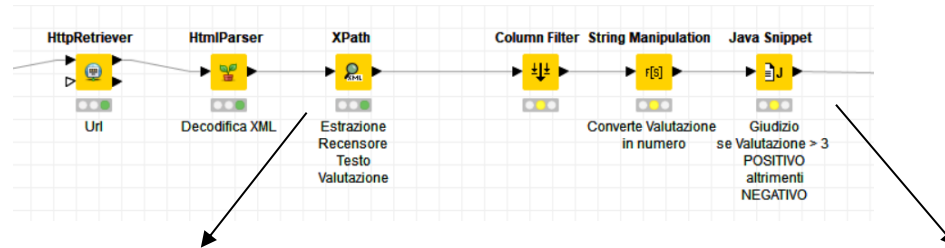
- Corpus

Recensore	Testo	Valutazione	Giudizio
M.VITTORIA	Ancora una volta J.K. Rowling non si smentisce! Libro divorato tutto d'un fiato, come tutti gli altri della saga d'altronde, con una trama avvincente e per nulla scontata nonostante la necessità per l'autrice di doversi confrontare con nuovi personaggi e con un pubblico di lettori sempre più vasto ed esigente. La versione di lettura "a copione" pur rispecchiando le esigenze teatrali rende il testo molto scorrevole e realistico, regalando al lettore la possibilità di immaginarsi sul palco durante lo svolgimento di ogni atto.	4	POSITIVO
martinavatteone	Nulla è più bello che tornare ad Hogwarts, quando lì ci sei cresciuta anche un po' tu! Per chi con Harry, Ron ed Hermione è cresciuto questo libro è imperdibile. Si trovano i vari personaggi, cresciuti, maturati, sposati e con figli che vi faranno rivivere una nuova storia. Il libro è diverso dai precedenti, scritto come un copione ma sempre bello e avvincente. Dove c'è J.K. Rowling c'è magia!	5	POSITIVO
julie.soul	Questo "libro" ha fatto arrabbiare molti e lo capisco benissimo perché ha fatto arrabbiare anche me. Sin dalle prime pagine si capisce benissimo che è una presa in giro, un lavoro fatto all'unico scopo di vendere e ok, ci può stare, ma io a questo punto avrei cambiato totalmente la storia, parlato di altro, modificato tutto, invece si sono limitati a riproporre gli stessi concetti in tutti le salse e soprattutto (la cosa peggiore) snaturando tutti i personaggi di Harry Potter facendogli fare cose assurde che non appartengono al loro vero essere. Abbastanza ridicolo, utile sicuramente a fare soldi e ai nostalgici, ma come libro davvero una delusione.	2	NEGATIVO
ALESSANDRA	Questo non è un romanzo ma la sceneggiatura di un'opera teatrale quindi non ci si deve aspettare un racconto pieno di descrizione di paesaggi ma tutt'altro. Vengono indicati gli attori che entrano in scena e i dialoghi dei singoli interpreti. Questo fa sì che la lettura risulta essere molto veloce e gradevole. L'autrice è capace di far immergere il lettore nel racconto e di catapultarlo in un vero e proprio teatro.	4	POSITIVO



Utilizzo del software KNIME – Modulo3_Esempio6 (Classificazione)

- Raccolta dati



XPath summary

Column name	XPath query	Type
Recensore	//*[@itemprop="author"]	String(Multiple Rows)
Testo	//*[@class="text"]	String(Multiple Rows)
Valutazione	//*[@itemtype="http://schema.org/Rating"]	String(Multiple Rows)

Selected XPath: /*

Buttons: Add XPath, Edit XPath, Remove XPath

XML-Cell Preview

```

1331 <span itemtype="http://schema.org/Person" ·
1332 itemscope="itemscope" itemprop="author"><span ·
1333 itemprop="name">MANOLA</span></span></small><meta ·
1334 content="2016-09-28" itemprop="datePublished"/><small>il 28
1335 settembre 2016</small><div class="text"><p ·
1336 itemprop="description">Non si sa mai! esclamò Ron ·
1337 guardando il libriccino con apprensione. «#034;Fra i libri ·
1338 confiscati dal ministero... mi ha detto papà... ce n'era ·
1339 uno che ti bruciava gli occhi. E quelli che leggevano ·
1340 Sonetti di uno stregone dopo parlavano in versi per tutta ·
1341 la vita. itemtype="http://schema.org/Rating"><span ·
1342 style="display: none;" itemprop="ratingValue">5</span>
    
```

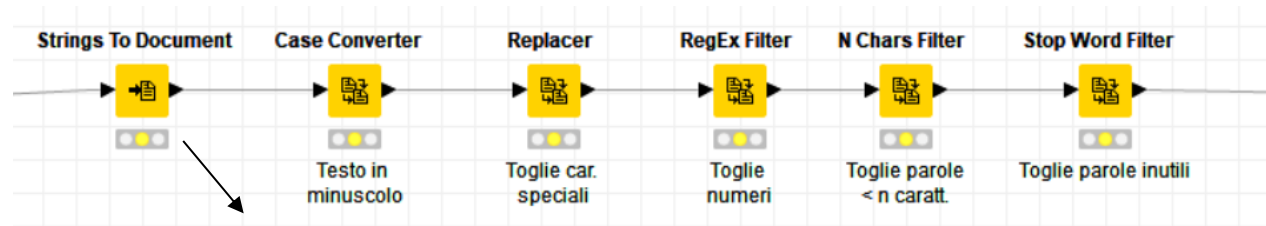
```

27
28 // expression start
29 // Enter your code here:
30
31
32 if (c_Valutazione >= 4)
33 {out_Giudizio = "POSITIVO";}
34 else
35 {out_Giudizio = "NEGATIVO";}
36
37 // expression end
38
39
40
    
```



Utilizzo del software KNIME – Modulo3_Esempio6 (Classificazione)

Preparazione



Title
 Column Row ID Empty string
Title column: \$ Testo

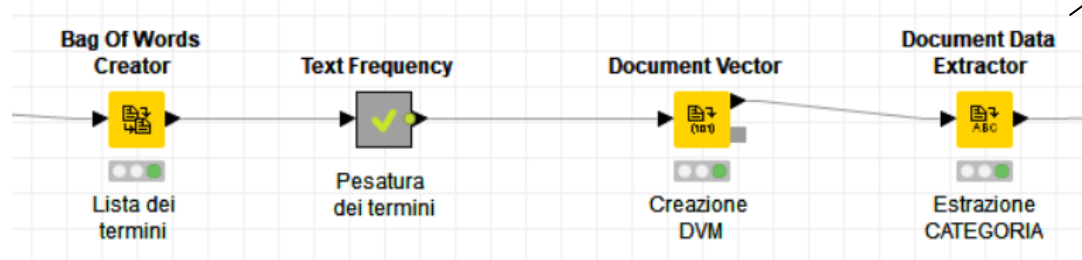
Text
Full text: \$ Testo

Meta Information
Document source:
 Use sources from column Document source column: \$ Giudizio
Document category:
 Use categories from column Document category column: \$ Giudizio

Document column: Document

Data extractors:
Title
Abstract
Text
Document body text
Author
Author set
Category
Category set

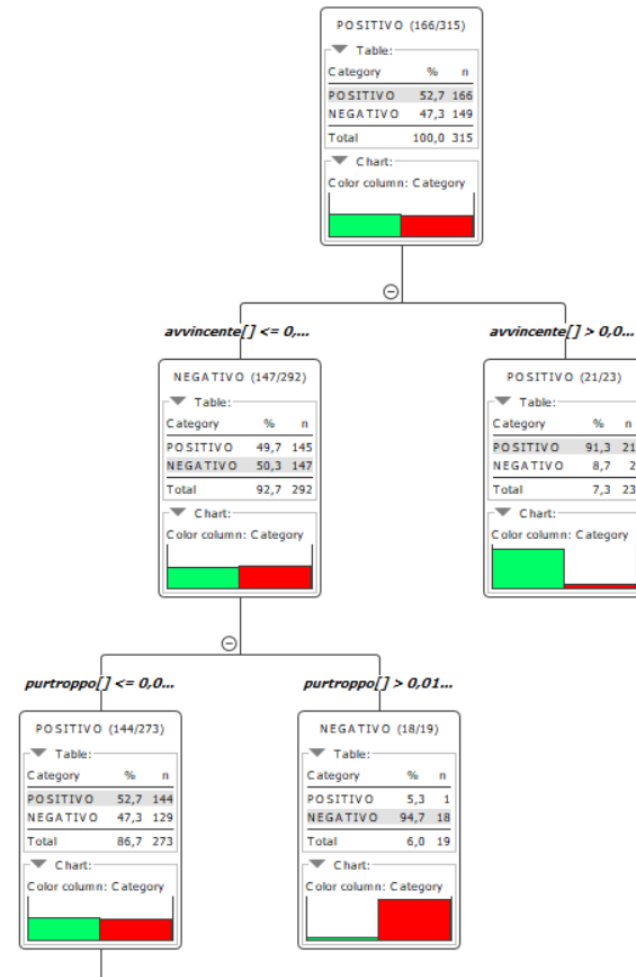
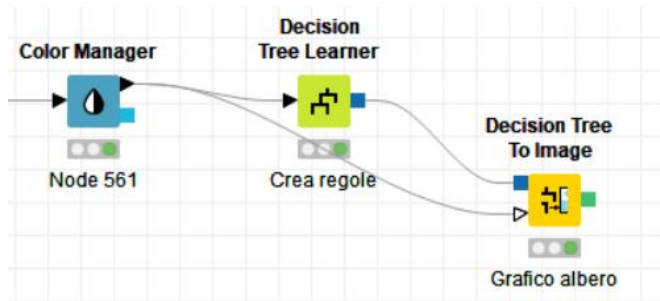
Trasformazione





Utilizzo del software KNIME – Modulo3_Esempio6 (Classificazione)

- Alberi decisionali
 - ✓ **Nodo Tree Learner**
Usa Giudizio come variabile target.



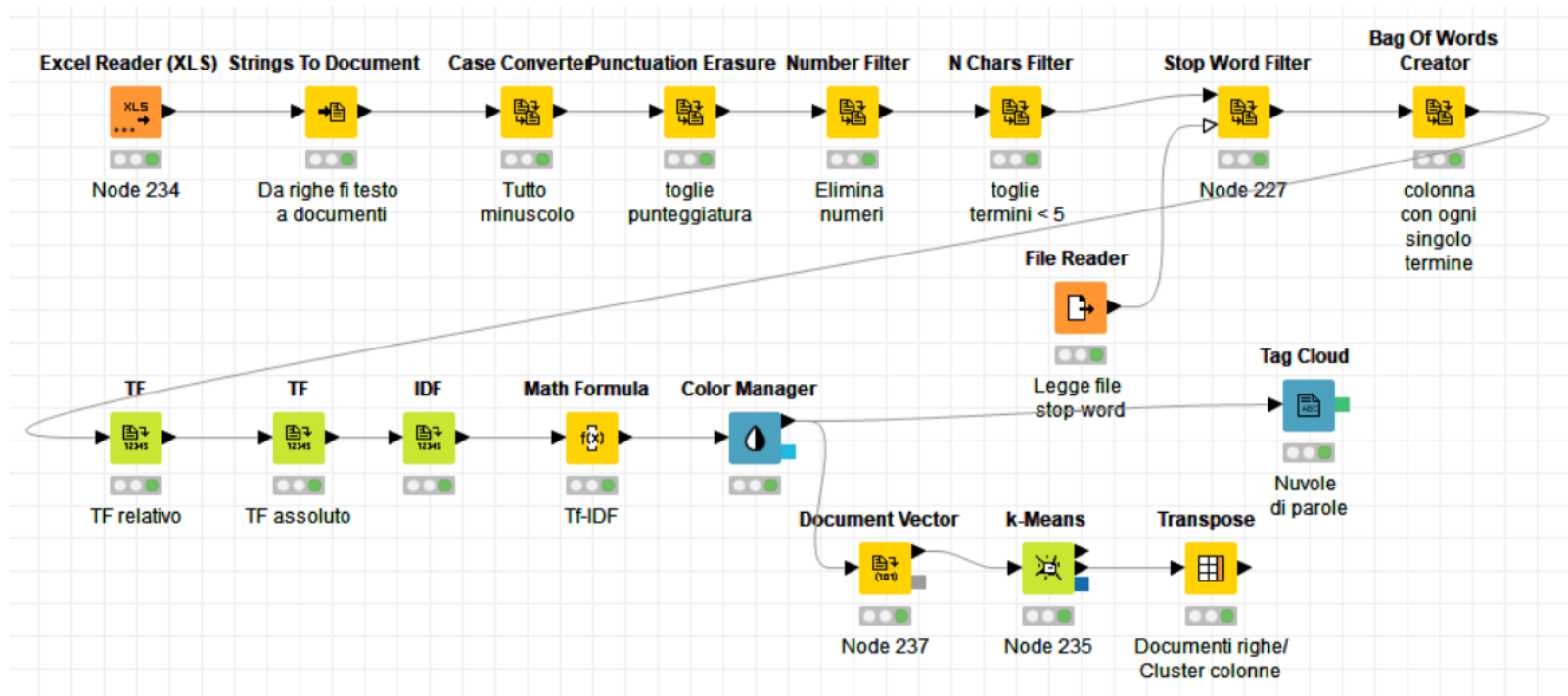


■ Utilizzo del software KNIME – Modulo3_Esercizio1 (parte 3)

- visualizzare le nuvole di parole con il nodo **Tag Clouds**, al massimo 100 righe per la colonna *TF rel*;
- interpretare i risultati;
- collegare all'uscita del nodo **Color Manager** il nodo **Document Vector** con il vector value selezionato con la colonna *TF-IDF*; togliere le selezioni nei campi "Bitvector" e "As collection cell";
- collegare il nodo **k-Means** con numero di cluster uguale a 3; usare un seme casuale (p.e. 12345) e portare le iterazioni a 200;
- dalla seconda porta d'uscita collegare il nodo **Transpose**;
- interpretare i risultati.



■ Soluzione Modulo3_Esercizio1





Installazione dell'integrazione Palladian

<https://nodepit.com/node-installation-guide>

File -> Preferences -> Install/Update -> Available Software Sites

Aggiungere **<https://download.nodepit.com/palladian/5.3>**

Name: *Palladian*

File -> Install KNIME Extensions

Inserire la parola chiave Palladian e installare le estensioni





Software

<https://www.knime.org/>

Bibliografia

Technical Report - The KNIME Text Processing Feature:

https://www.knime.com/sites/default/files/inline-images/knime_text_processing_introduction_technical_report_120515.pdf

Gary Miner, Dursun Delen, John Elder, Andrew Fast, Thomas Hill and Robert A. Nisbet, "*Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*", Academic Press, Waltham, Mass., 2012



L'esperienza è il miglior maestro

Contatti

[alfredo.roccato\(at\)fastwebnet.it](mailto:alfredo.roccato(at)fastwebnet.it)

www.alfredoroccatto.it