Text Mining Course for KNIME Analytics Platform

Michele Marsili - michele.marsili@invalsi.it

Paolo Tamagnini - paolo.tamagnini@knime.com

1st International Conference
on Measurement in STEM Education (MESE1)
University of Naples Federico II
PLS for Statistics and Physics

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Table of Contents

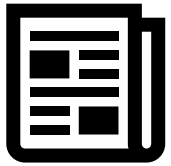
1. Text Mining Concepts
2. KNIME Analytics Platform
3. The Text Processing Extension
4. Importing Text
5. Enrichment
6. Preprocessing
7. Transformation
8. Classification
9. Overview Text Mining Use Cases

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Text Mining Concepts

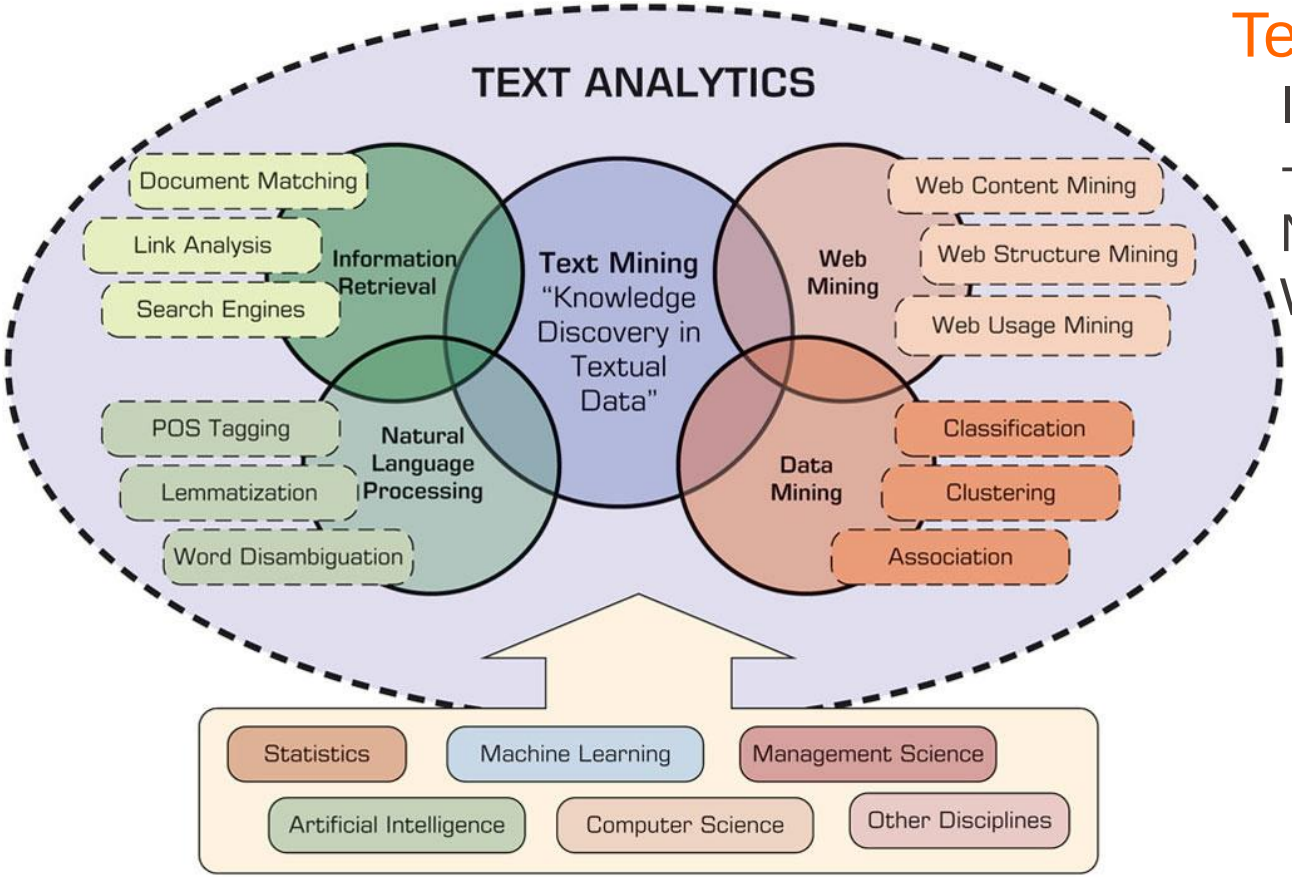
Too Many Terms

- Text Mining
- Text Analytics
- Text Processing
- Information Retrieval
- Information Extraction
- Natural Language Processing
- Computational Linguistics
- Unstructured Data Mining
- …



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Text Mining versus Text Analytics



Text Analytics =

Information Retrieval

+

NLP + Data Mining +

Web Mining

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Why Text Mining?

- Roughly 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months…
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: text mining / text analytics / text processing
  - A semi-automated process of extracting information and discovering knowledge from unstructured data sources
  - It is also described as text data mining or knowledge discovery in textual databases

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Data Mining versus Text Mining

- Both seek for novel and useful patterns
- Both are semi-automated processes
- Difference is the nature of the data
  - Structured versus unstructured data
  - Structured data: in databases
  - Unstructured data: Word documents, PDF files, text excerpts, XML files, and so on
- To perform text mining – first, impose structure to the data, then mine the structured data.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Use of Text Mining

- Benefits of text mining are obvious especially in text-rich data environments
  - Such as law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.
- Electronic communication records (e.g., Email)
  - Spam filtering
  - Email prioritization and categorization
  - Automatic response generation

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Text Mining Terminology

- Unstructured data
- Corpus (and corpora)
- Words and Terms
- Concepts
- Stemming vs. Lemmatization
- Stop [include] words/terms
- Synonyms (and homonyms)
- Morphology
- Tokenization
- Part-of-speech tagging
- Bag of Words
- Lexicon (word dictionary)
  - WordNet, SentiWordNet
- Term-by-document matrix
  - Word/document frequency
  - Numeric, binary, TF/IDF
- Dimensional reduction: Singular value decomposition (SVD)
- Topic Modeling
  - Latent Dirichlet Allocation (LDA)
- Word Embedding
  - Word2Vec

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Stemming vs. Lemmatization

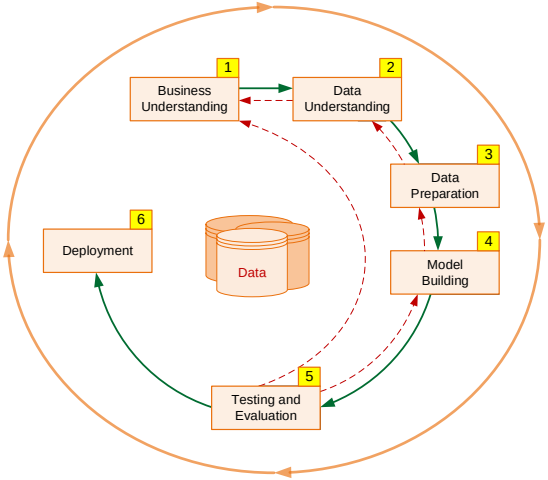
- Common goal: to generate the root form of the words
- Difference:
  - Stem results in truncated/chopped words (not necessarily a complete word)
    - Stemming is syntactic and fast - follows an algorithm
    - Example: amuse, amusing, amused → **amus**
  - Lemma results in an actual language word (inflection free)
    - Lemmatization is semantic and slow - follows linguistic dictionary
    - Example: am, is, are → **be**

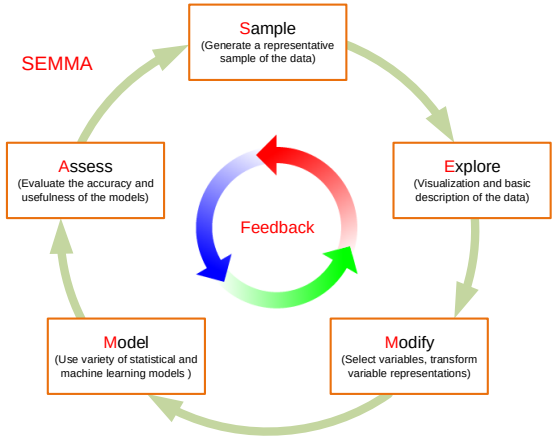
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

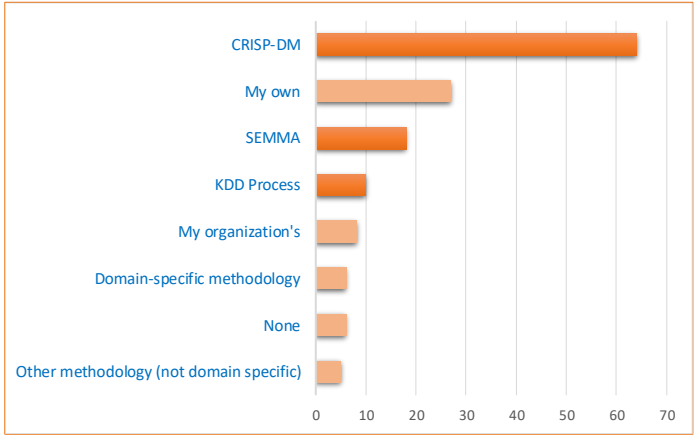
Text Mining Process

Text Mining Process

- A standard process: the manifestation of the “best” practices
- Standard process for data mining:
  - Cross industry process for data mining (CRISP-DM)
  - Sample, Explore, Modify, Model, Assess (SEMMA)

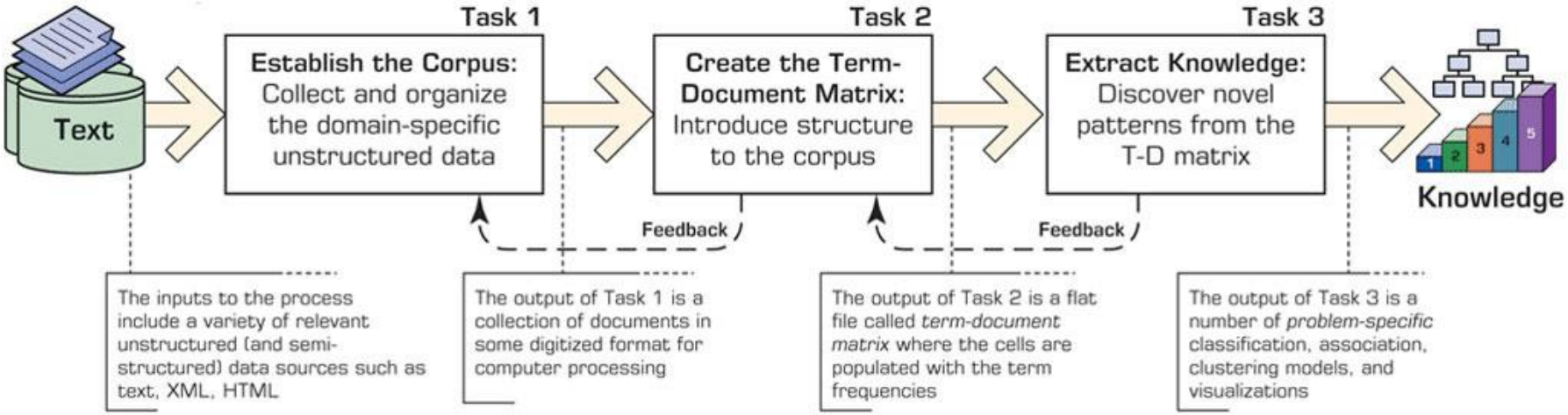






These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Text Mining Process



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

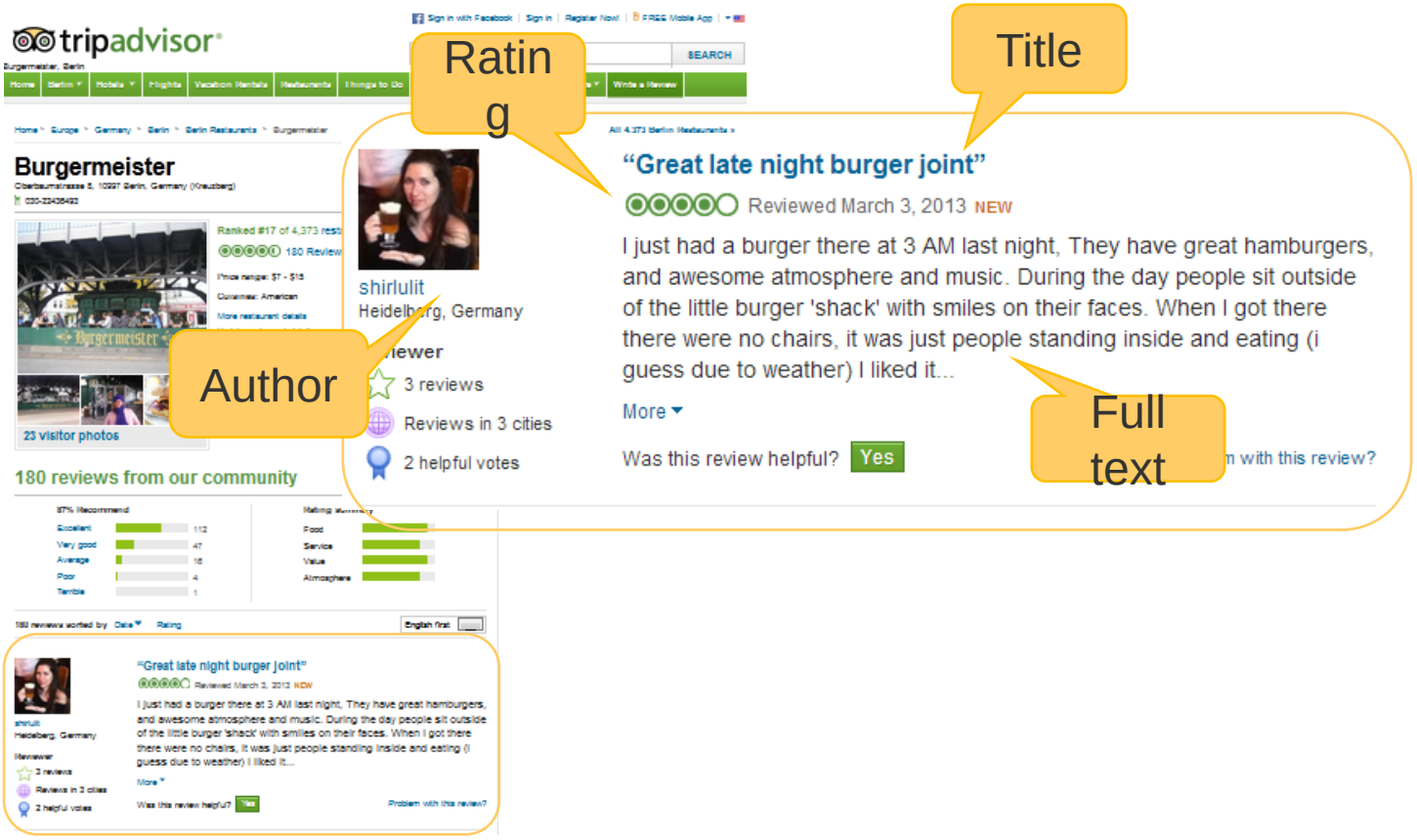
Today’s Example

Today's Example

- Classification of free-text documents is a common task in the field of text mining.
- It is used to categorize documents, i.e. assign pre-defined topics, or it can be used for sentiment analysis.
- Today we want to construct a workflow that reads and preprocesses text documents, transforms them into a numerical representation and builds a predictive model to assign pre-defined labels to documents.
- Additional tasks:
  - Sentiment analysis
  - Visualization of documents
  - Document clustering

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Today's Example



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Today's Example

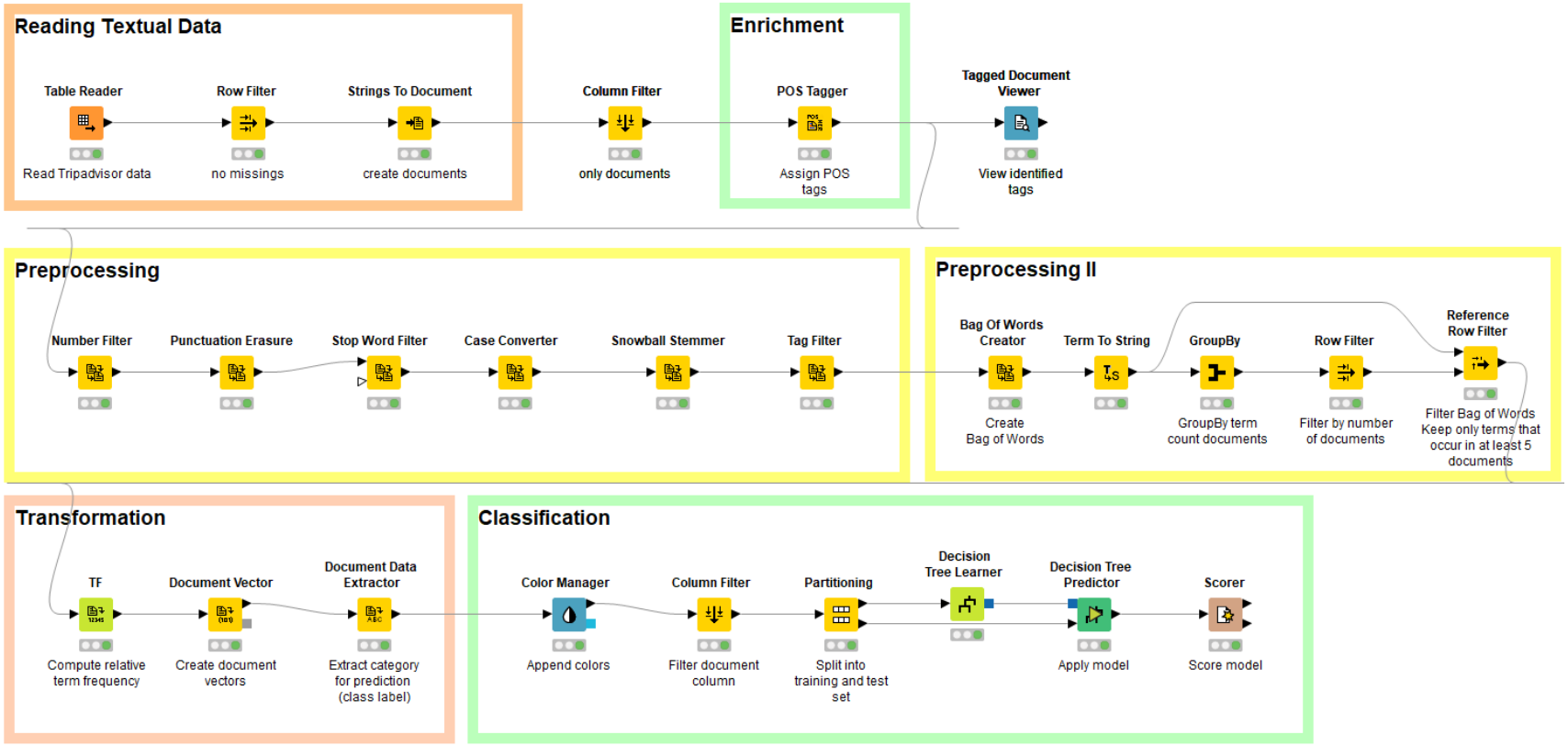
Goal:

- Build a **classifier** to distinguish between reviews about Italian or Chinese restaurants.

Review about an Italian or a Chinese restaurant?

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Today's Example

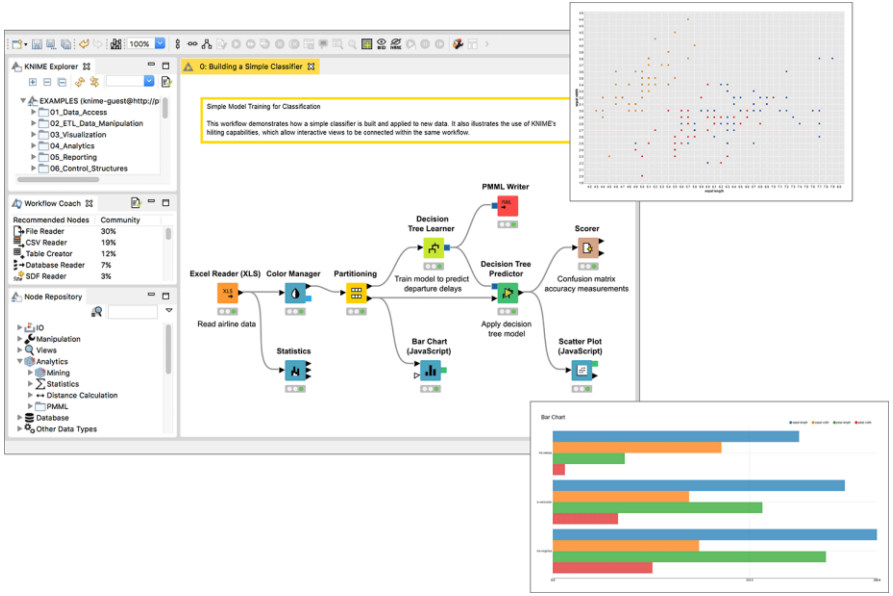


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Overview
KNIME Analytics Platform

What is KNIME Analytics Platform?

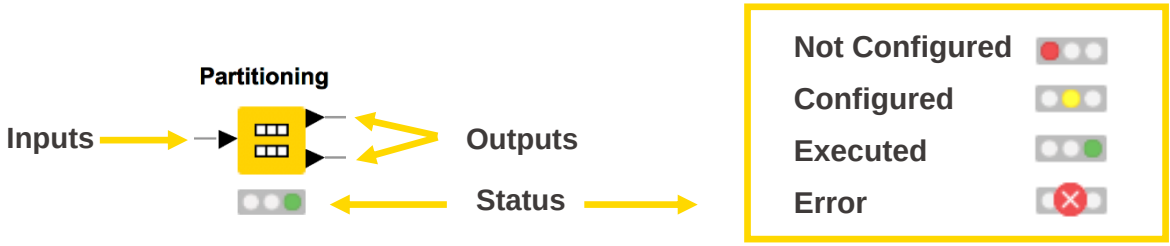
- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
  - Text Mining
  - Network Mining
  - Cheminformatics
  - Many integrations,
    such as Java, R, Python,
    Weka, Keras, Plotly, H2O, etc.



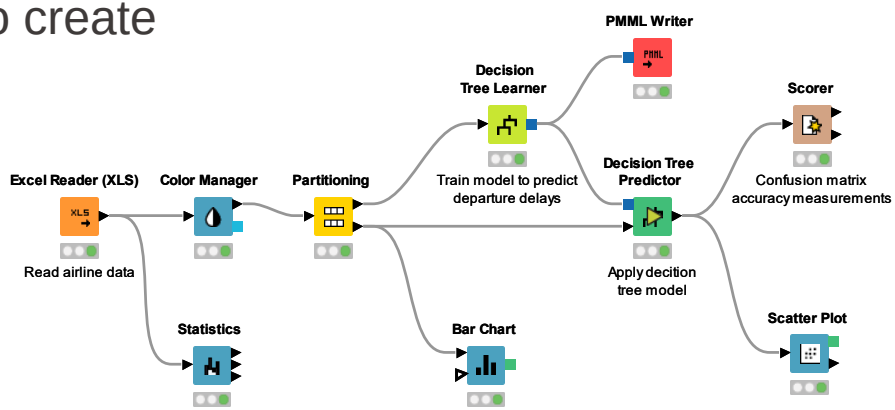
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Visual KNIME Workflows

NODES perform tasks on data

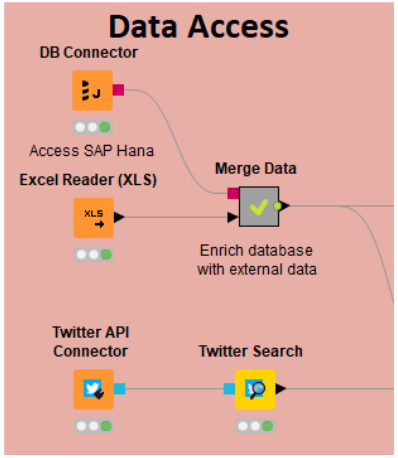


Nodes are combined to create **WORKFLOWS**



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

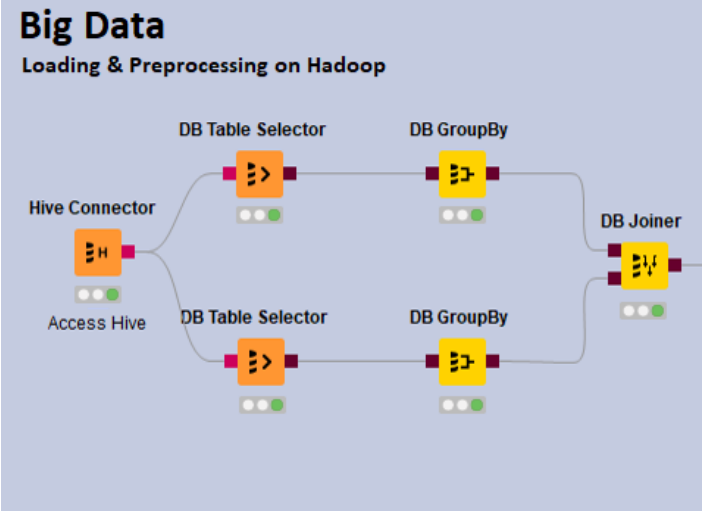
Data Access



- Databases
  - MySQL, PostgreSQL, Oracle
  - Theobald
  - any JDBC (DB2, MS SQL Server)
  - Amazon DynamoDB
- Files
  - CSV, txt, Excel, Word, PDF
  - SAS, SPSS
  - XML, JSON, PMML
  - Images, texts, networks
- Other
  - Twitter, Google
  - Amazon S3, Azure Blob Store
  - Sharepoint, Salesforce
  - Kafka
  - REST, Web services

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

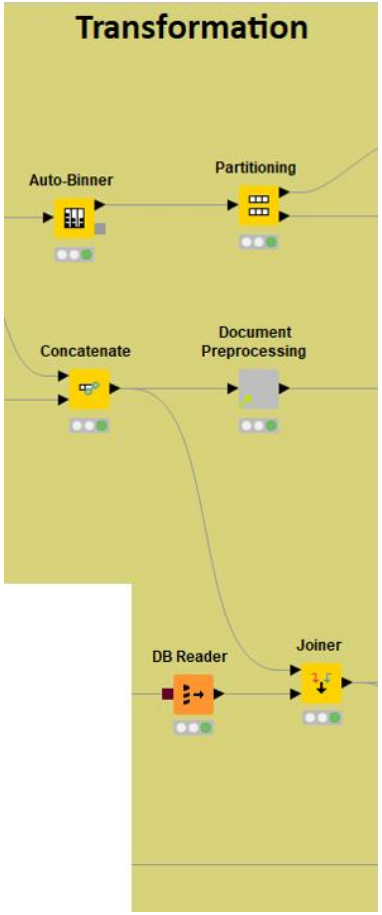
Big Data



- Spark & Databricks
- HDFS support
- Hive
- Impala
- In-database processing

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

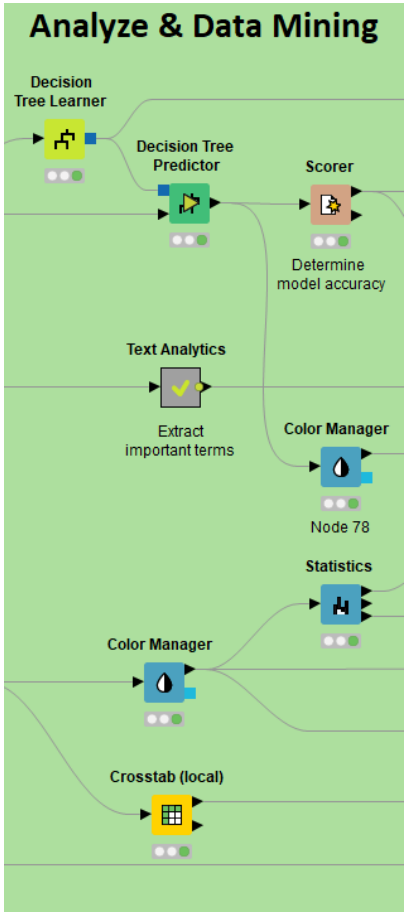
Transformation



- Preprocessing
  - Row, column, matrix based
- Data blending
  - Join, concatenate, append
- Aggregation
  - Grouping, pivoting, binning
- Feature creation and selection

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

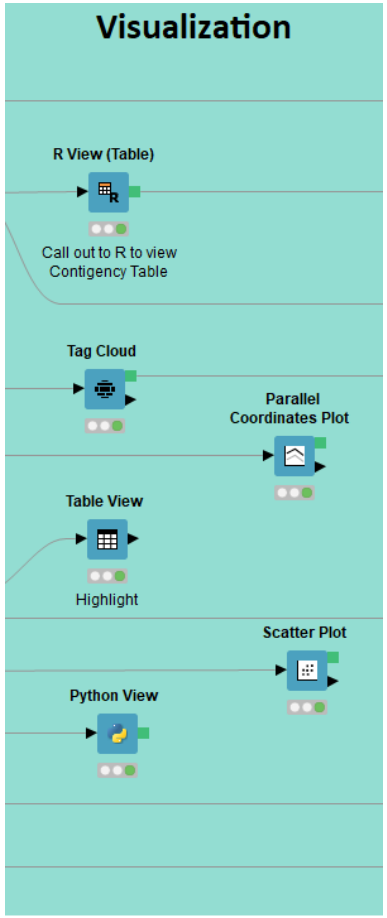
Analysis & Data Mining

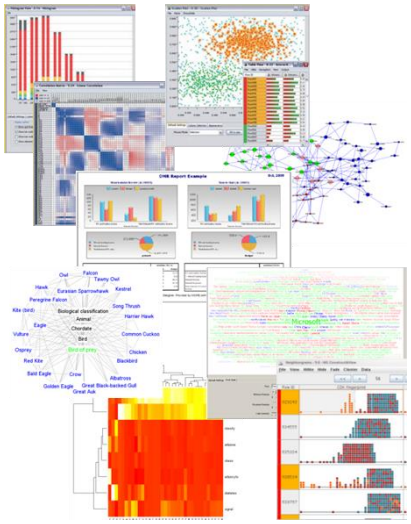


- Regression
  - Linear, regression tree
- Classification
  - Decision tree, ensembles, SVM, MLP, Naïve Bayes, logistic regression
- Clustering
  - k-means, DBSCAN, hierarchical
- Validation
  - Cross-validation, scoring, ROC
- Deep Learning
  - Keras, DL4J
- External
  - R, Python, Weka, H2O, Keras

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Visualization

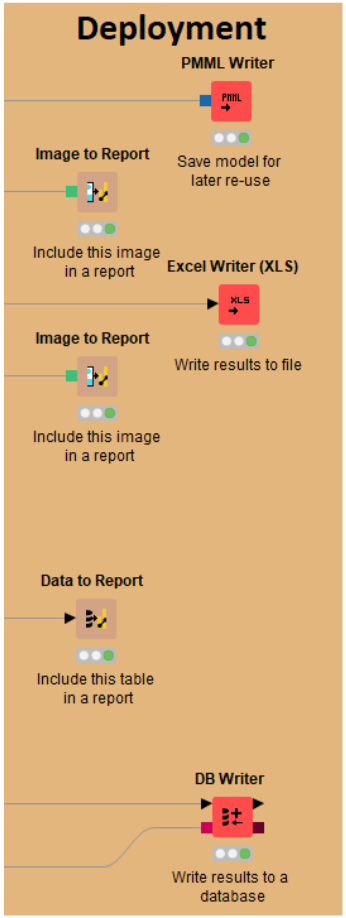




- Interactive visualizations
- JavaScript-based nodes
  - Scatter Plot, Box Plot, Line Plot
  - Networks, ROC Curve, Decision Tree
  - Plotly Integration
  - Adding more with each release!
- Misc
  - Tag cloud, open street map, molecules
- Script-based visualizations
  - R, Python

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

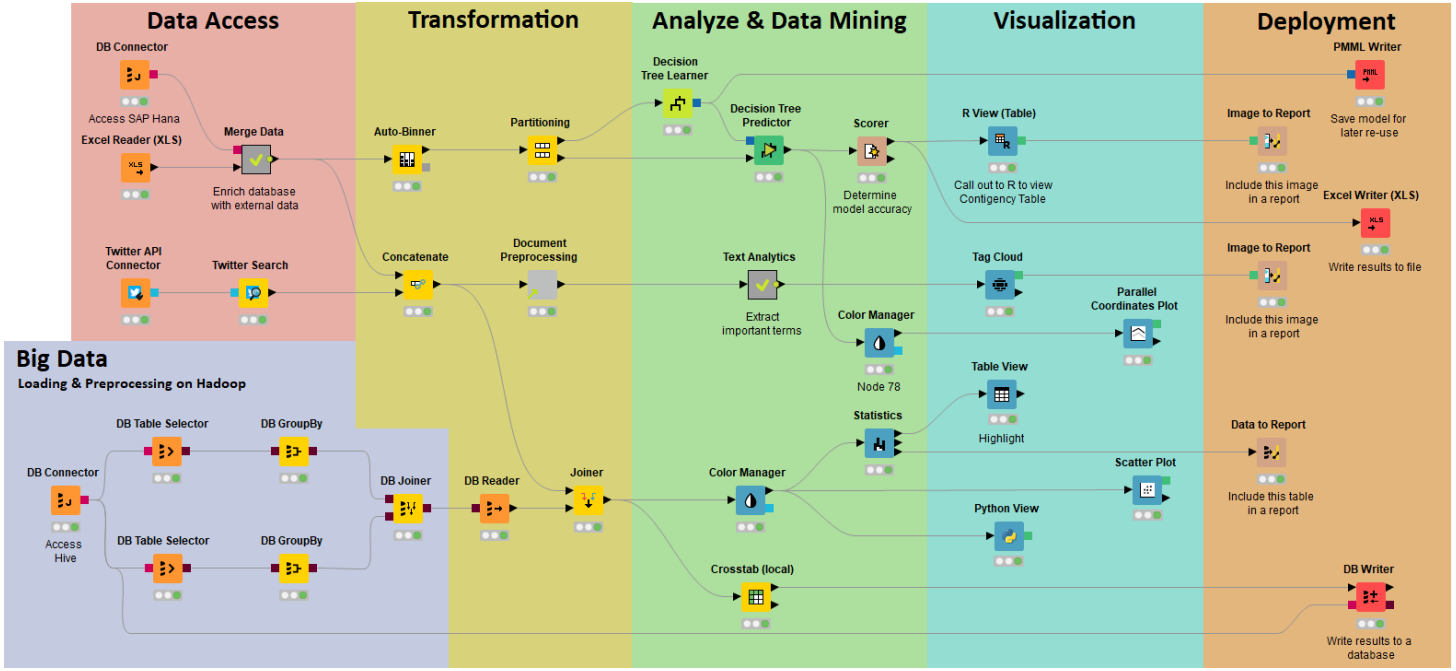
Deployment



- Database
- Files
  - Excel, CSV, txt
  - XML
  - PMML
  - to: local, KNIME Server, Amazon S3, Azure Blob Store
- BIRT reporting

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Over 4000 Native and Embedded Nodes Included:



Data Access
MySQL, Oracle, ...
SAS, SPSS, ...
Excel, Flat, ...
Hive, Impala, ...
XML, JSON, PMML
Text, Doc, Image, ...
Web Crawlers
Industry Specific
Community / 3rd

Transformation
Row
Column
Matrix
Text, Image
Time Series
Java
Python
Community / 3rd

Analysis & Mining
Statistics
Data Mining
Machine Learning
Web Analytics
Text Mining
Network Analysis
Social Media Analysis
R, Weka, Python
Community / 3rd

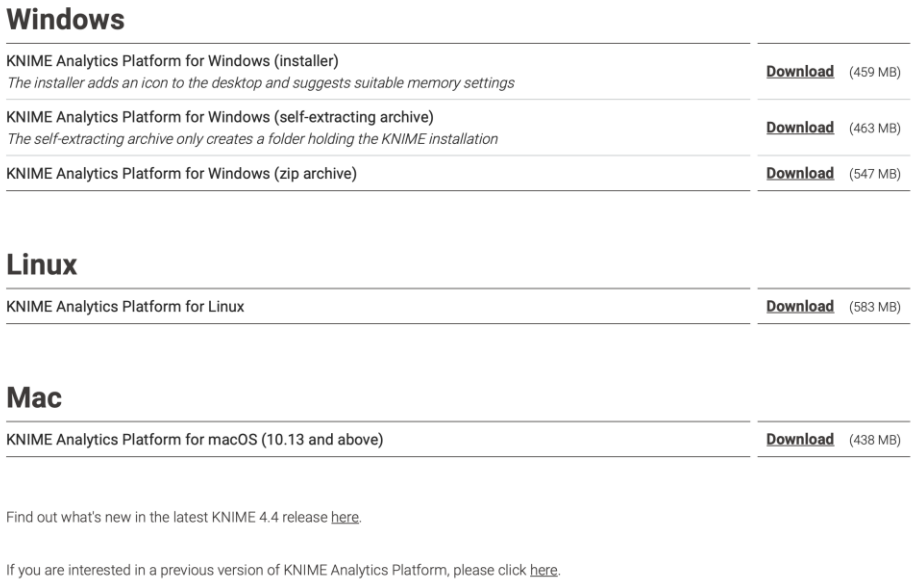
Visualization
R
JFreeChart
JavaScript
Plotly
Community / 3rd

Deployment
via BIRT
PMML
XML, JSON
Databases
Excel, Flat, etc.
Text, Doc, Image
Industry Specific
Community / 3rd

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Install KNIME Analytics Platform

- Select the KNIME version for your computer:
  - Mac
  - Windows – 32 or 64 bit
  - Linux
- Download the archive and extract the file, or download installer package and run it



Windows

KNIME Analytics Platform for Windows (installer) *The installer adds an icon to the desktop and suggests suitable memory settings*	**Download** (459 MB)
KNIME Analytics Platform for Windows (self-extracting archive) *The self-extracting archive only creates a folder holding the KNIME installation*	**Download** (463 MB)
KNIME Analytics Platform for Windows (zip archive)	**Download** (547 MB)

Linux

KNIME Analytics Platform for Linux	**Download** (583 MB)

Mac

KNIME Analytics Platform for macOS (10.13 and above)	**Download** (438 MB)

Find out what's new in the latest KNIME 4.4 release here.

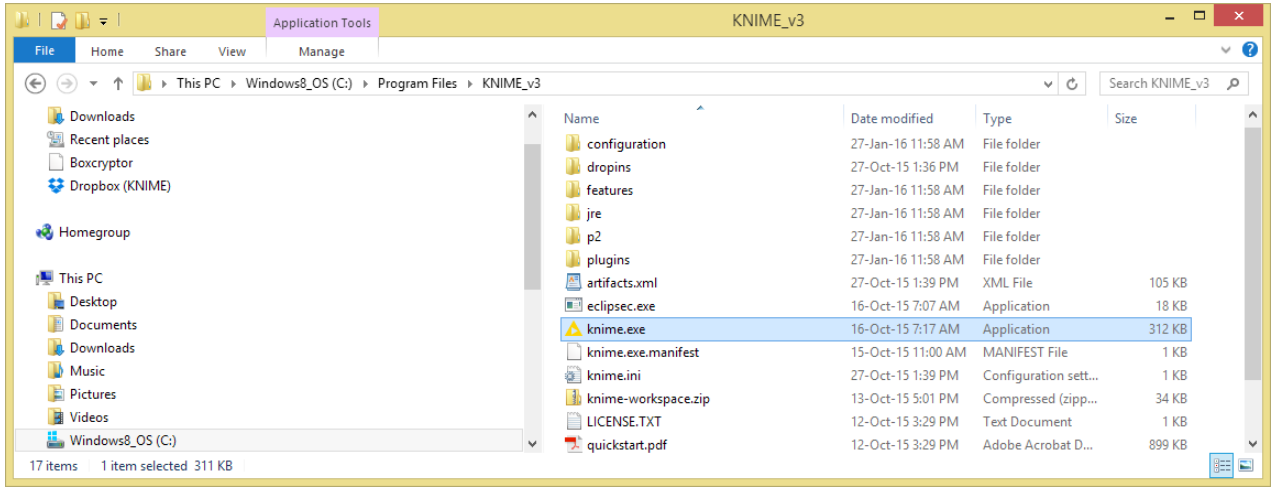
If you are interested in a previous version of KNIME Analytics Platform, please click here.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Start KNIME Analytics Platform

- Use the shortcut created by the installer

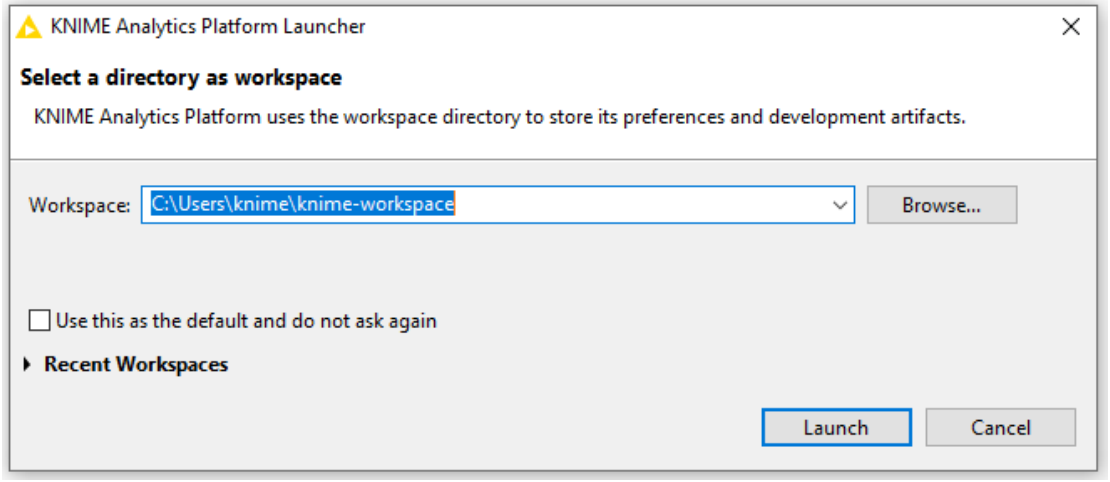
- Or go to the installation directory and launch KNIME via the knime.exe



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

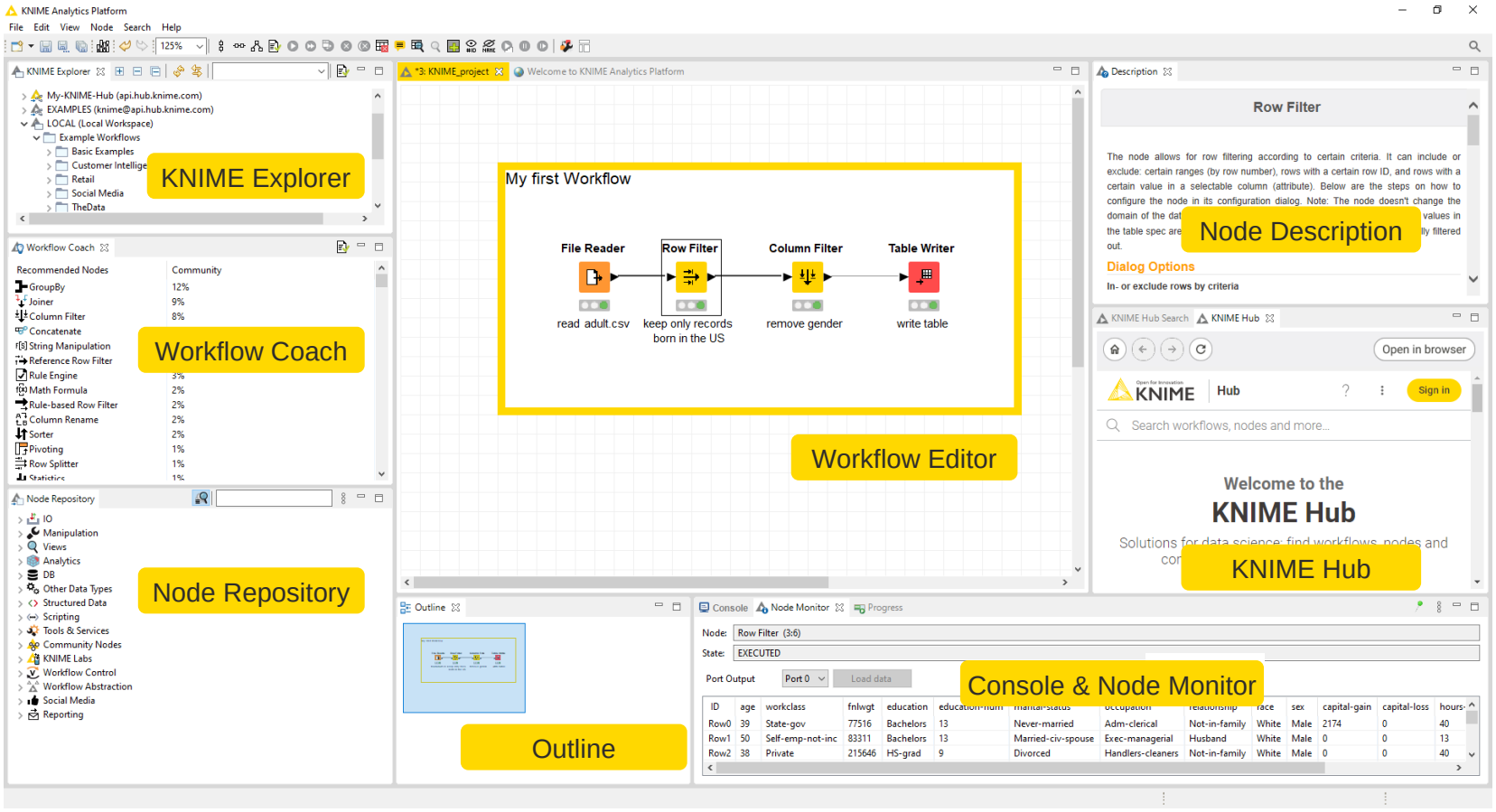
The KNIME Workspace

- The workspace is the **folder/directory** in which workflows (and potentially data files) are stored for the current KNIME session
- Workspaces are portable (just like KNIME)



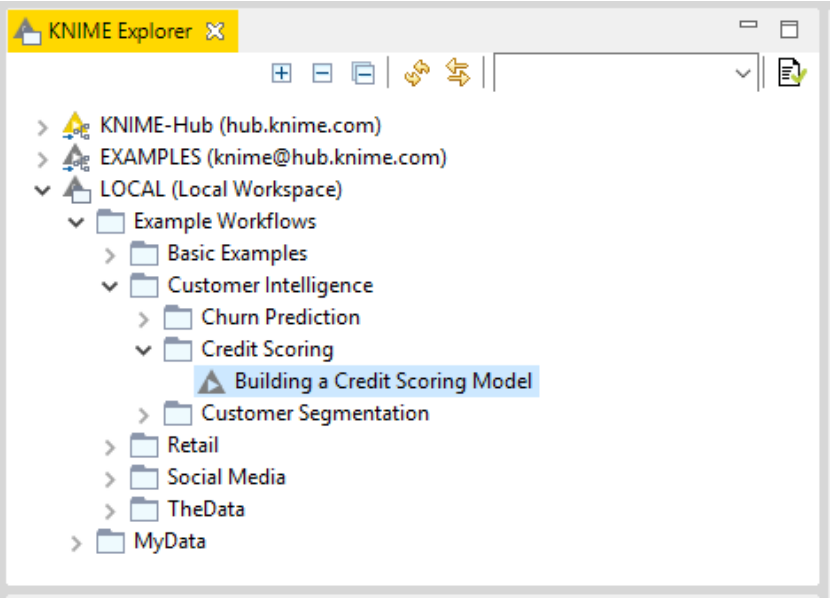
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The KNIME Analytics Platform Workbench



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

KNIME Explorer

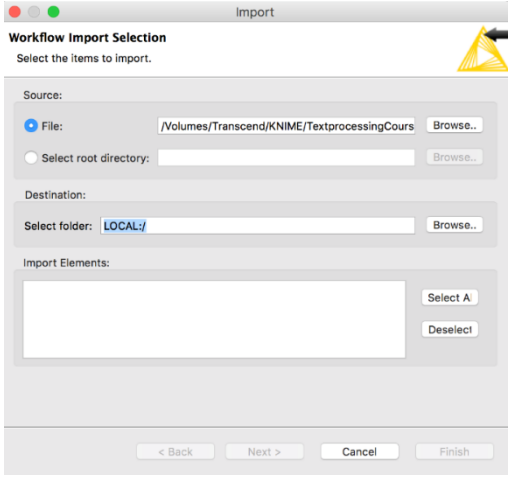


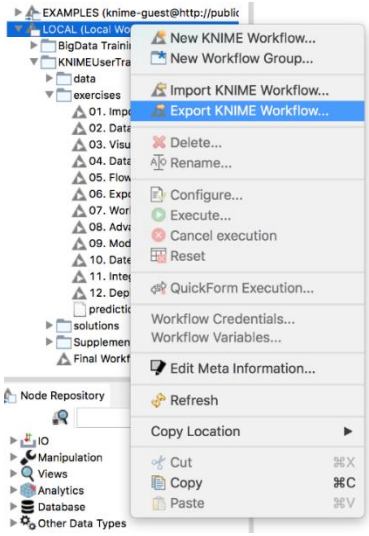
- In LOCAL you can access your own workflow projects.
- Other mountpoints allow you to connect to
  - EXAMPLE Server
  - KNIME Hub
  - KNIME Server
- The Explorer toolbar on the top has a search box and buttons to
  - select the workflow displayed in the active editor
  - refresh the view
- The KNIME Explorer can contain 4 types of content:
  - Workflows
  - Workflow groups
  - Data files
  - Shared Components

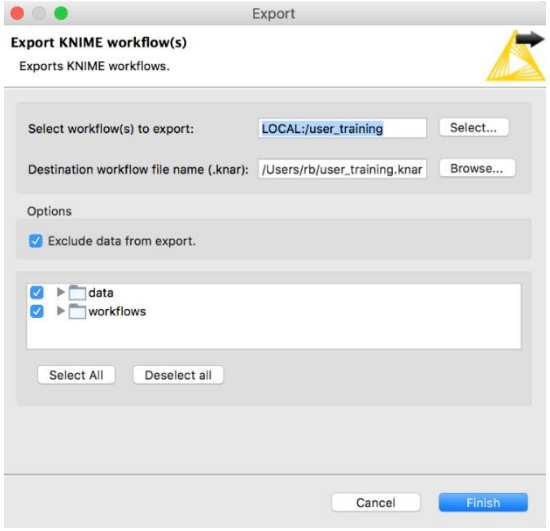
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Creating New Workflows, Importing and Exporting

- Right-click inside the KNIME Explorer to create a new workflow or a workflow group, or to import a workflow
- Right-click the workflow or workflow group to export





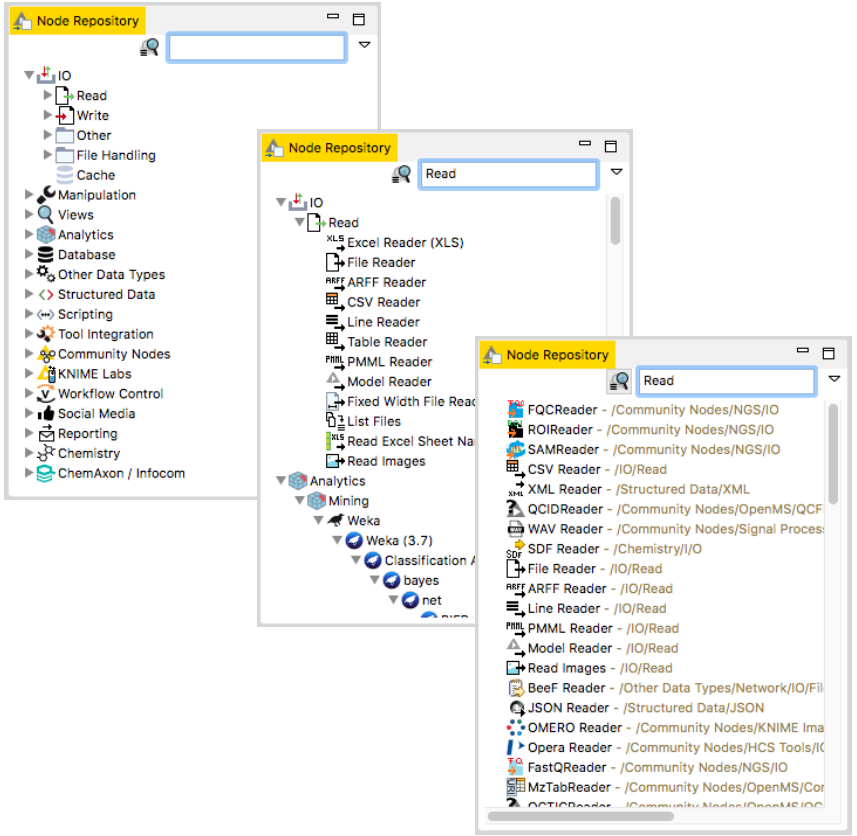


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Node Repository

- The Node Repository lists all KNIME nodes

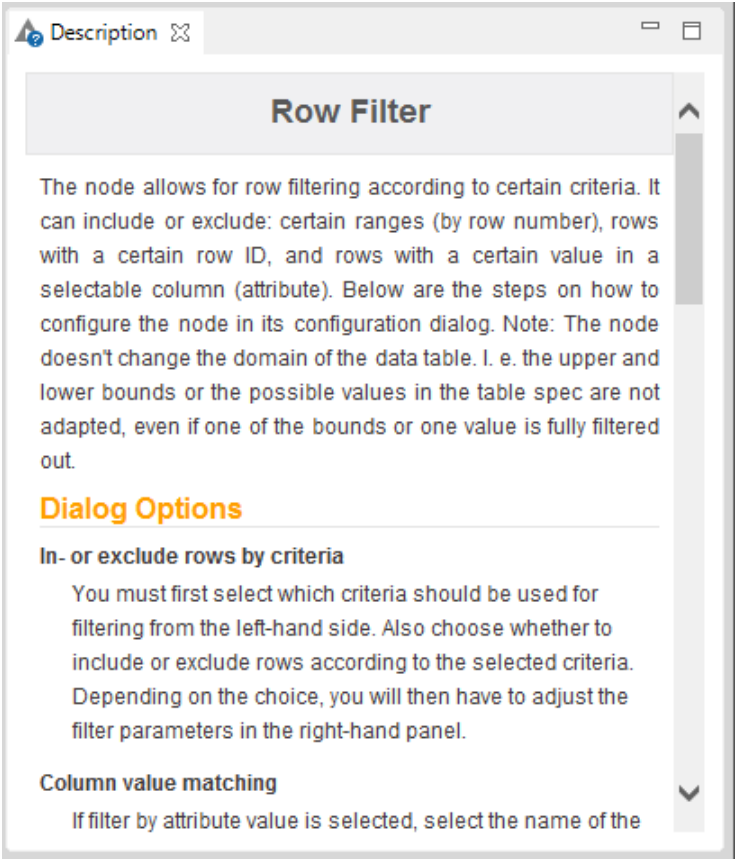
- The search box has 2 modes
  - Standard Search – exact match of node name
  - Fuzzy Search – finds the most similar node name



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Description

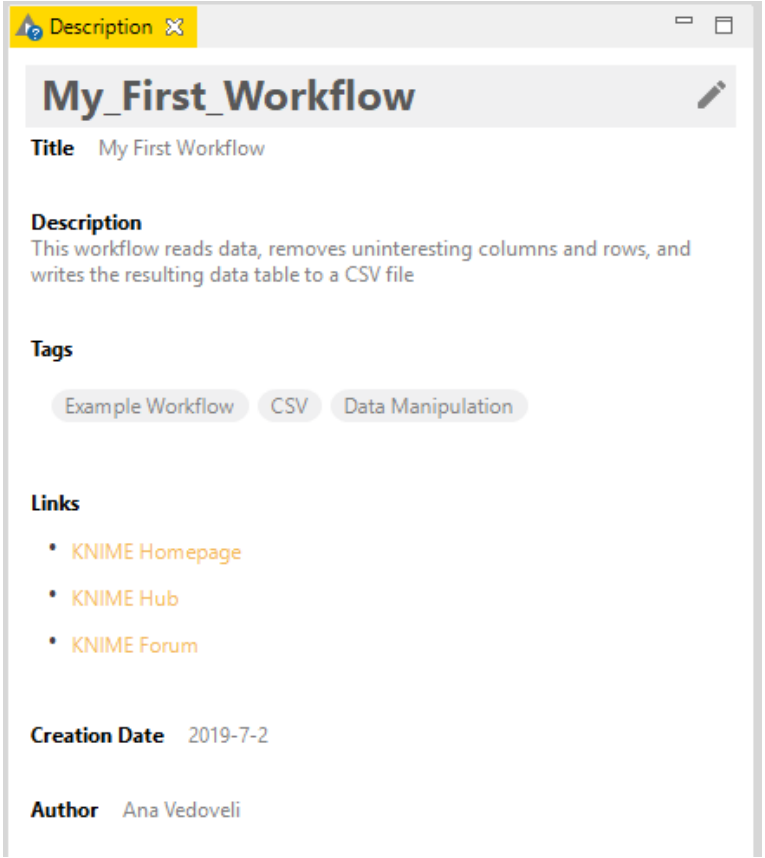
- The Description window gives information about:
  - Node Functionality
  - Input & Output
  - Node Settings
  - Ports
  - References to literature



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Workflow Description

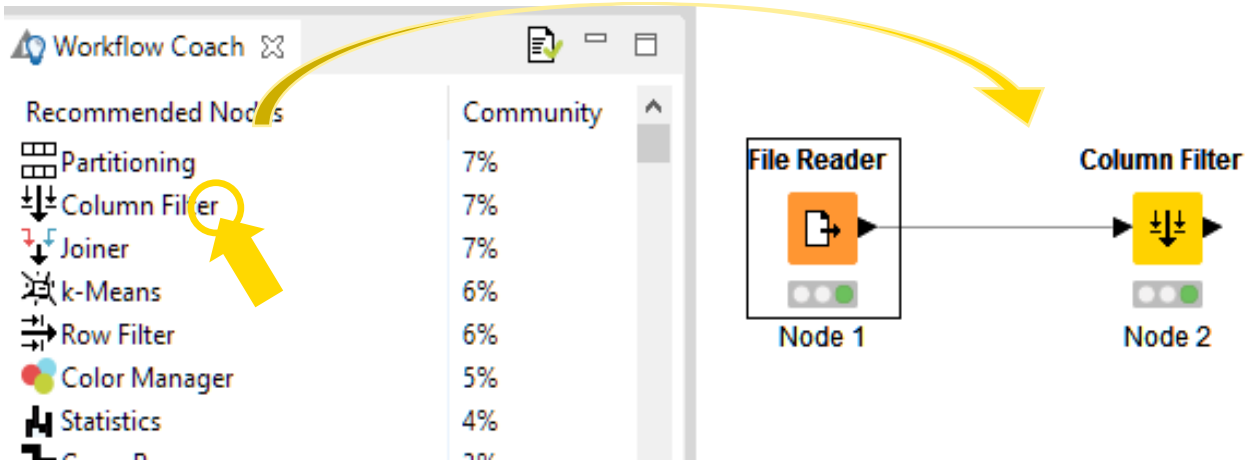
- When selecting the workflow, the Description window gives information about the workflow's:
  - Title
  - Description
  - Associated Tags and Links
  - Creation Date
  - Author



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Workflow Coach

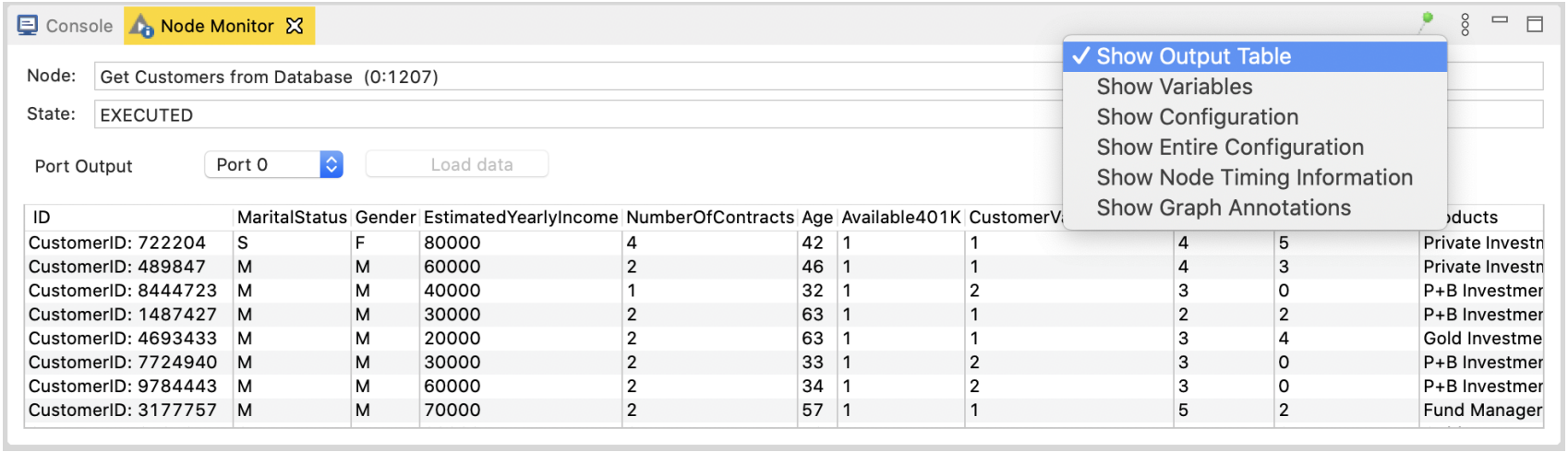
- Node recommendation engine
  - Gives hints about which node to use next in the workflow
  - Based on KNIME communities' usage statistics
  - Based on your own KNIME workflows



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Node Monitor

- By default the Node Monitor shows you the output table of the node selected in the workflow editor
- Click on the three dots on the upper right to show the flow variables, configuration, etc.



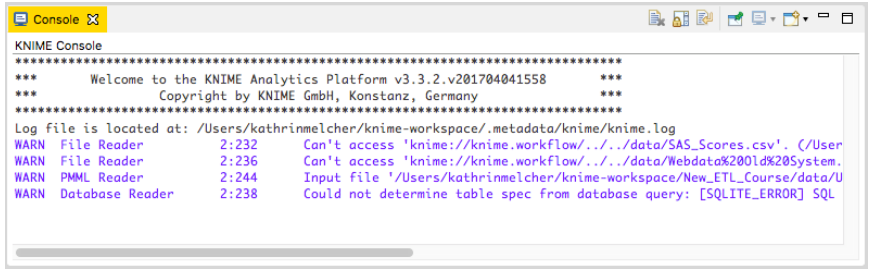
ID	MaritalStatus	Gender	EstimatedYearlyIncome	NumberOfContracts	Age	Available401K	CustomerV			oducts
CustomerID: 722204	S	F	80000	4	42	1	1	4	5	Private Investn
CustomerID: 489847	M	M	60000	2	46	1	1	4	3	Private Investn
CustomerID: 8444723	M	M	40000	1	32	1	2	3	0	P+B Investmer
CustomerID: 1487427	M	M	30000	2	63	1	1	2	2	P+B Investmer
CustomerID: 4693433	M	M	20000	2	63	1	1	3	4	Gold Investme
CustomerID: 7724940	M	M	30000	2	33	1	2	3	0	P+B Investmer
CustomerID: 9784443	M	M	60000	2	34	1	2	3	0	P+B Investmer
CustomerID: 3177757	M	M	70000	2	57	1	1	5	2	Fund Manager

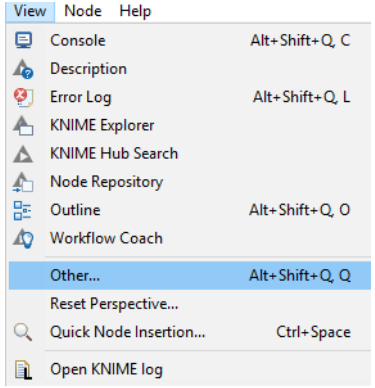
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

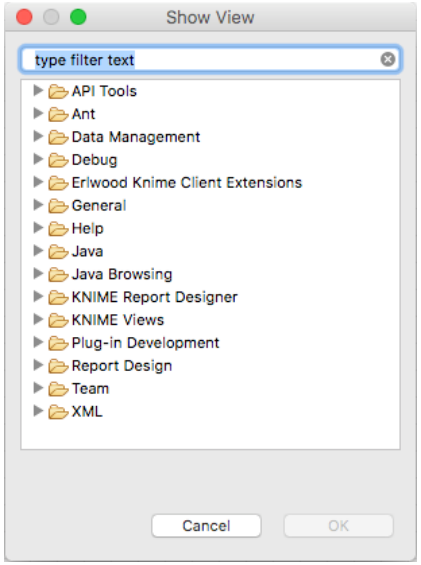
Console and Other Views

- Console view prints out error and warning messages about what is going on under the hood

- Click on View and select Other… to add different views
  - Node Monitor, Licenses, etc.



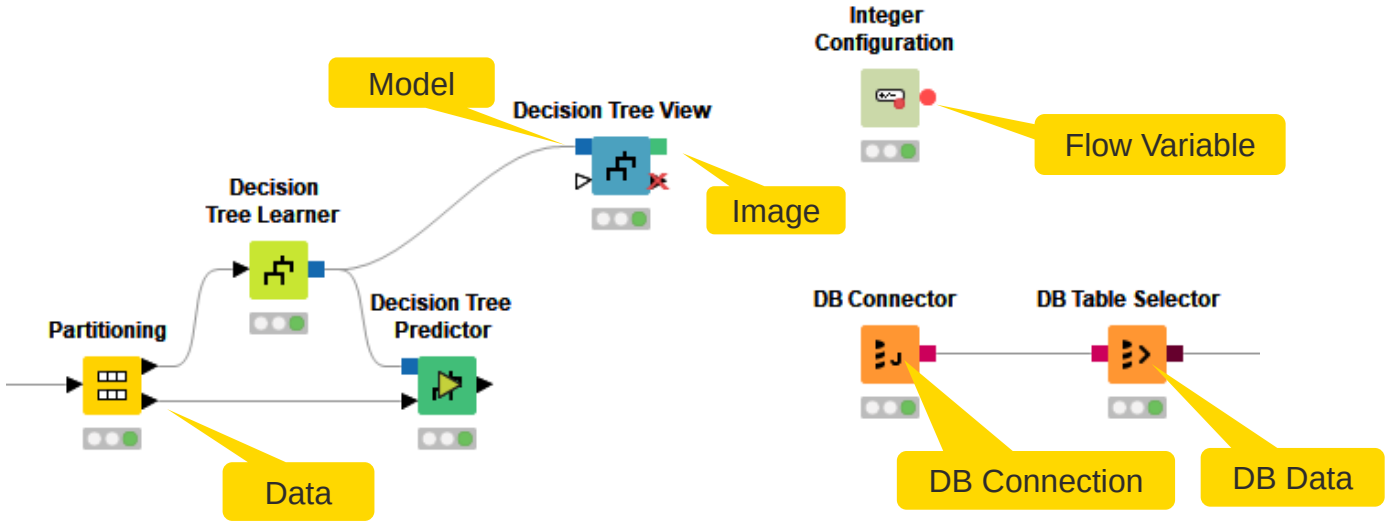




These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Inserting and Connecting Nodes

- Insert nodes into workspace by dragging them from Node Repository or by double-clicking in Node Repository
- Connect nodes by left-clicking the output port of Node A and dragging the cursor to the (matching) input port of Node B
- Common port types:



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

More on Nodes…

- A node can have 4 states:

Not Configured:
The node is waiting for configuration or incoming data

Configured:
The node has been configured correctly and can be executed

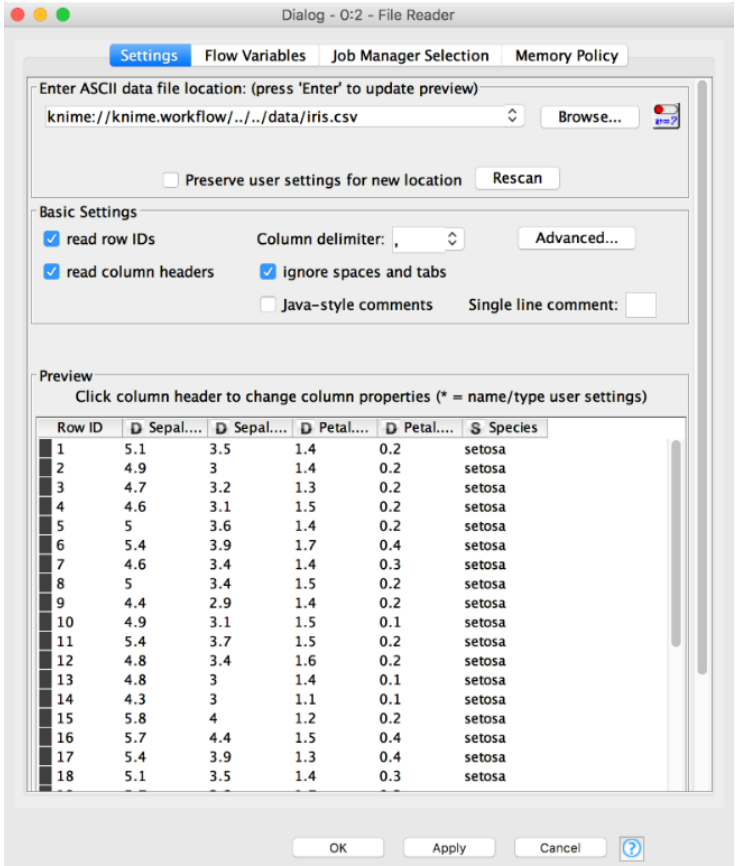
Executed:
The node has been successfully executed. Results may be viewed and used in downstream nodes

Error:
The node has encountered an error during execution

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Node Configuration

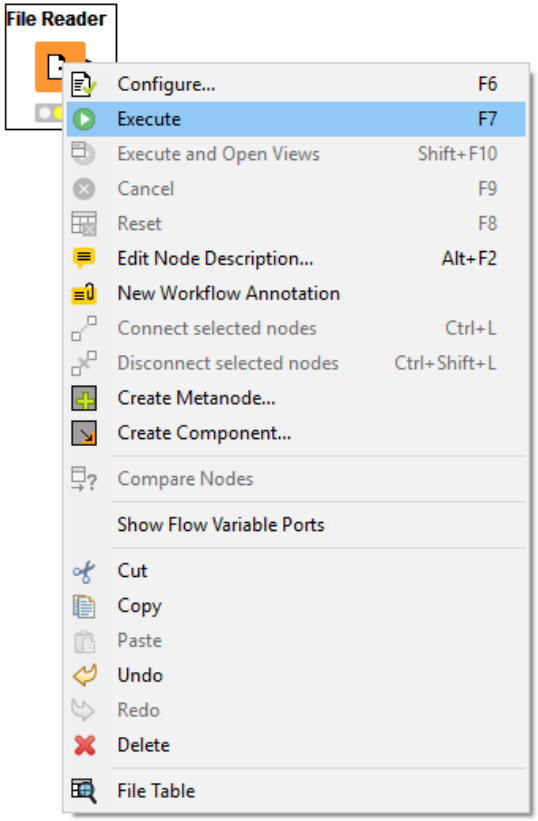
- Most nodes require configuration
- To access a node configuration window:
  - Double-click the node
  - Right-click -> Configure



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

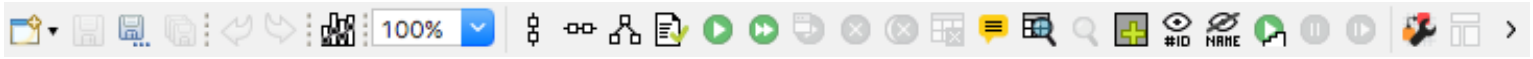
Node Execution

- Right-click node
- Select Execute in the context menu
- If execution is successful, status shows green light
- If execution encounters errors, status shows red light



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tool Bar



- The buttons in the toolbar can be used for the active workflow. The most important buttons are:

Execute selected and executable nodes (F7)

Execute all executable nodes

Execute selected nodes and open first view

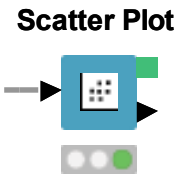
Cancel all selected, running nodes (F9)

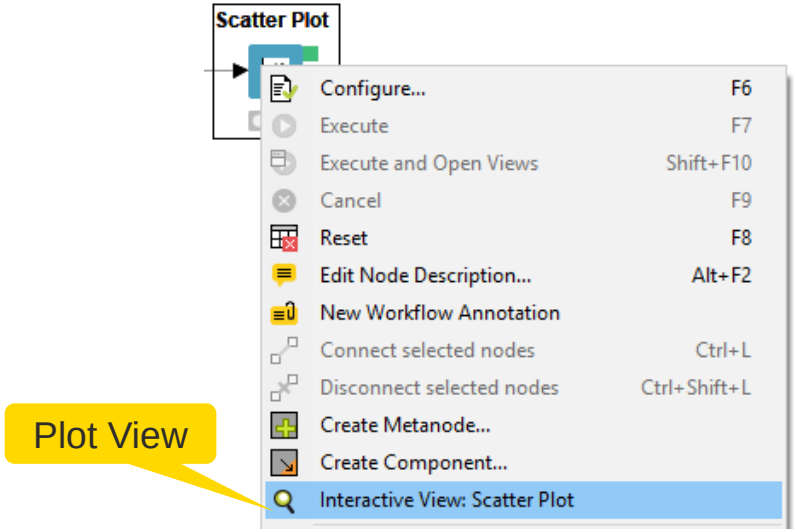
Cancel all running nodes

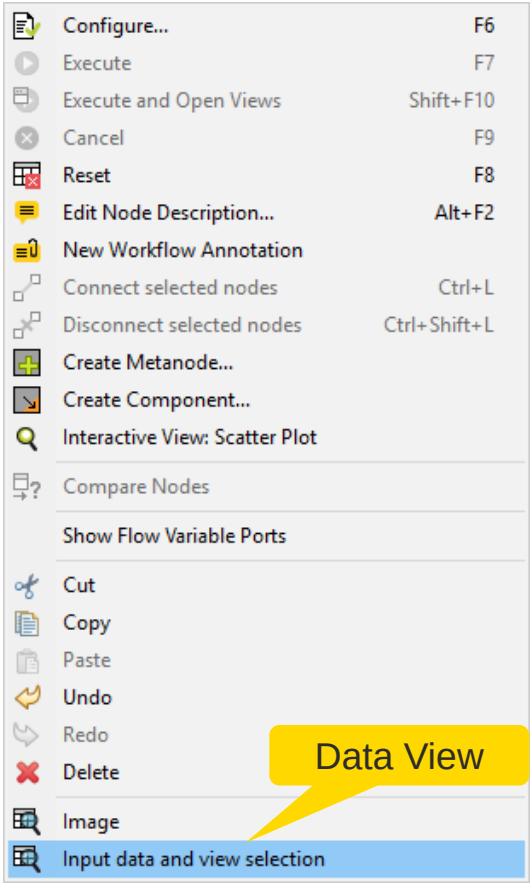
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Node Views

- Right-click node to inspect the execution results by
  - selecting output ports (last option in the context menu) to inspect tables, images, etc.
  - selecting Interactive View to open visualization results in a browser







These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

KNIME File Extensions

Dedicated file extensions for workflows and workflow groups associated with KNIME Analytics Platform

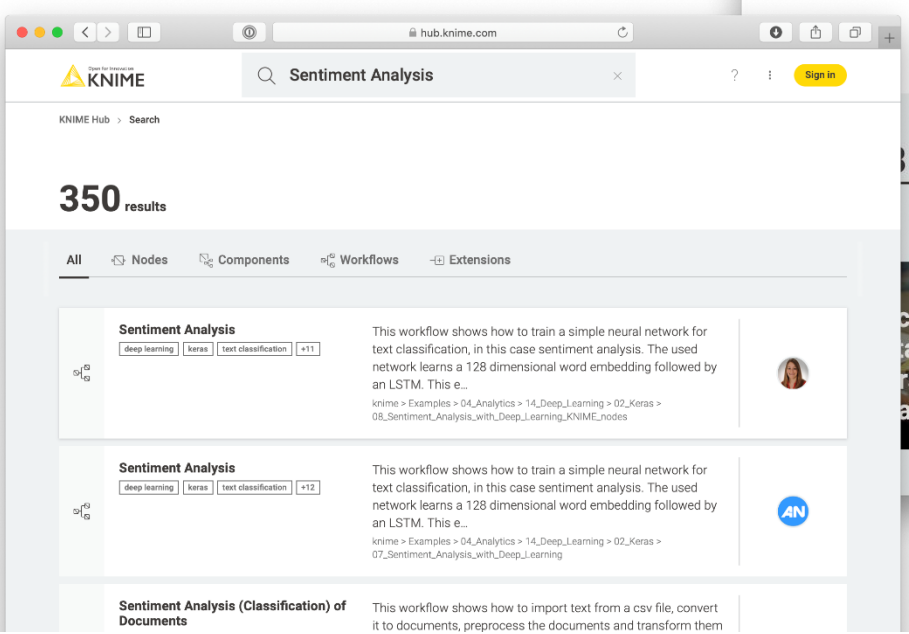
- ***.knwf** for KNIME Workflow Files

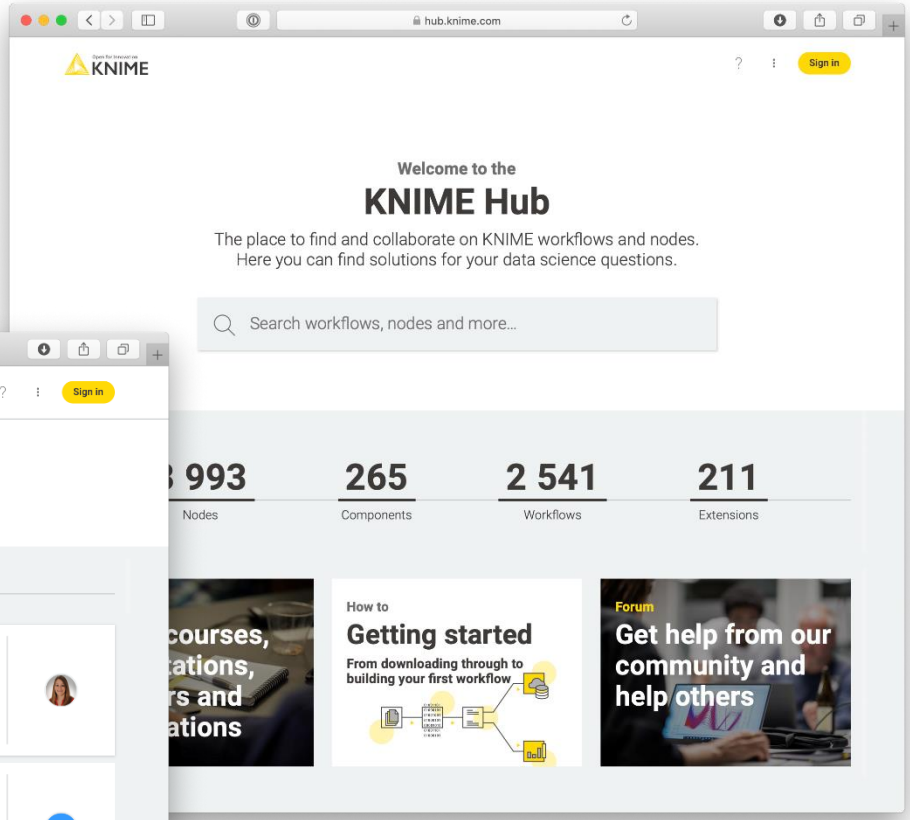
- ***.knar** for KNIME Archive Files

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Getting Started: KNIME Hub

- Place to search and share
  - Workflows
  - Nodes
  - Components
  - Extensions



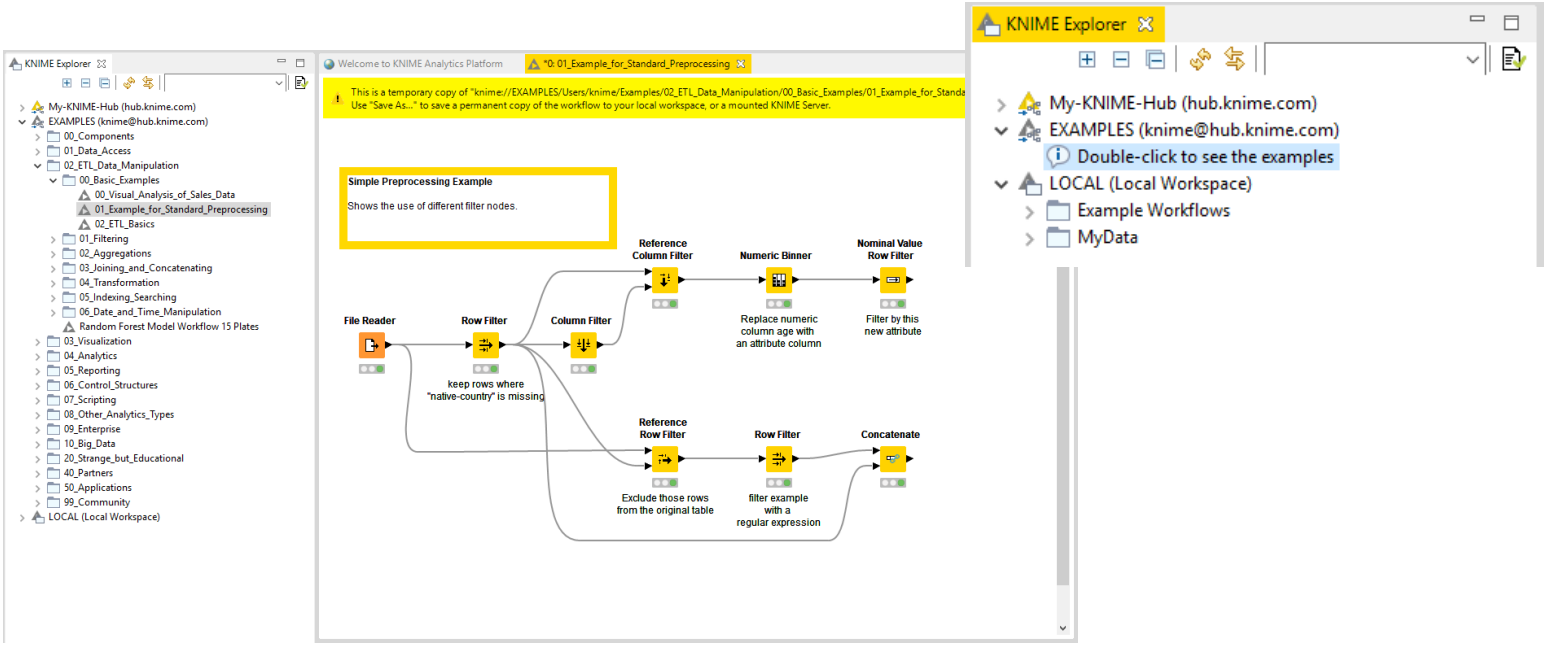


https://hub.knime.com

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Getting Started: KNIME Example Server

- Connect via KNIME Explorer to a public repository with large selection of example workflows for many, many applications



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

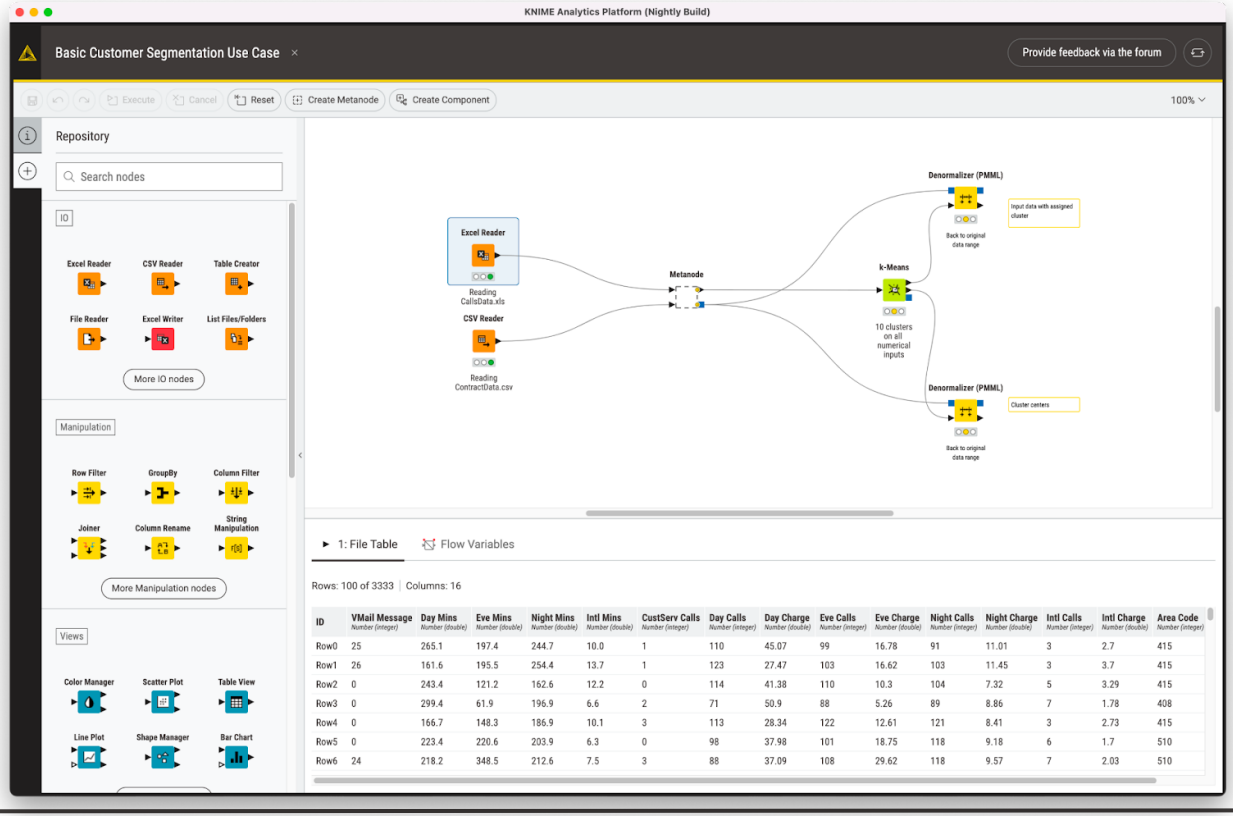
Hot Keys (for Future Reference)

Task	Hot key	Description
Node Configuration	F6	opens the configuration window of the selected node
Node Execution	F7	executes selected configured nodes
	Shift + F7	executes all configured nodes
	Shift + F10	executes all configured nodes and opens all views
	F9	cancels selected running nodes
	Shift + F9	cancels all running nodes
Node Connections	Ctrl + L	connects selected nodes
	Ctrl + Shift + L	disconnects selected nodes
Move Nodes and Annotations	Ctrl + Shift + Arrow	moves the selected node in the arrow direction
	Ctrl + Shift + PgUp/PgDown	moves the selected annotation in the front or in the back of all overlapping annotations
Workflow Operations	F8	resets selected nodes
	Ctrl + S	saves the workflow
	Ctrl + Shift + S	saves all open workflows
	Ctrl + Shift + W	closes all open workflows
Metanode	Shift + F12	opens metanode wizard

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

KNIME Modern UI Preview (Labs)

- Preview KNIME Analytics Platform's makeover
  - Install KNIME Modern UI Preview extension and click the "Open KNIME Modern UI Preview"



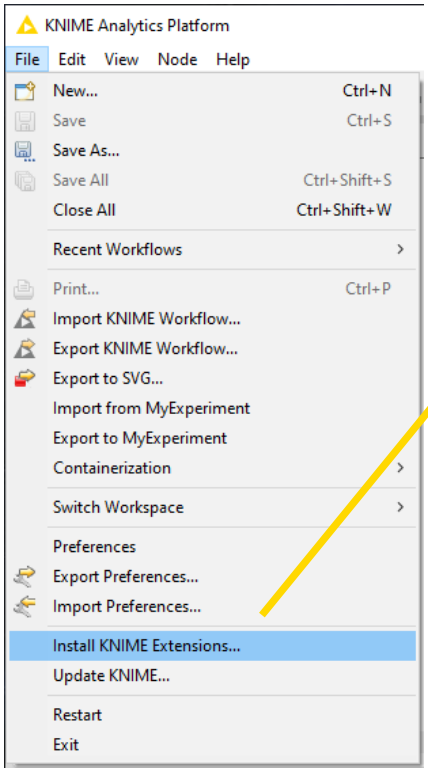
ID	VMail Message	Day Mins	Eve Mins	Night Mins	Intl Mins	CustServ Calls	Day Calls	Day Charge	Eve Calls	Eve Charge	Night Calls	Night Charge	Intl Calls	Intl Charge	Area Code
	Number (integer)	Number (double)	Number (double)	Number (double)	Number (double)	Number (integer)	Number (integer)	Number (double)	Number (integer)	Number (double)	Number (integer)	Number (double)	Number (integer)	Number (double)	Number (integer)
Row0	25	265.1	197.4	244.7	10.0	1	110	45.07	99	16.78	91	11.01	3	2.7	415
Row1	26	161.6	195.5	254.4	13.7	1	123	27.47	103	16.62	103	11.45	3	3.7	415
Row2	0	243.4	121.2	162.6	12.2	0	114	41.38	110	10.3	104	7.32	5	3.29	415
Row3	0	299.4	61.9	196.9	6.6	2	71	50.9	88	5.26	89	8.86	7	1.78	408
Row4	0	166.7	148.3	186.9	10.1	3	113	28.34	122	12.61	121	8.41	3	2.73	415
Row5	0	223.4	220.6	203.9	6.3	0	98	37.98	101	18.75	118	9.18	6	1.7	510
Row6	24	218.2	348.5	212.6	7.5	3	88	37.09	108	29.62	118	9.57	7	2.03	510

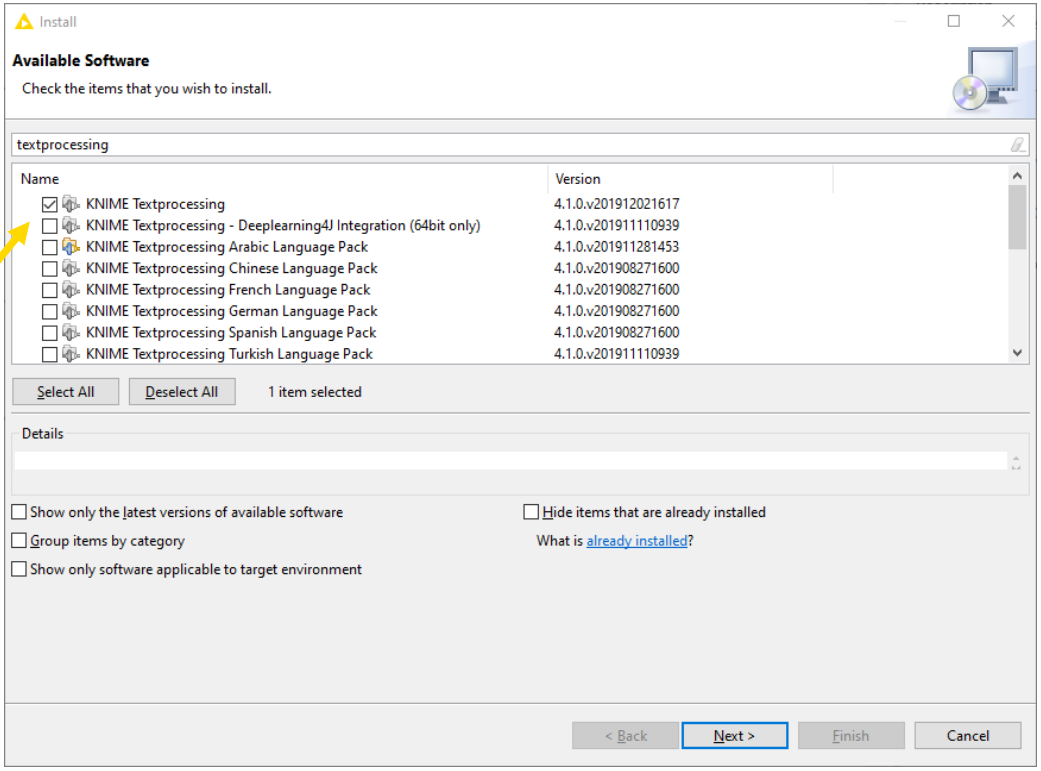
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The KNIME Text Processing Extension

Installation

Install KNIME Textprocessing Extension from menu…

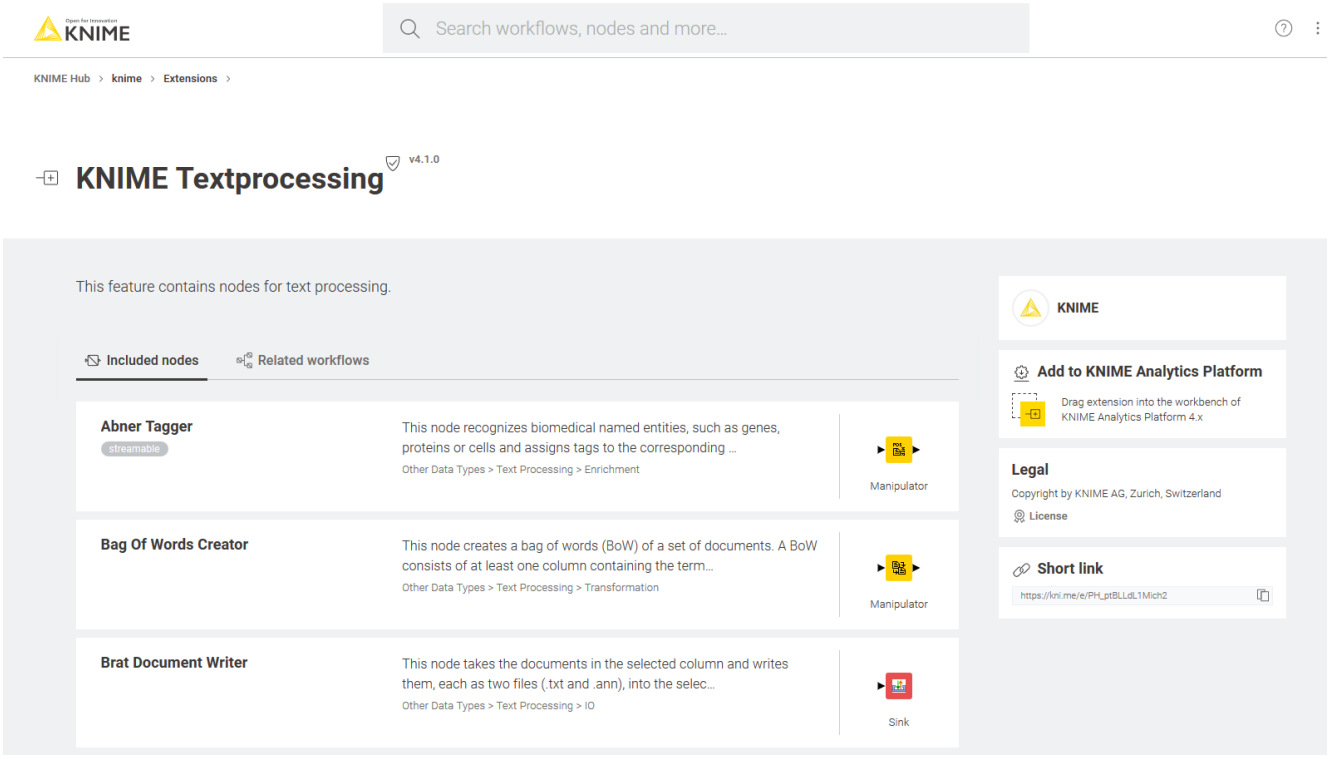




These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Installation

…or drag-and-drop from KNIME Hub



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

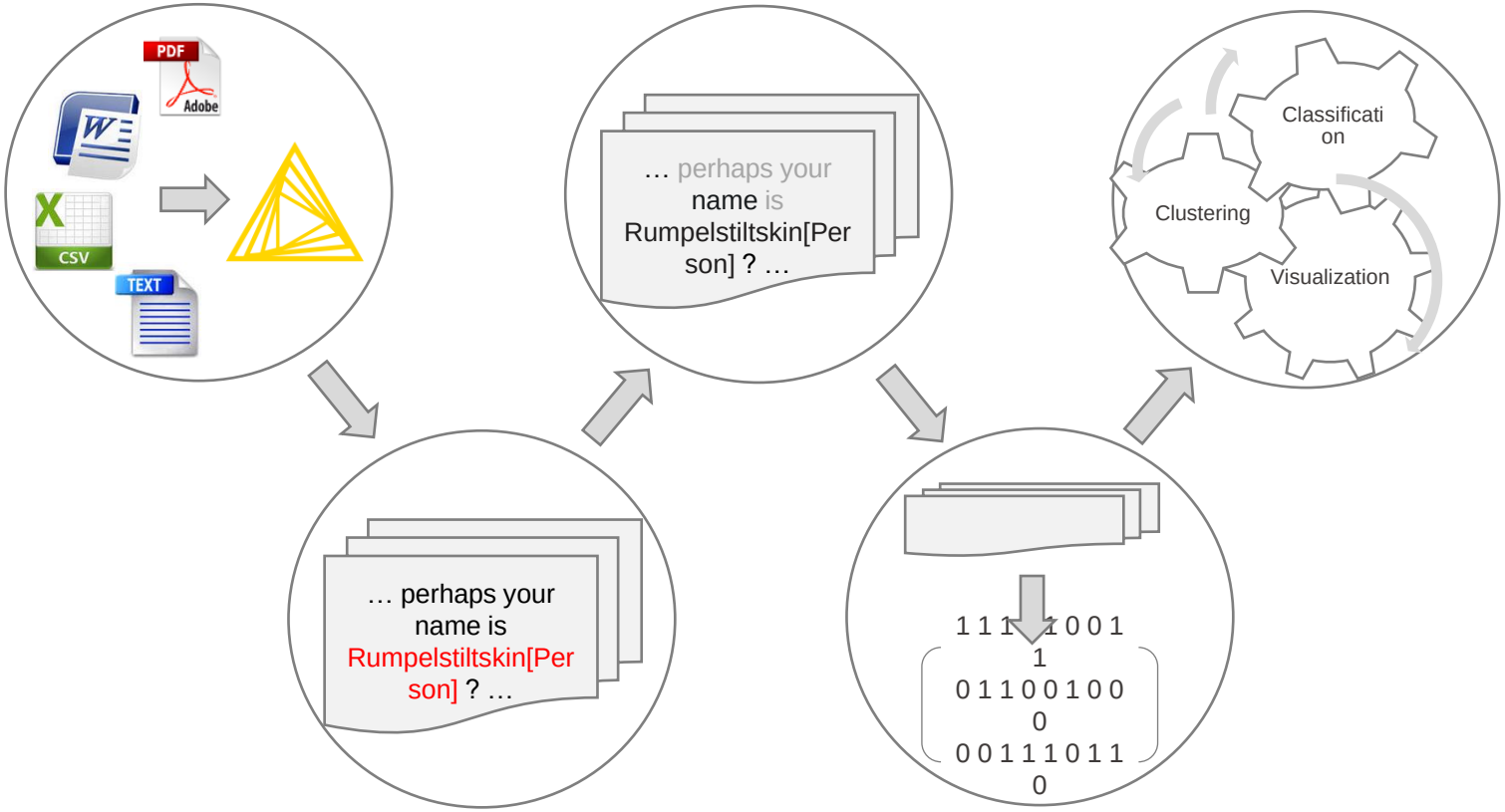
Tips for Better Performance

- Increase maximum memory for KNIME
- Edit knime.ini
  - Add “-Xmx3G” as last line of knime.ini file
  - Replace 3 by the amount of gigabytes allocated for KNIME

- Useful additional extensions
  - XML-Processing (KNIME extension)
    - Parsing and processing of XML documents
  - KNIME JavaScript Views (Labs)
    - Tagged Document Viewer

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Philosophy



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Additional Data Types: Document

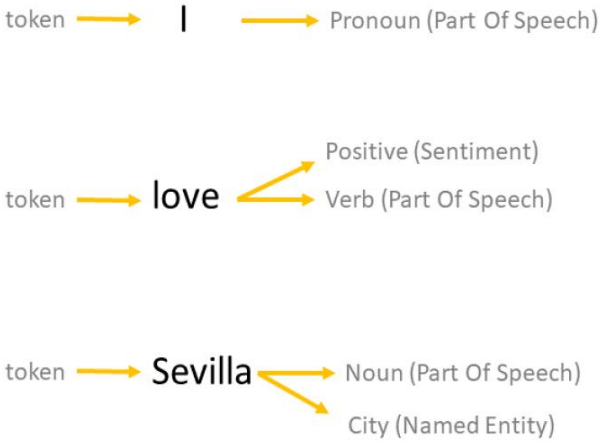
- KNIME uses a composite/aggregate data type to represent textual content
- Fields include:
  - Title
  - Text
  - Source
  - Category
  - Author(s)
  - Date, …
  - Generic Meta Data

Document
"Great food , interesting service"
"Excellent Lunch Destination"
"Hidden treasure near KaDaWe"
"Excellent Food Very Reasonable !"
"Good food , great prices !"
"Nice food at a reasonable price"

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Additional Data Types: Term

- KNIME uses a composite/aggregate data type to represent terms [keywords]
- Fields include:
  - Sentiment
  - POS tag
  - City
  - Person name
  - Etc.



Term
Very[RB(POS)]
good[JJ(POS)]
Thai[NNP(POS)]
food[NN(POS)]
![SYM(POS)]
Been[NNP(POS)]
there[EX(POS)]
on[IN(POS)]
Monday[NNP(POS)]
and[CC(POS)]
had[VBD(POS)]
a[DT(POS)]
great[JJ(POS)]
time[NN(POS)]

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Part of Speech Tags

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

Source: Penn-Treebank Project.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Data Table Structures

- Document table
  - List of documents
- Bag of words
  - Tuples of documents and terms
- Document vectors
  - Numerical representations of documents

Table "default" - Rows: 440 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Document
Row0_1_Row...	"Great food , interesting service"
Row0_2_Row...	"Excellent Lunch Destination"
Row0_3_Row...	"Hidden treasure near KaDaWe"
Row0_4_Row...	"Excellent Food Very Reasonable !"
Row0_5_Row...	"Good food , great prices !"

Table "default" - Rows: 21290 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Term	Document
Row7804	Nice[NNP(POS)]	"Nice Thai Food"
Row7805	Thai[NNP(POS)]	"Nice Thai Food"
Row7806	Food[NNP(POS)]	"Nice Thai Food"
Row7807	Good[NNP(POS)]	"Nice Thai Food"
Row7808	food[NN(POS)]	"Nice Thai Food"

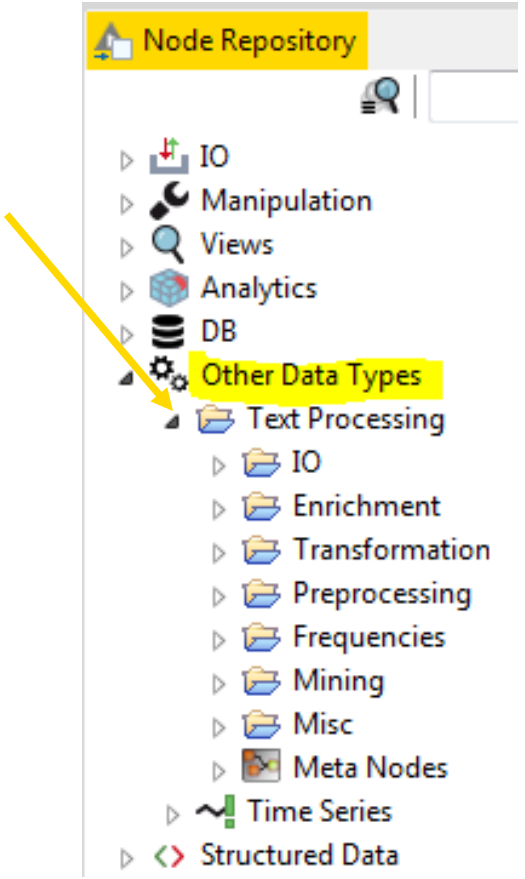
Table "default" - Rows: 440 | Spec - Columns: 2613 | Properties | Flow Variables

Row ID	Document	afford	truly	elegant	restaur...
1	"...if you can afford...	1	1	1	1
2	"2 michelin stars wel...	0	0	0	1
3	"A Michelin Star exp...	0	0	1	1
4	"A Unique Place to ...	0	0	0	0
5	"A fun place to hav...	0	0	0	1

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Exercise

- Open KNIME
- Install KNIME Textprocessing Extension
- Import workflows from the USB stick (or download from provided URL)

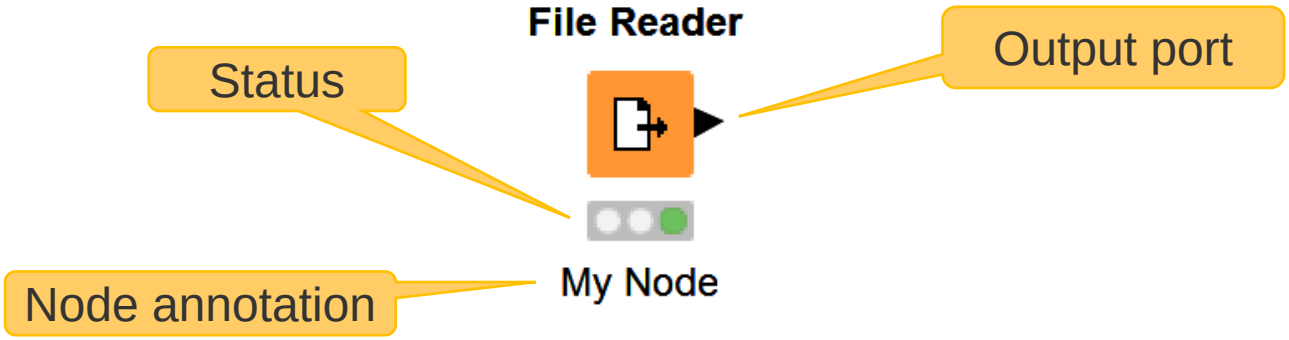


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Importing Text

Data Source Nodes

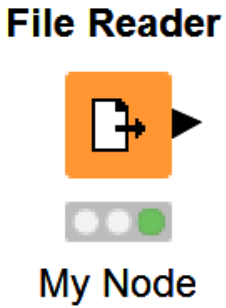
- Typically characterized by:
  - Orange color
  - No input ports, 1 output port



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

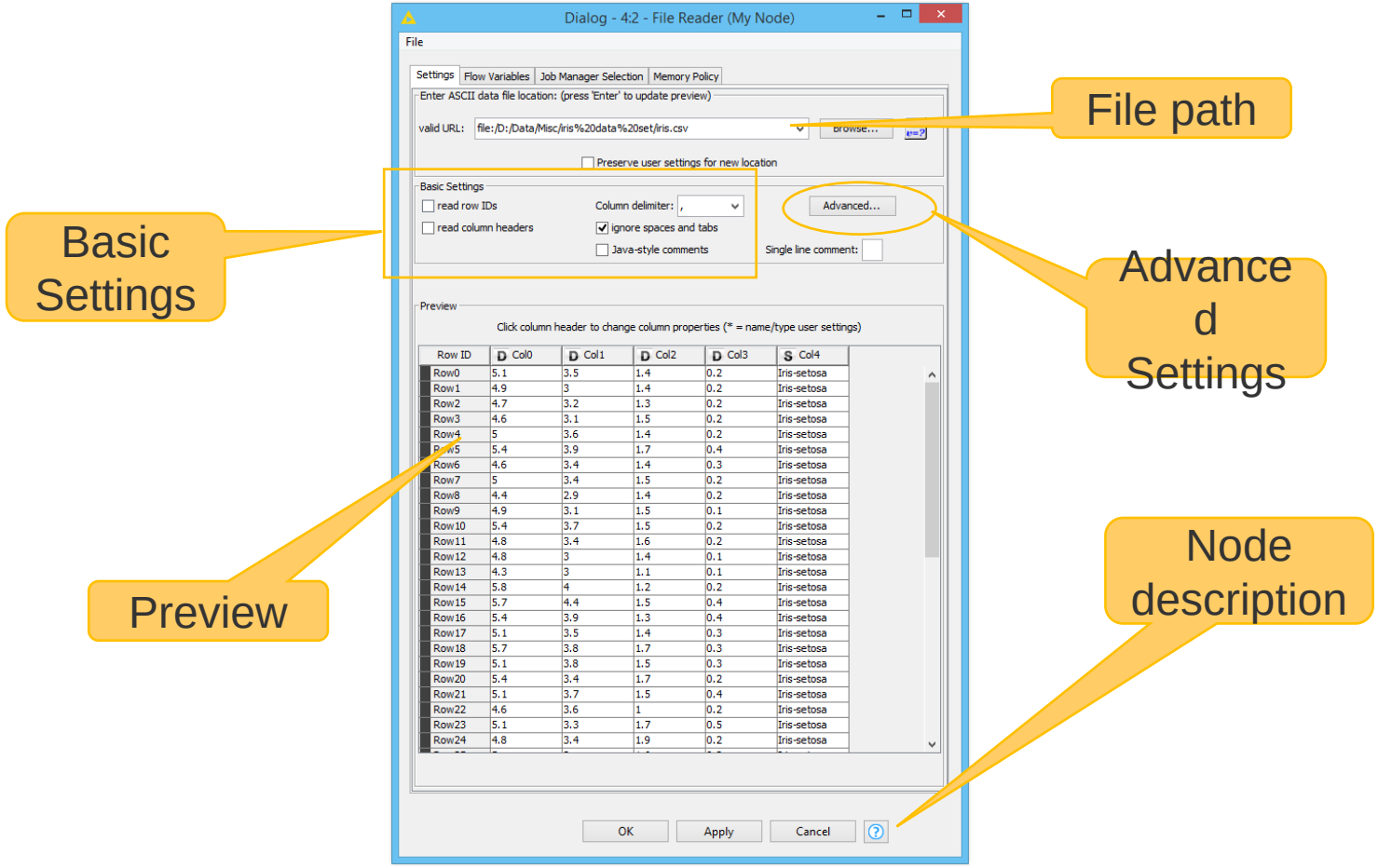
New Node: File Reader

- Workhorse of the KNIME Source nodes
  - Reads text based files
  - Many advanced features allow it to read most ‘weird’ files



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

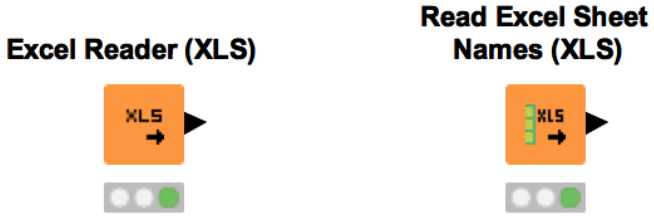
File Reader: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

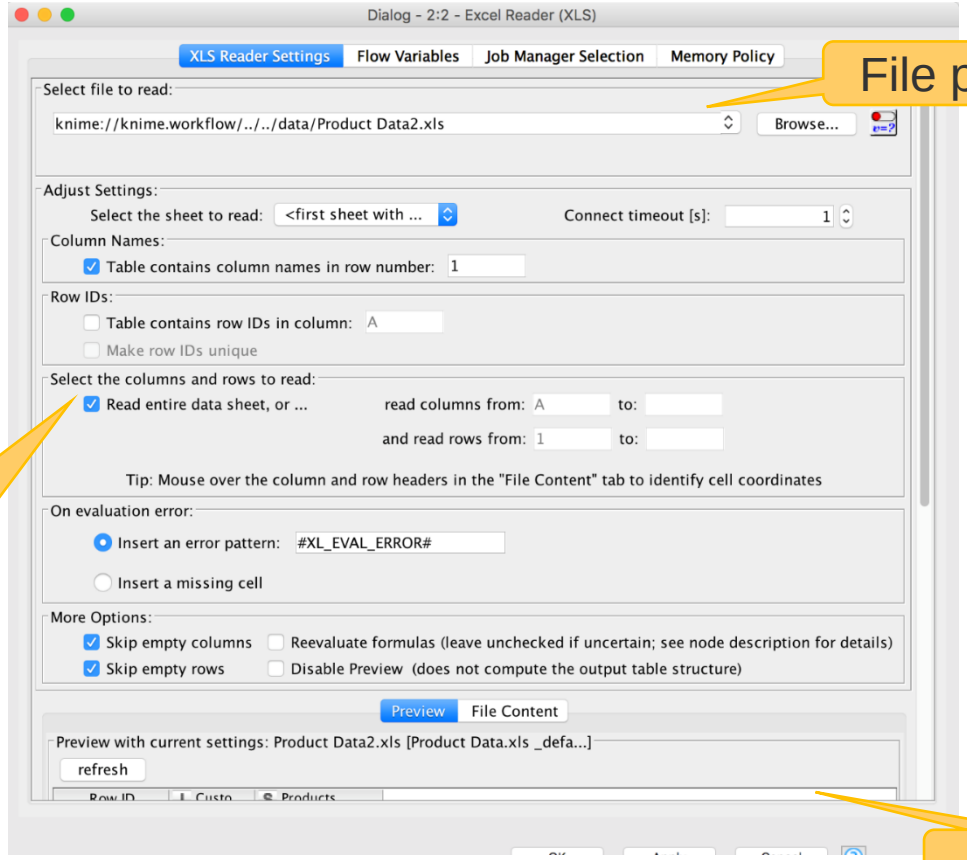
New Node: Excel Reader (XLS)

- Reads .xls and .xlsx file from Microsoft Excel
  - Supports reading from multiple sheets



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

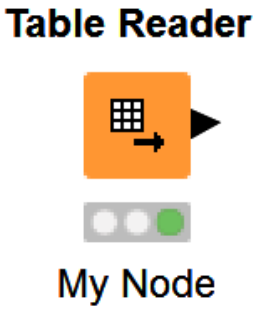
Excel Reader Configuration

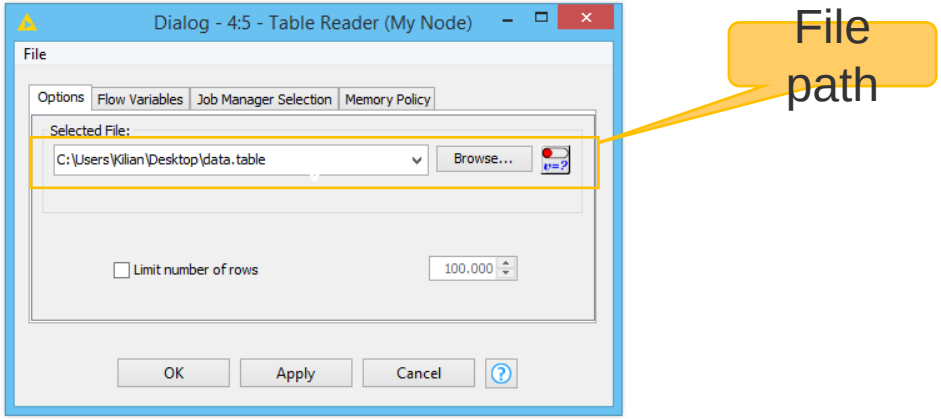


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Table Reader

- Reads tables from the native KNIME Format
- Maximum performance
- Minimum configuration





These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

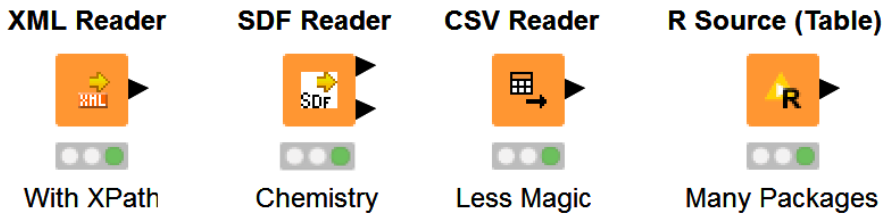
New Node: Database Reader

- Connectors for Common DB types
  - MySQL, Postgres, SQLite
- Also works with any JDBC driver
- Common nodes for SQL Query Building
  - GroupBy, Join, Filter, Sort

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Other Interesting Nodes

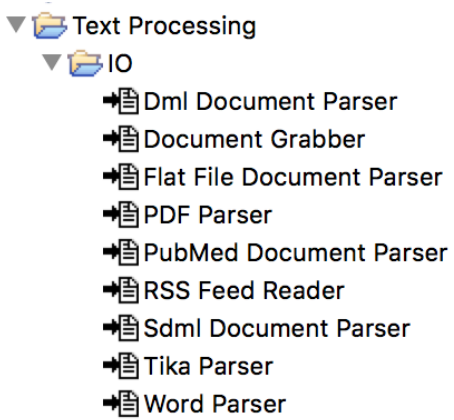
- PMML Reader – reads standard predictive models
- XML Reader with XPATH support
- REST/SOAP, and many more



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Parser Nodes

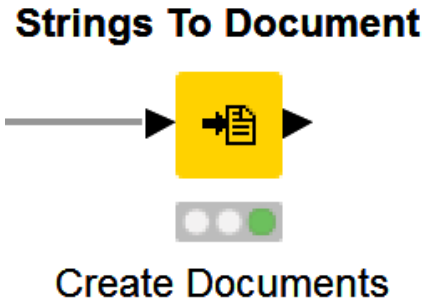
- Node Repository: Other Data Types/Text Processing/IO
- Available Parser Nodes
  - Flat File Document Parser
  - PDF Parser
  - Word Parser
  - Document Grabber
  - …



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

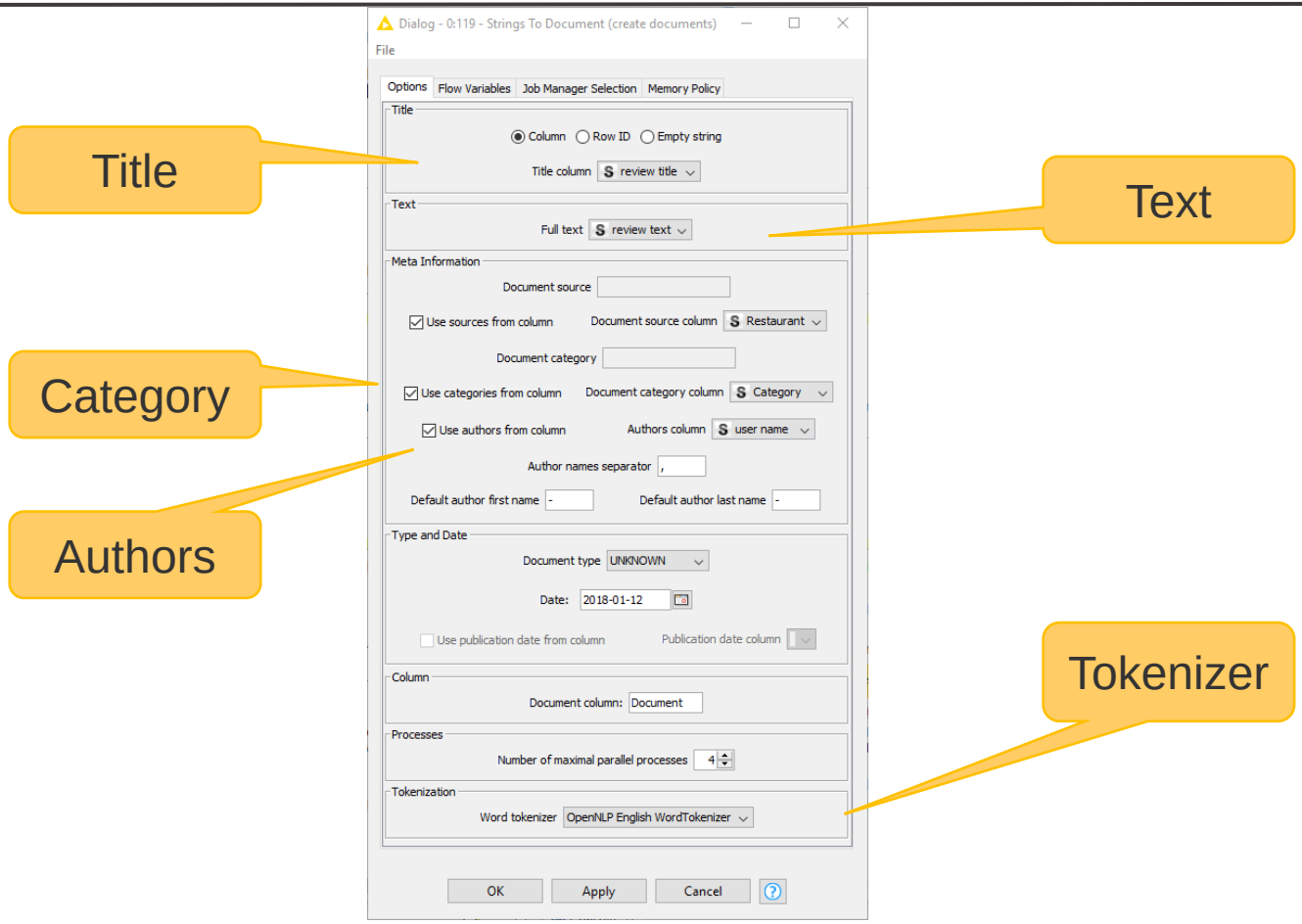
New Node: Strings To Document

- Creation of document cells from strings
  - Converts string cells to document cells
  - Useful in combination with File Reader, XLS Reader, database nodes



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

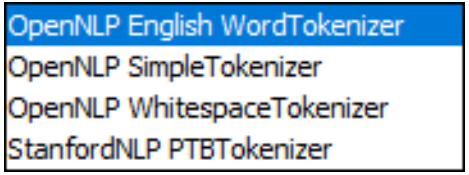
Strings To Document: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tokenizers

- Different tokenizers are available leading to slightly different terms extracted from the document



Example: "*I'm enjoying the tutorial*"

English WordTokenizer

WhitespaceTokenizer

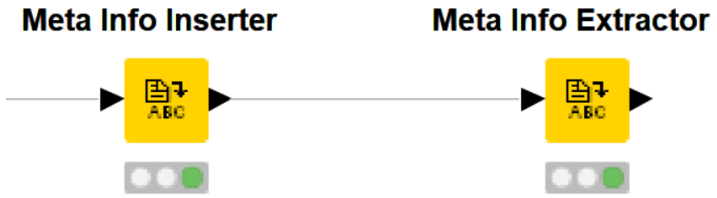
"*I*", "*'m*", "*enjoying*", "*the*", "*tutorial*"

"*I'm*", "*enjoying*", "*the*", "*tutorial*"

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Meta Info Inserter & Extractor

- Inserter allows adding document meta info
  - Adds meta info to documents as key value pairs
  - Helpful if more meta info available than covered by Strings to Documents node
- Extractor brings data back from document cell into table columns
  - Each key results in a column, containing the specific values for each document related to that key.



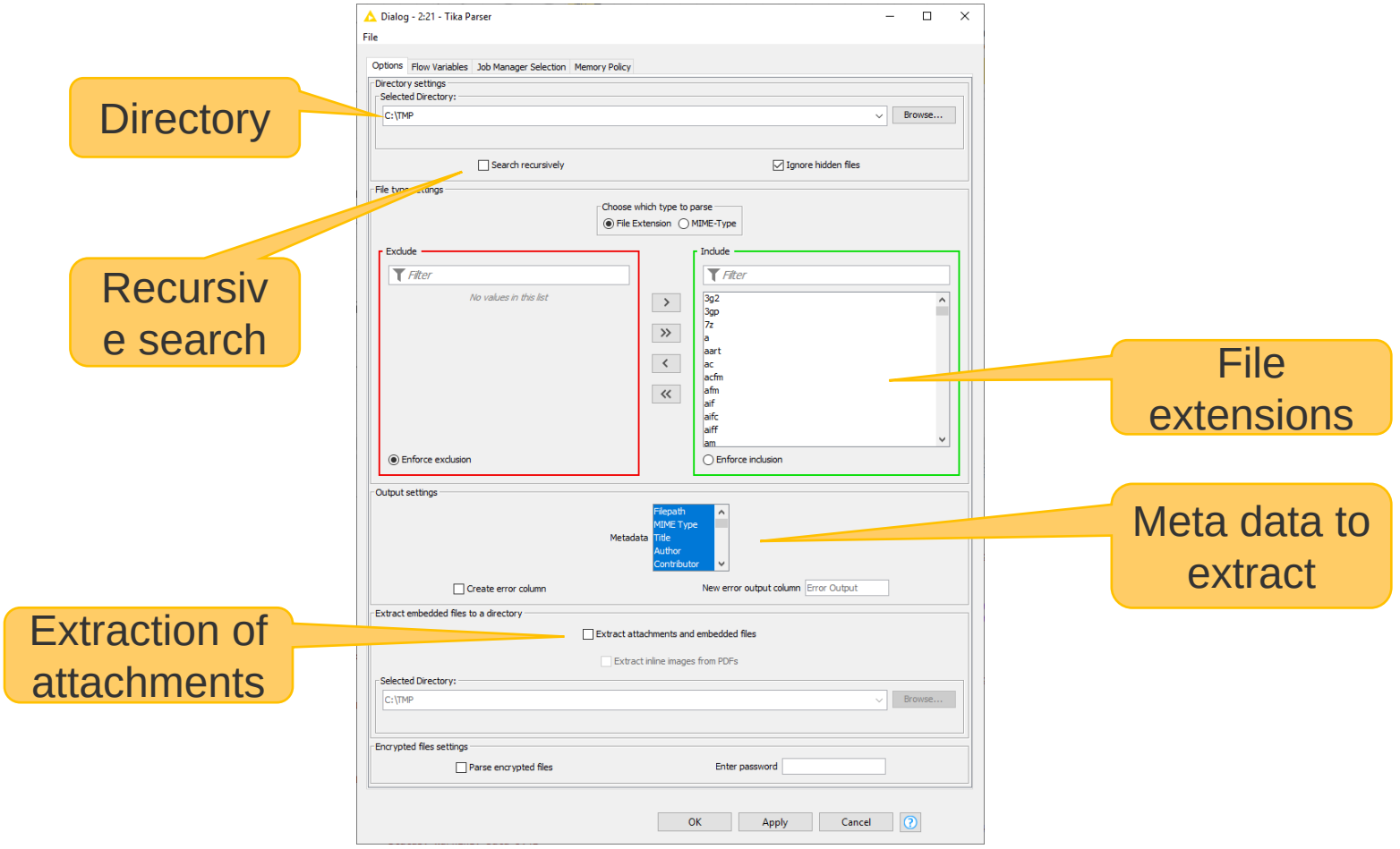
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Tika Parser

- Reads files of various formats from directory
  - Searches for all files with specified extension in directory
  - Creates one document for each file
  - Extracts specified (meta) information

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tika Parser: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Exercise

- Start with “Exercise: Importing text”
  - Import string data from:
    - TripadvisorReviews-SanFranciscoRestaurants-ItalianChineseFood.table
  - Filter rows with missing titles
  - Convert strings to documents
  - Filter all columns except the document column

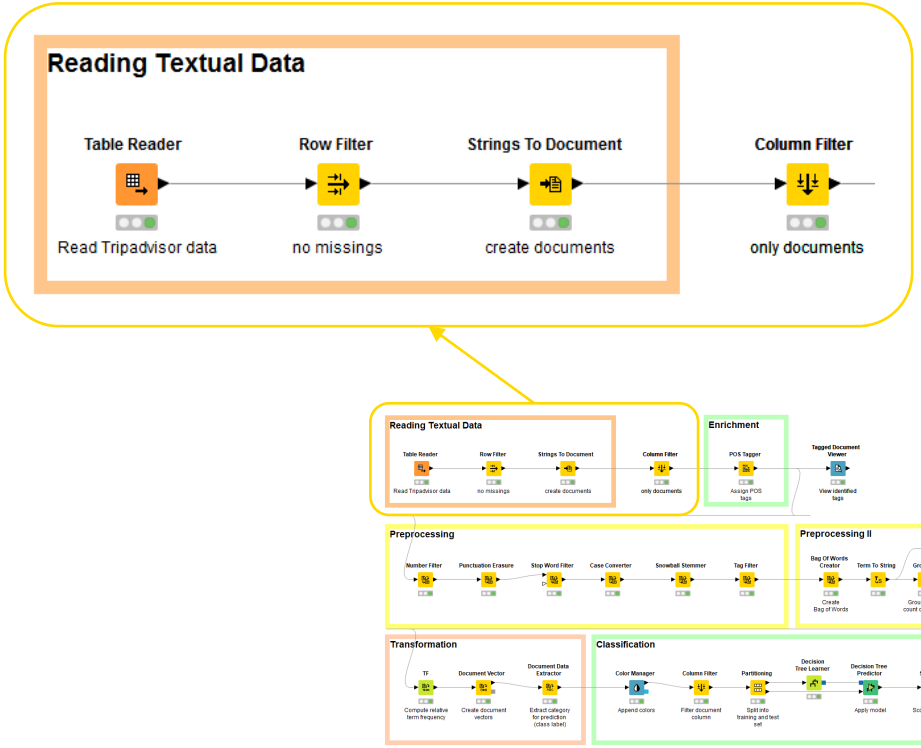
You can download the training workflows from the KNIME Hub: https://hub.knime.com/knime/spaces/Education/latest/02%20KNIME%20Text%20Mining%20Course/

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Solution

Import text

- Table Reader
- Row Filter
- Strings to Documents
- Column Filter

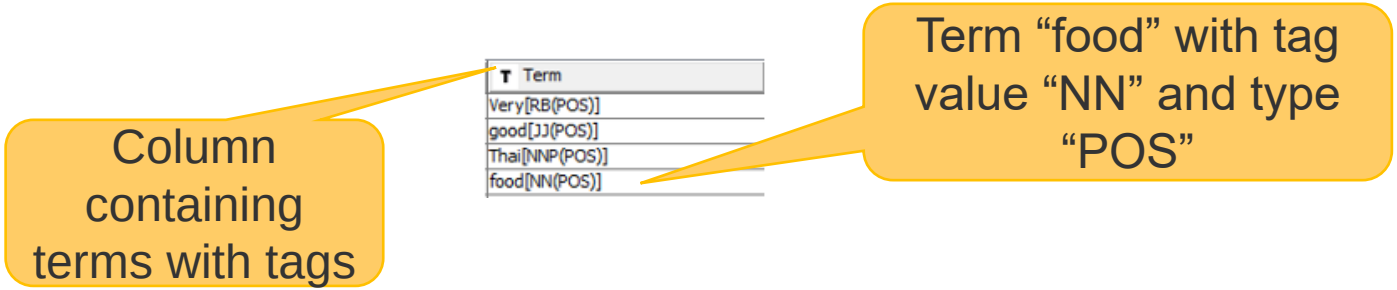


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Enrichment

Enrichment

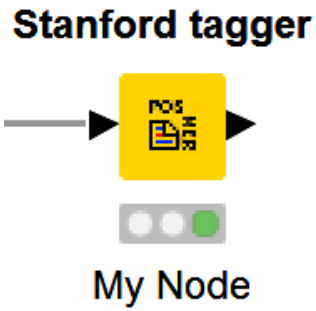
- Semantic information is indicated by a tag assignment
  - Part of speech, named entities (persons, organizations, genes, …), sentiments
- A tag consists of a type and a value
  - Type represents the class or set of tags
    - e.g. POS (part of speech), NE (named entity)
  - Value represents the actual tag value
    - e.g. NN (noun), PERSON



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tagger Nodes

- Typically characterized by:
  - Yellow color
  - 1 to 2 input ports (requiring one document column), 1 output port
  - Assignment of semantic information (tags) to terms



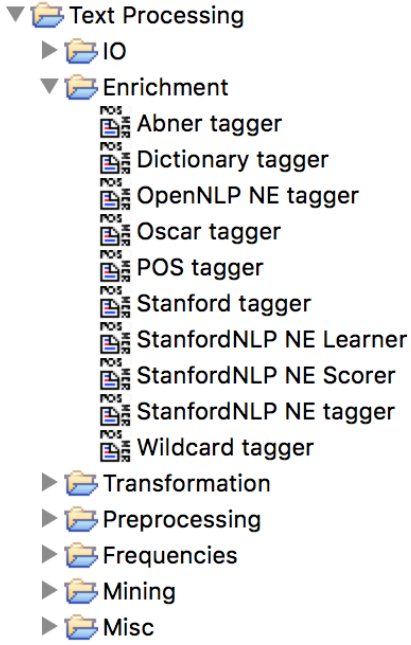
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tagger Nodes

- Node Repository:

  Other Data Types/Text Processing/Enrichment

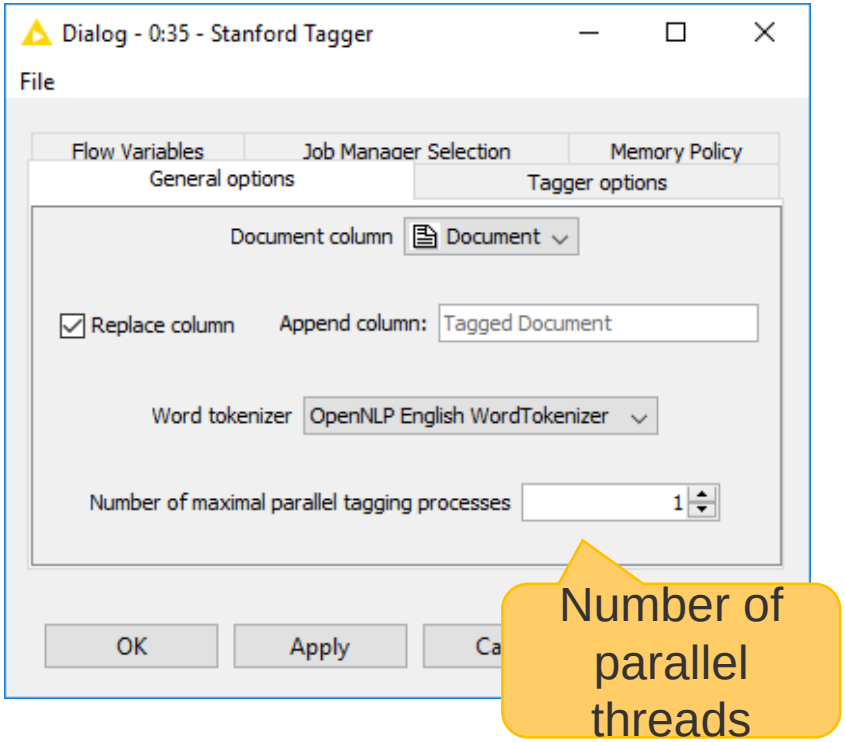
- Available Tagger Nodes
  - Stanford tagger
  - Dictionary (& Wildcard) tagger
  - OpenNLP tagger
  - Abner tagger
  - Amazon Comprehend
  - …



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tagger Nodes

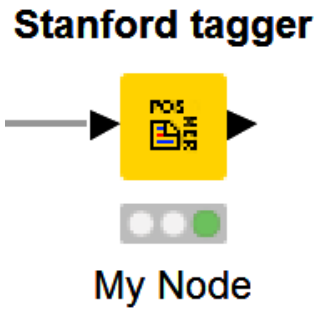
- Allows to specify the number of parallel threads.
- Note: each thread will load a separate model into memory!
- Tagged terms are set unmodifiable (we'll explain this shortly).



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

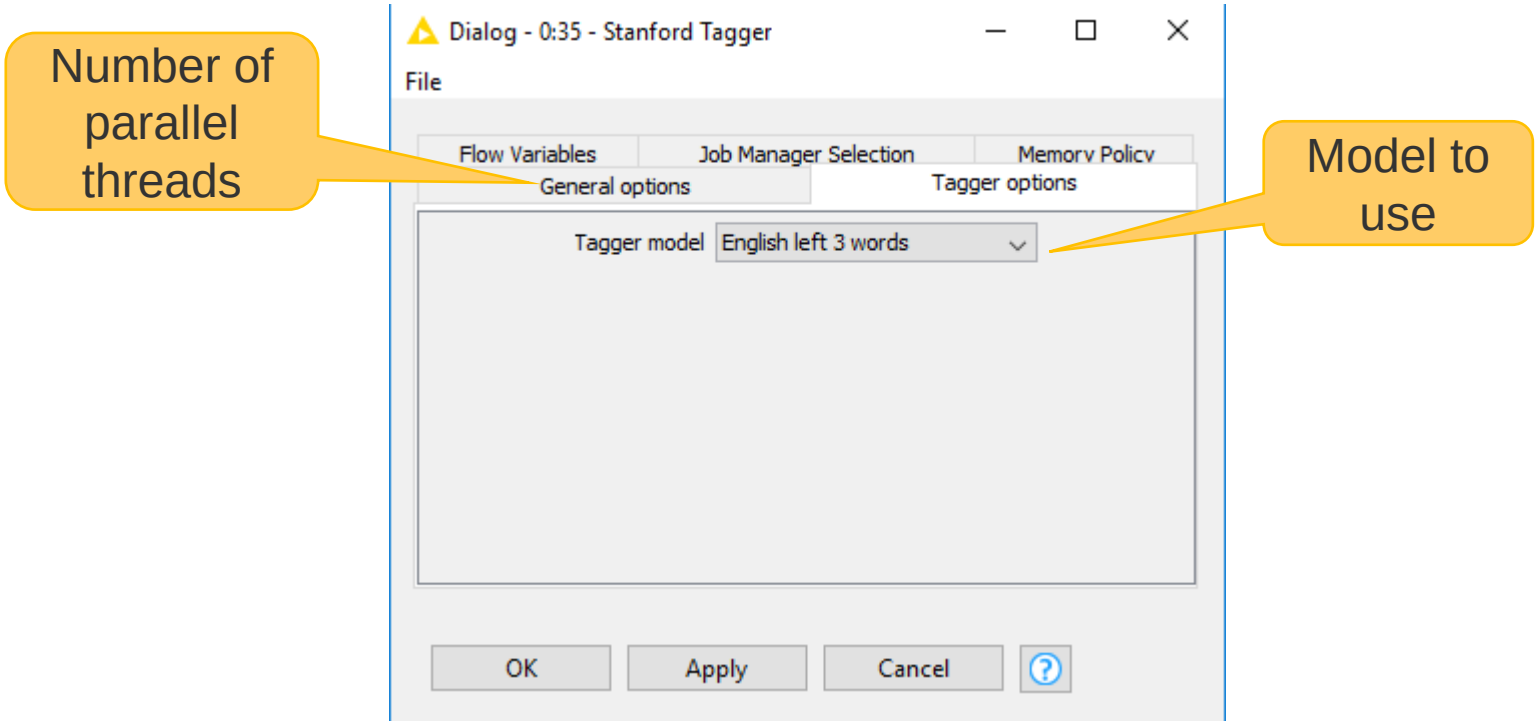
New Node: Stanford tagger

- Assigns part of speech tags to terms
  - Models for English, German, French (from Stanford NLP Group)
  - Alternative node: POS tagger
    - Model only for English (from OpenNLP)



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

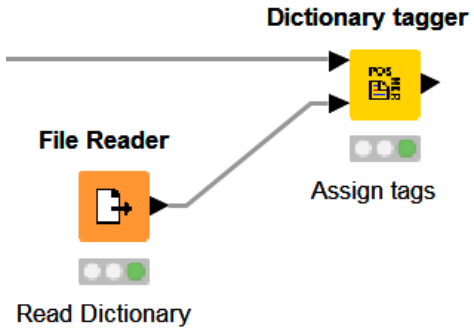
Stanford Tagger: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

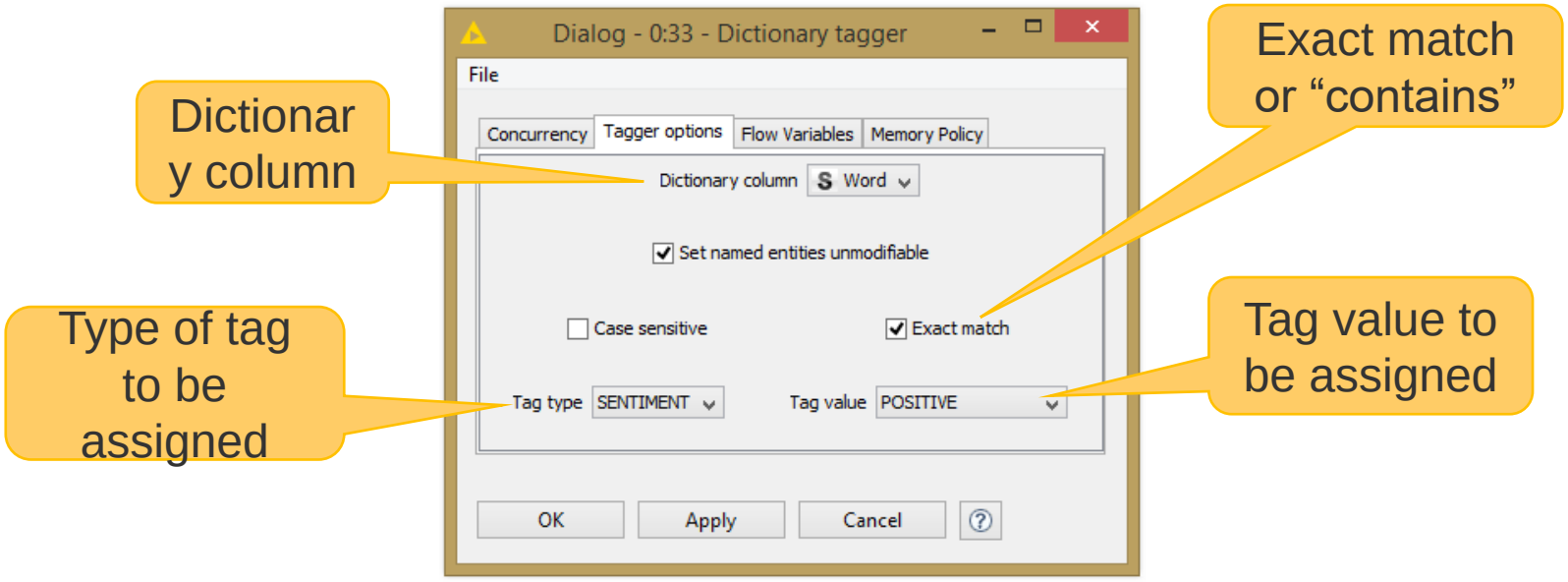
New Node: Dictionary Tagger

- Assigns selected tag to matching terms
  - Matches terms in documents against terms in dictionary
  - Tag to be assigned to matching terms is specified in the dialog
  - Alternative node: Wildcard tagger
    - Terms in dictionary may contain wild cards and regular expressions



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Dictionary Tagger: Configuration



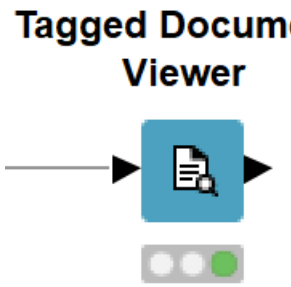
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Tagged Document Viewer

- Displays documents with tags highlighted:
  - Takes a column with documents as input
  - Allows to inspect tags assigned to documents

Document column

Row ID	Document
Row…	"Italian With Pizzazz"
Row…	"Who Doesn't like Italian?"
Row…	"Great Italian Food Served by Italian People"
Row…	"Fantastic Italian"
Row…	"Outstanding Italian!"



Document with tags highlighted

Italian With Pizzazz

ItalianWithPizzazz

I'm a particular eater. I tried because of the great reviews and was not disappointed. The food literally melted in my mouth and I may have said "wow" out loud.

The energy is good & the Italian waiters were too. They all wore a inspiring tee and were cheery and helpful. It's very popular and was quite loud with lots...

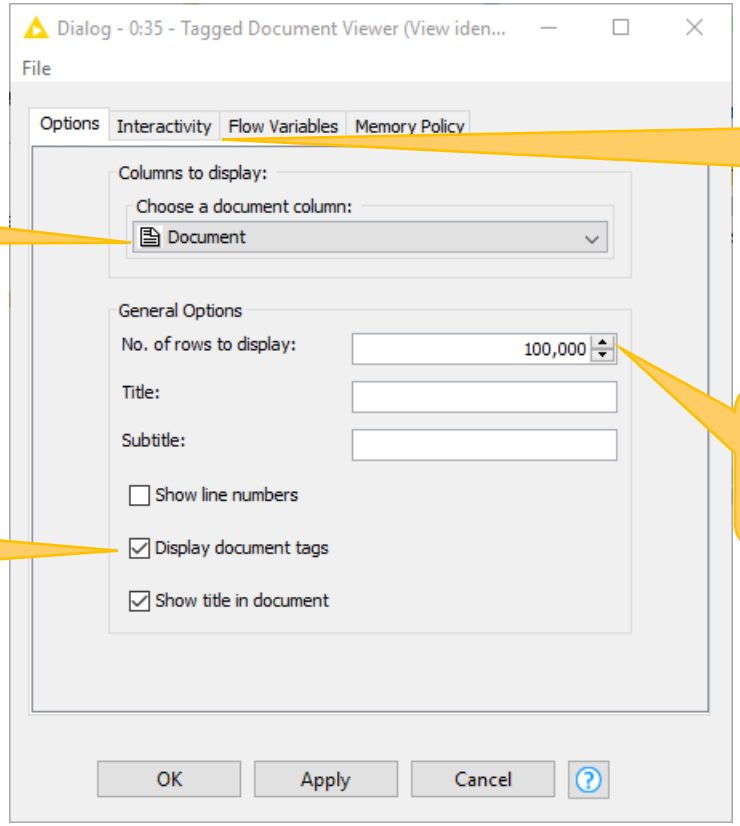
More

Showing 1 to 1 of 594 entries

Previous 1 2 3 4 5 … 594 Next

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tagged Document Viewer: Configuration



Document column

View and interactivity configuration

Enable display of tags

Number of documents to display

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Exercise

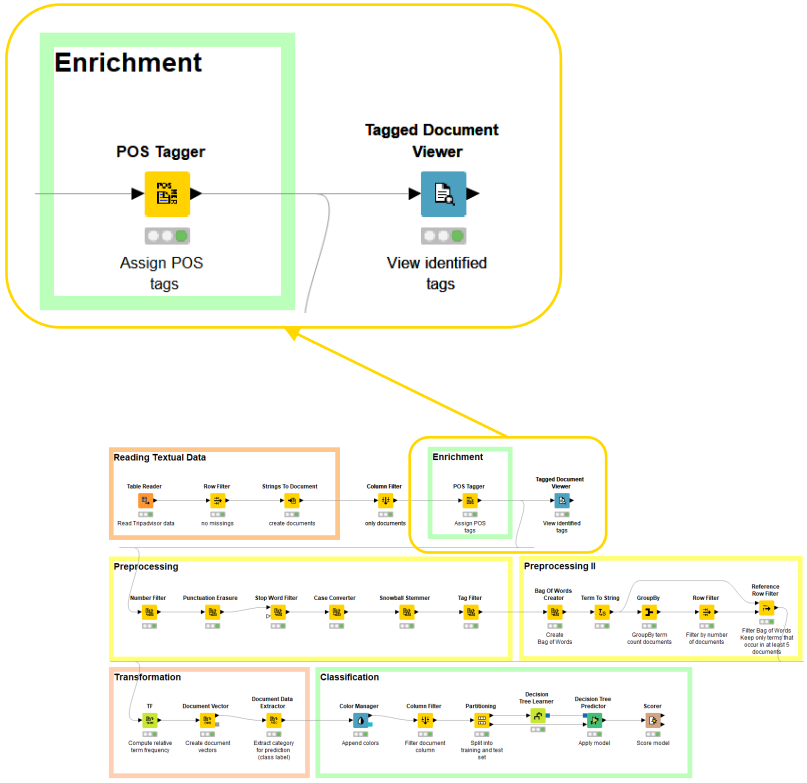
- Start with “Exercise: Enrichment”
  - Assign (English) POS tags
  - View tagged documents

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Solution

Enrichment

- POS tagger
- Tagged Document Viewer



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

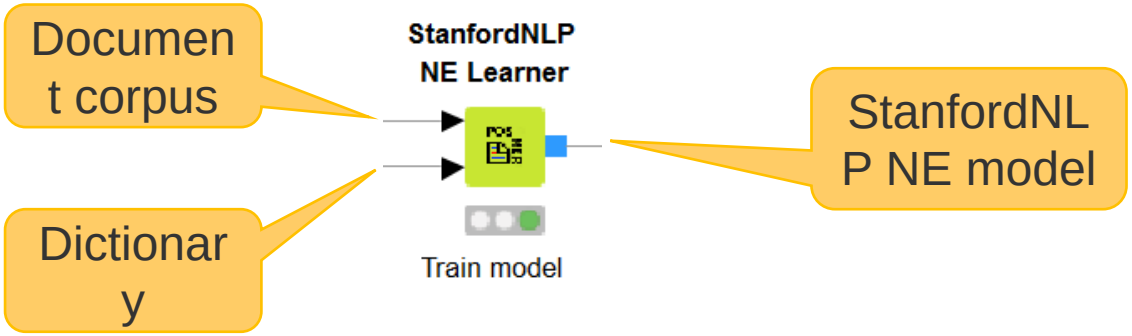
Custom NER models

- The provided NER models of *OpenNLP NE tagger* and *StandfordNLP NE tagger* are trained for a few types of entities and English language only.
- For more specific applications and other languages custom models are needed.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

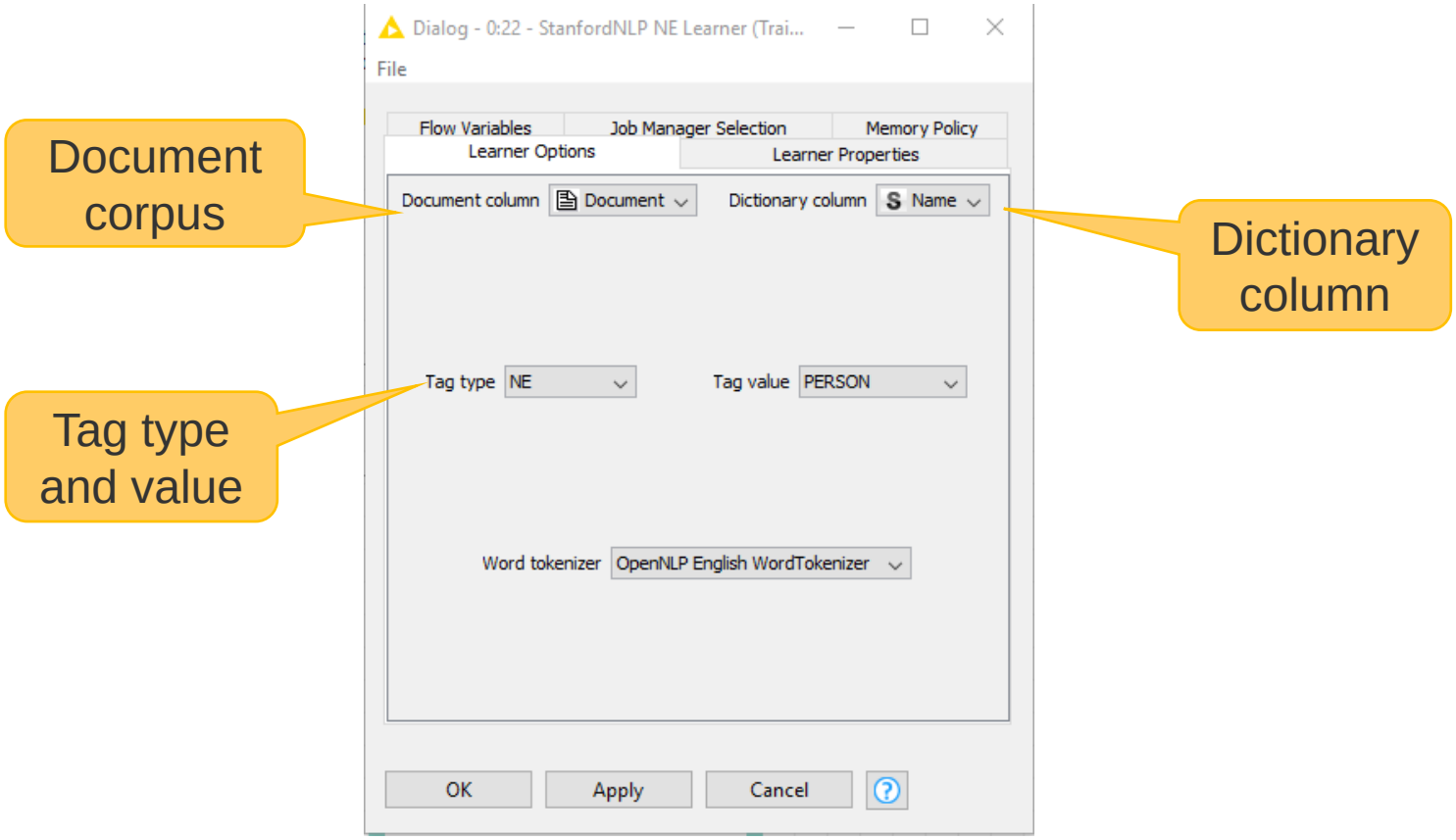
New Node: StanfordNLP NE Learner

- Trains a NER model based on the input dictionary and corpus
  - Tag type and value can be set in the dialog
  - Creates tagged corpus based on input documents and dictionary. Trains model with tagged corpus.



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

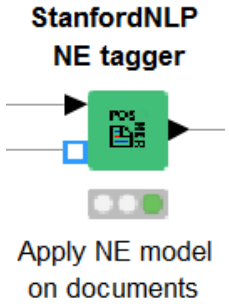
Stanford Tagger: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

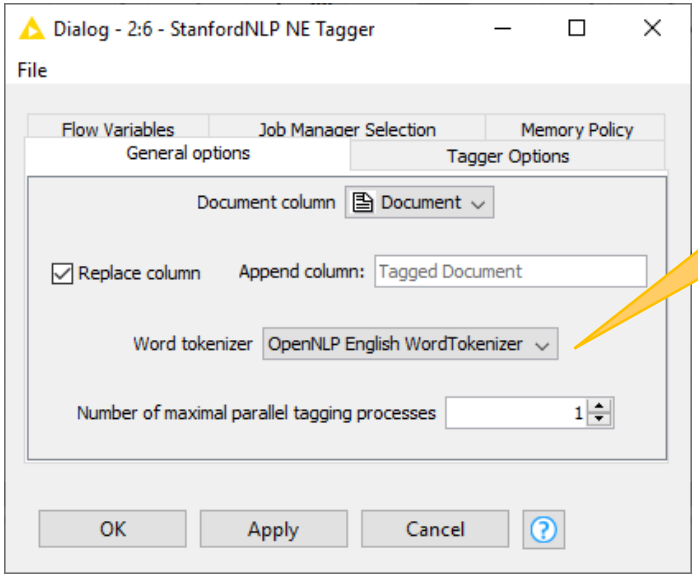
New Node: StanfordNLP NE tagger

- Tags documents based on input NER model.
  - NER model can be specified in dialog, built-in or model from input port



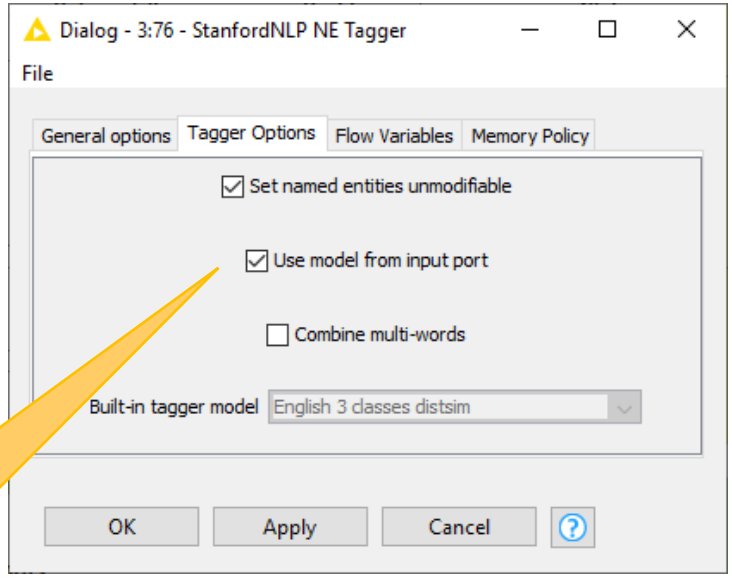
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

StanfordNLP NE tagger: Configuration



Parameters for built-in model

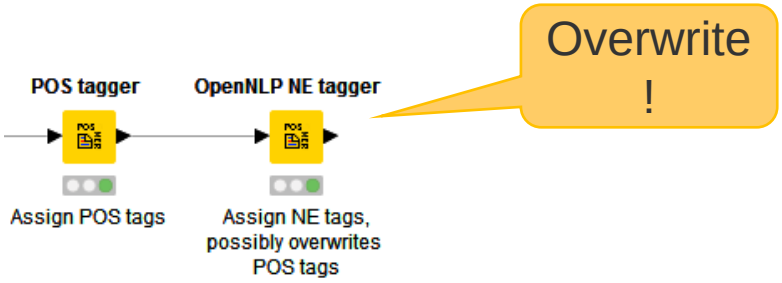
Use model from input port or built-in models



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tagging Conflicts

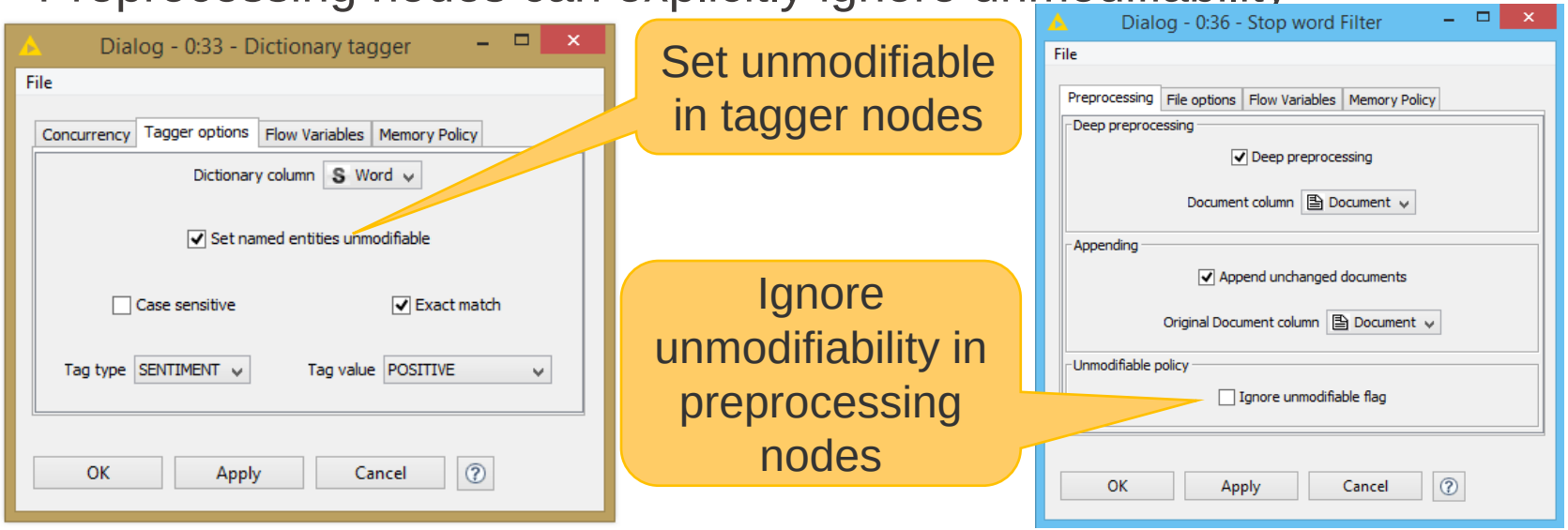
- In case of tag intersections, the last node overwrites.
- “Serbian-American inventor Nikola Tesla developed the …”
  1. POS tagger: “Serbian-American\NNP inventor\NNP Nikola\NNP Tesla\NNP developed\VBD the\DT…”
  2. NE tagger: “Serbian-American\NNP inventor\NNP Nikola Tesla\Person developed\VBD the\DT …”



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Unmodifiable Terms

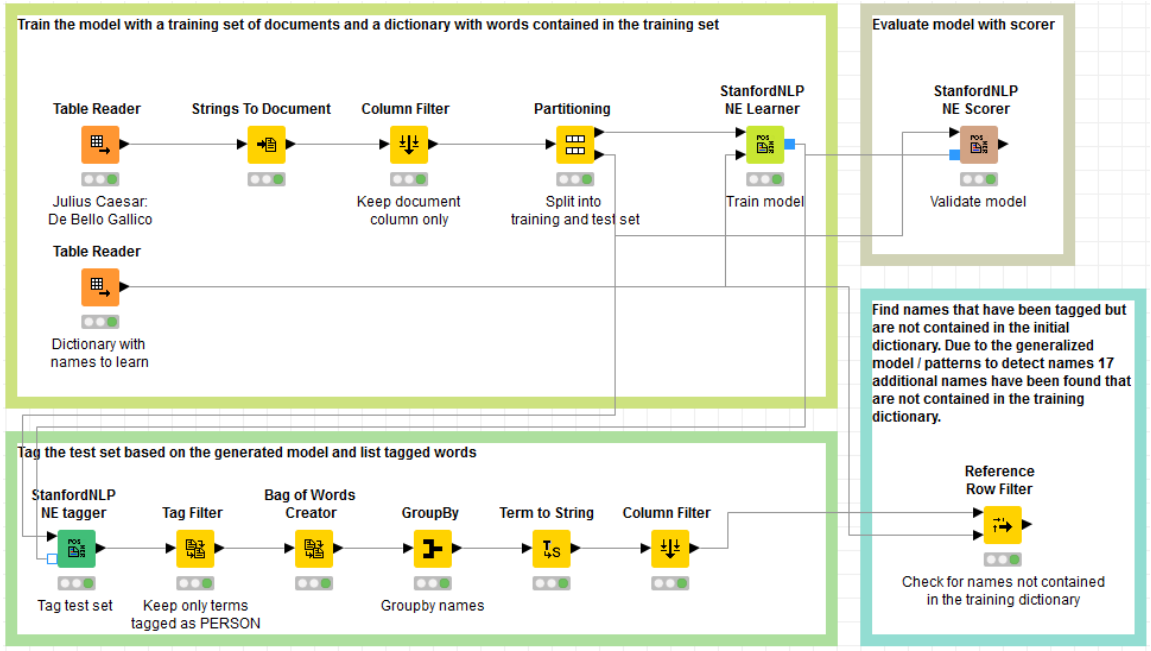
- Tagged terms can be set unmodifiable
- Unmodifiable terms are not affected by any preprocessing node
- Preprocessing nodes can explicitly ignore unmodifiability



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Supplementary Workflows: NER Tagger Model Training

- Trains NER model for latin and gallic names based on “De Bello Gallico” from Julius Caesar.



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Preprocessing

Preprocessing

- Reduction of feature space (terms)
- Filtering of unnecessary terms
  - Stop words, based on POS tags, dictionaries, RegEx, …
- Normalization of terms
  - Stemming, case conversion

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Preprocessing Nodes

- Typically characterized by:
  - Yellow color
  - 1 to 2 input ports (requiring one document column), 1 output port
  - For filtering and normalizing terms of documents and bags of words

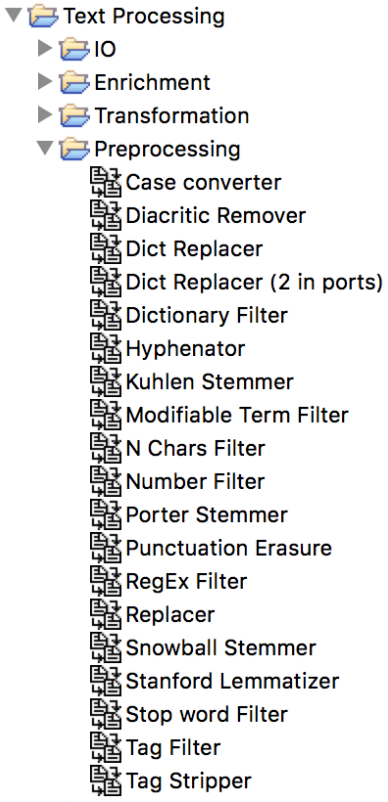
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Preprocessing Nodes

- Node Repository:

  Other Data Types/Text Processing/Preprocessing

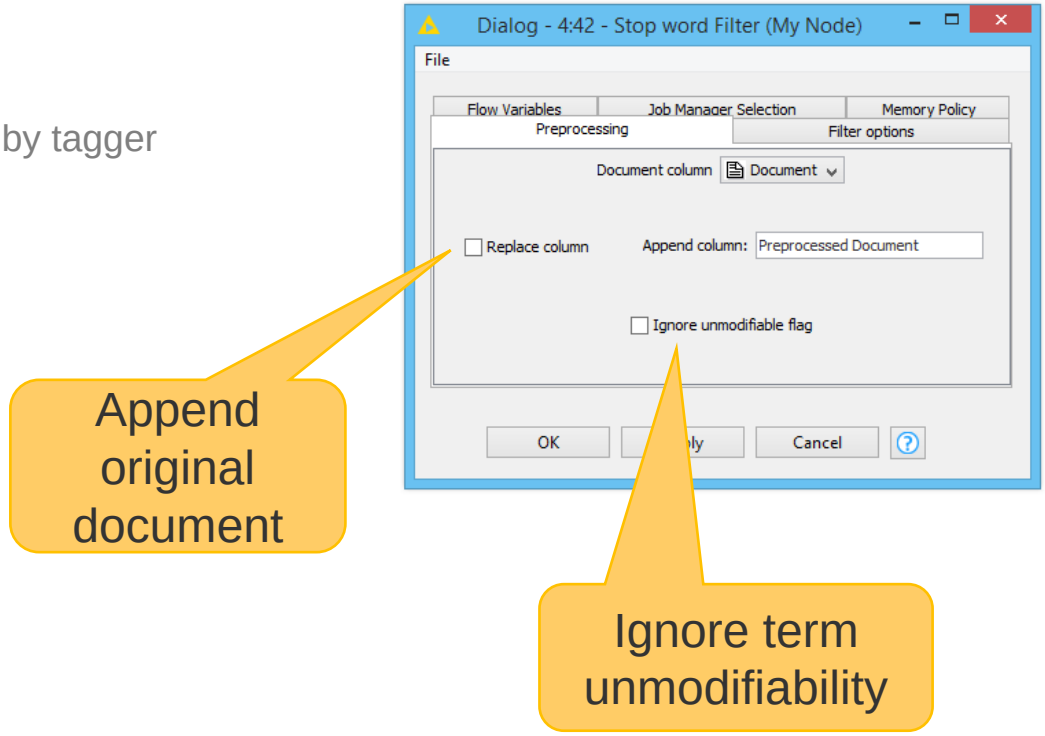
- Available Preprocessing Nodes
  - Stop Word Filter
  - Snowball Stemmer
  - Tag Filter
  - Case Converter
  - RegEx Filter
  - …



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Preprocessing Nodes

- Preprocessing tab in node dialog to specify:
  - Append original documents
  - Ignore term unmodifiability (set by tagger nodes).



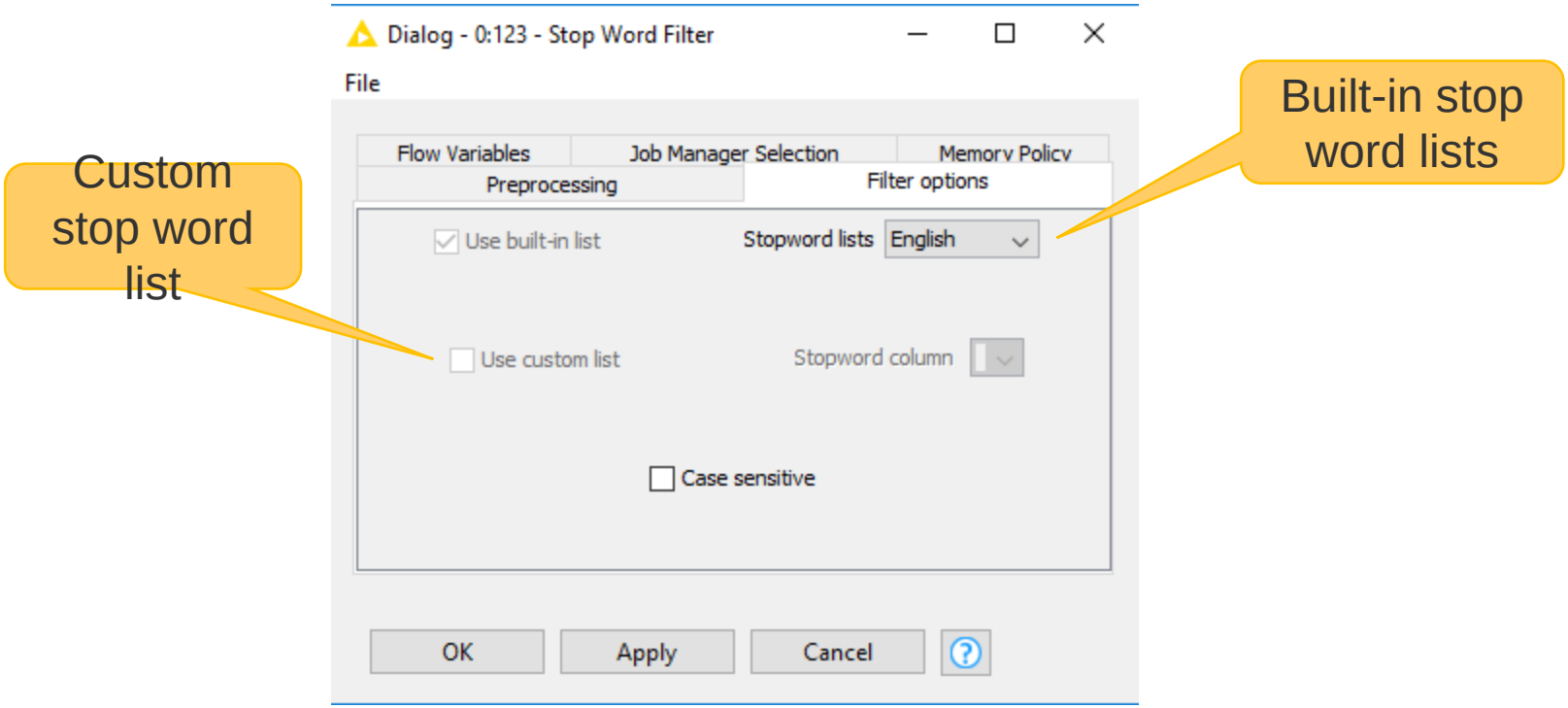
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Stop Word Filter

- Filters stop words
  - Built-in stop word lists: English, French, German, Italian, …
  - Alternatively load custom stop word list

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Stop Word Filter: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Stemming

Stemming reduces different forms of a word to its common base by sequential application of stemming rules.

Rule	Example
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

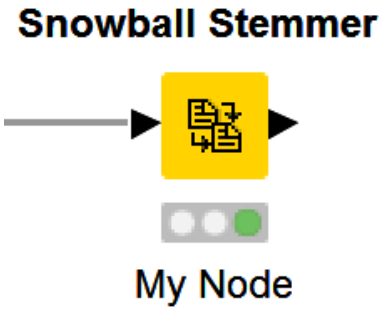
Original text:
Light caresses colours, sets them aglow, plays with nuances, shadows and structures

Porter stemmer:
Light caress colour, set them aglow, plai with nuanc, shadow and structure.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

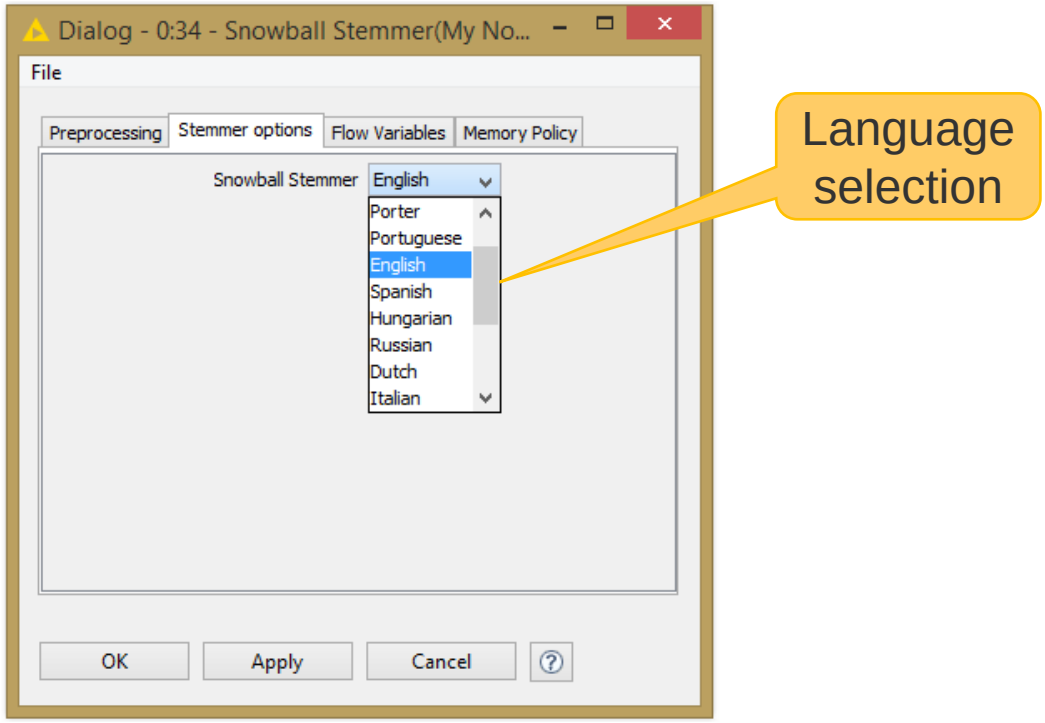
New Node: Snowball Stemmer

- Reduces terms to word stem
  - For various languages: English, German, French, Italian, …
  - Integration of Snowball stemming library
  - Alternative nodes: Porter Stemmer, Kuhlen Stemmer
    - For English only



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

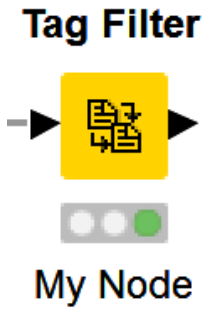
Snowball Stemmer: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

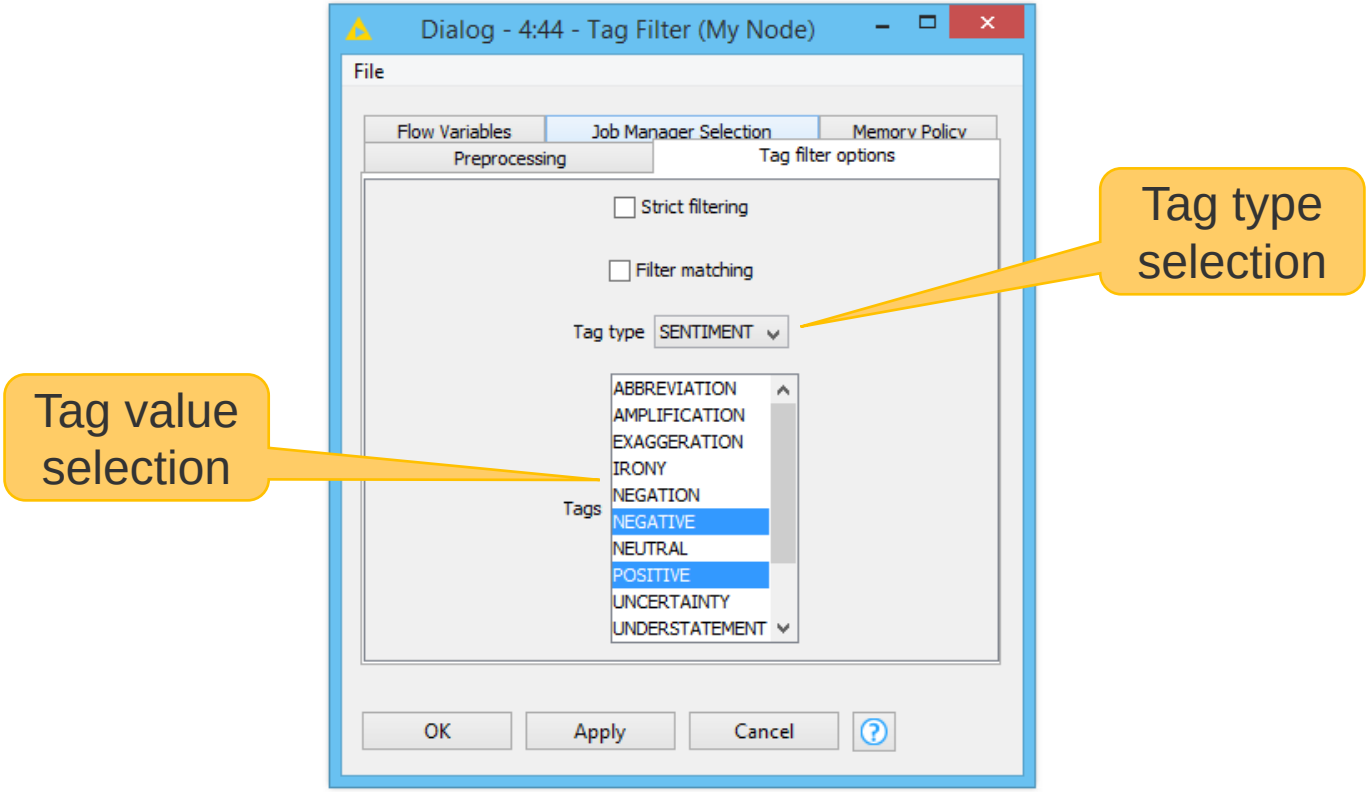
New Node: Tag Filter

- Filters terms based on specified tag values
  - For all tag types and values



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Tag Filter: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Exercise

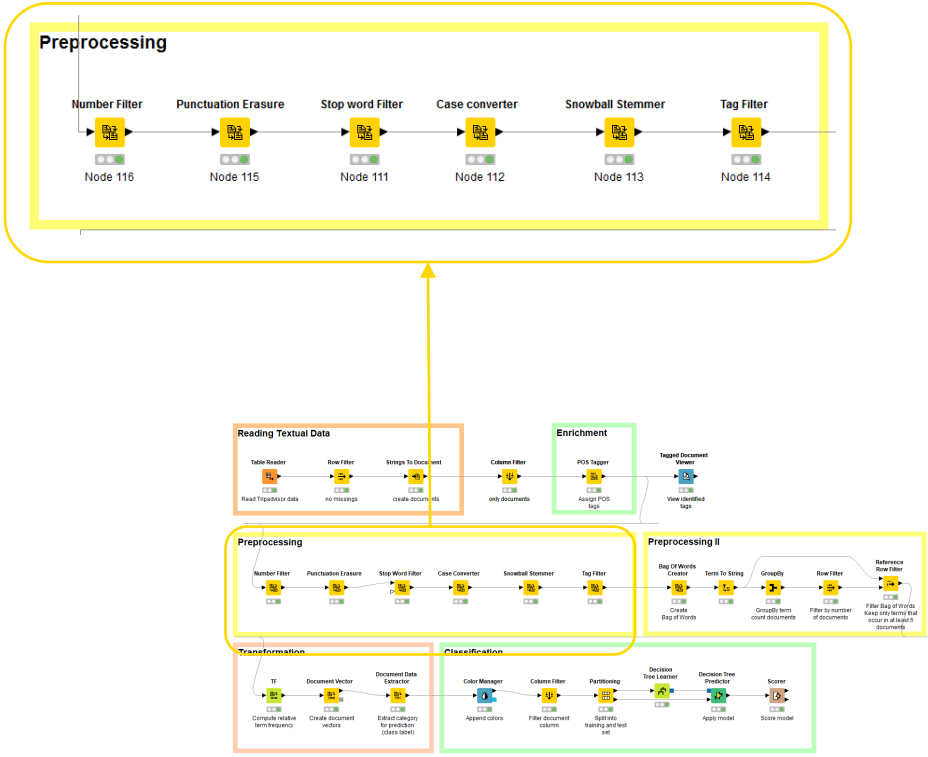
- Start with “Exercise: Preprocessing”
  - Filtering:
    - Numbers
    - Punctuation marks
    - Stop words
    - All terms except: nouns, verbs, adjectives
  - Stemming
  - To lower case

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Solution

Preprocessing

- Number Filter
- Punctuation Erasure
- Stop Word Filter
- Case Converter
- Snowball Stemmer
- POS Filter



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Transformation

Transformation

- Transformation of data table structures
  - List of documents ➔ bag of words
  - Bag of words ➔ document / term **vectors**
  - Extraction of document fields to string columns
  - Conversion of terms to strings

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Transformation Nodes

- Typically characterized by:
  - Yellow color
  - 1 input port, 1 output port

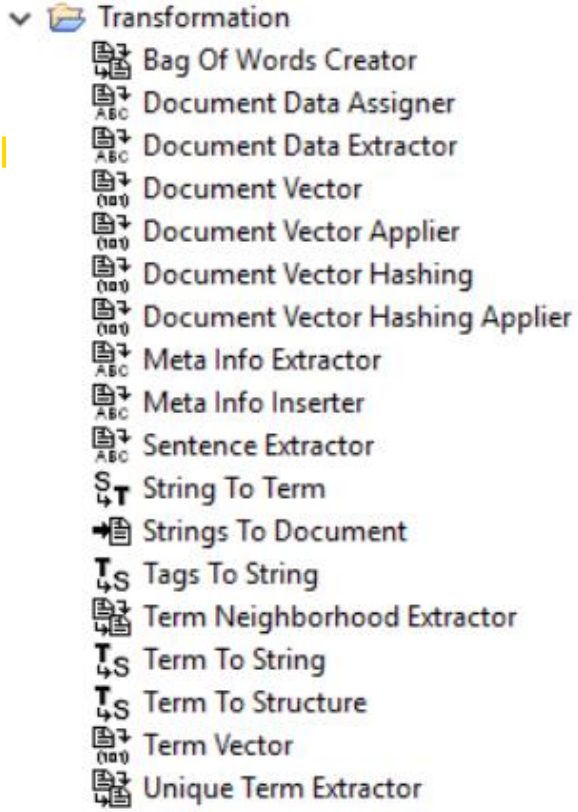
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Transformation Nodes

- Node Repository:

  Other Data Types/Text Processing/Transfor

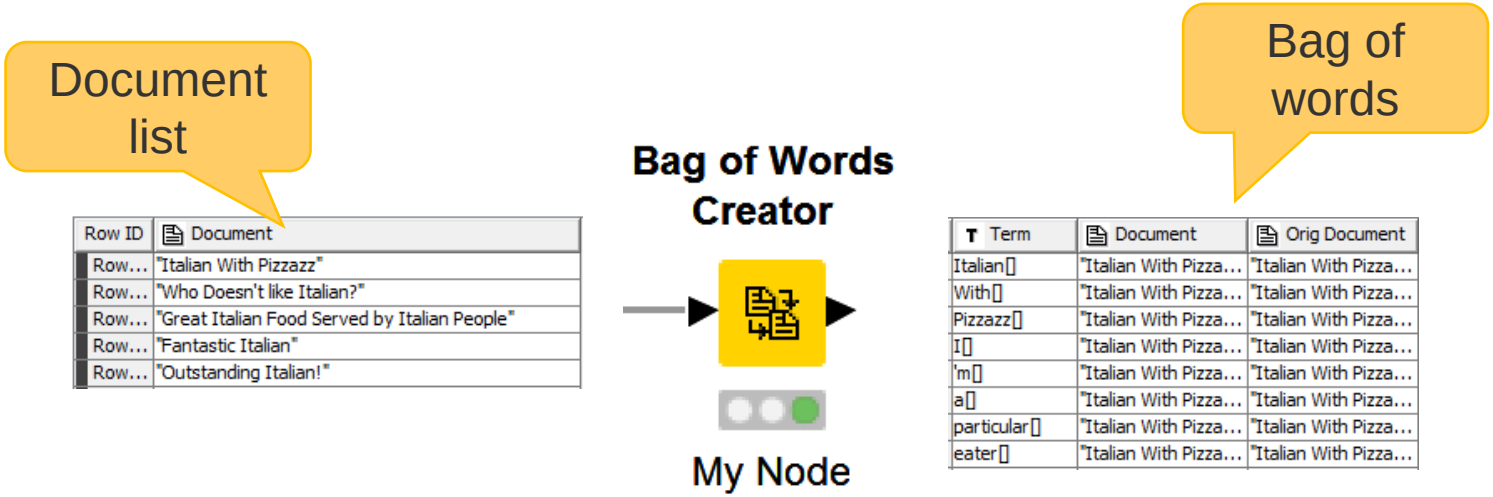
- Available Transformation Nodes
  - Bag of Words Creator
  - Document Vector
  - Strings to Document
  - Sentence Extractor
  - Document Data Extractor
  - Unique Term Extractor
  - …



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Bag of Words Creator

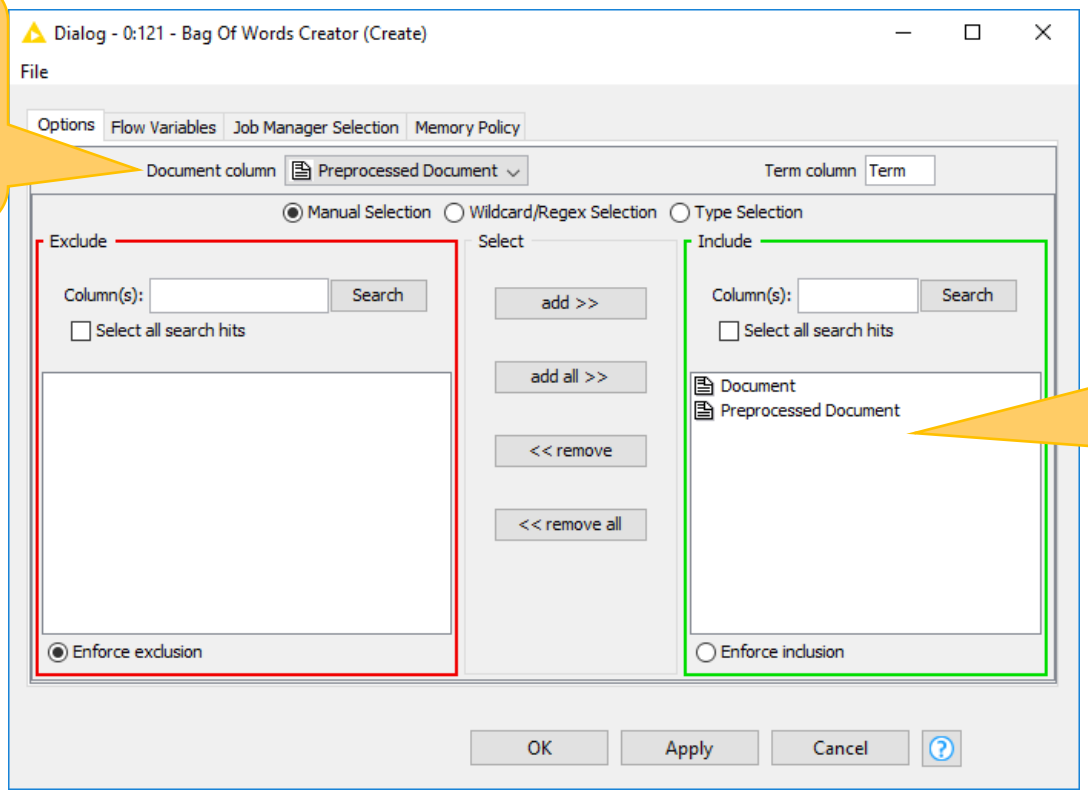
- Transforms list of documents into bag of words
  - Original documents can be appended in a column



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Bag of Words Creator: Configuration

Documents used to create bag of words



Original documents can be appended

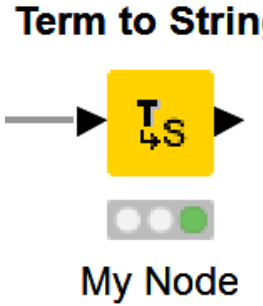
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Term to String

- Transforms term cells into string cells
  - Tag information will get lost

Bag of words

Term	Document	Orig Document
italian[JJ(POS)]	"italian pizza...	"Italian With Pizza...
pizzazz[NNP(POS)]	"italian pizza...	"Italian With Pizza...
particular[JJ(POS)]	"italian pizza...	"Italian With Pizza...
tri[VBN(POS)]	"italian pizza...	"Italian With Pizza...
review[NNS(POS)]	"italian pizza...	"Italian With Pizza...
disappoint[JJ(POS)]	"italian pizza...	"Italian With Pizza...
food[NN(POS)]	"italian pizza...	"Italian With Pizza...

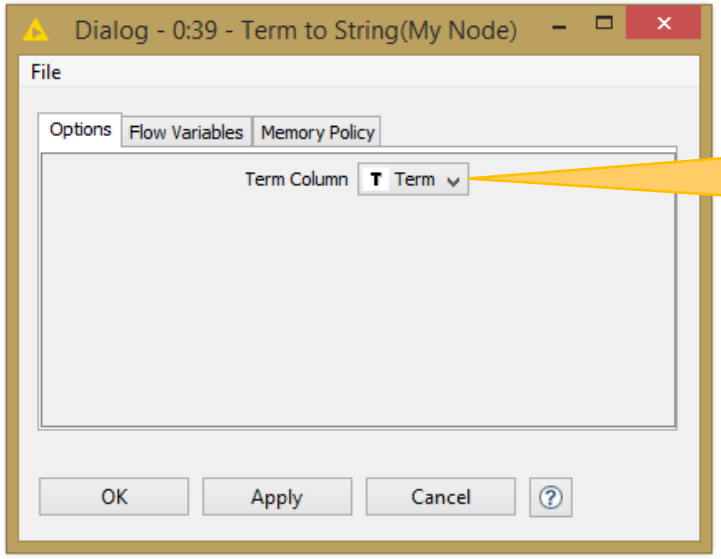


Bag of words with string column

Term	Document	Orig Document	Term a...
italian[JJ(PO...	"italian pizza...	"Italian With Pizza...	italian
pizzazz[NNP...	"italian pizza...	"Italian With Pizza...	pizzazz
particular[JJ...	"italian pizza...	"Italian With Pizza...	particular
tri[VBN(POS)]	"italian pizza...	"Italian With Pizza...	tri
review[NNS(...	"italian pizza...	"Italian With Pizza...	review
disappoint[J...	"italian pizza...	"Italian With Pizza...	disappoint
food[NN(POS)]	"italian pizza...	"Italian With Pizza...	food

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Term to String: Configuration



Terms to transform to strings

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Exercise

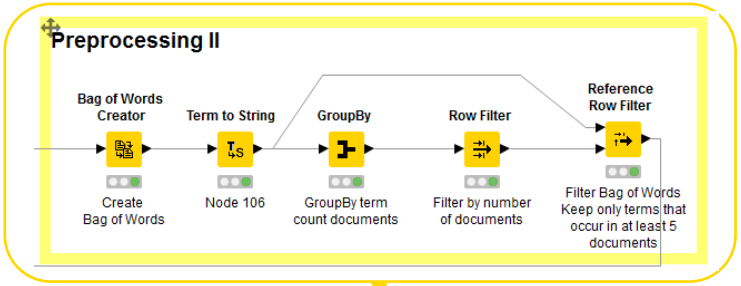
- Start with “Exercise: Preprocessing II”
  - Create bag of words
  - Filter terms that occur in less than 5 documents
  - Tip: Bag of Words, GroupBy, and Reference Row Filter

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Solution

Preprocessing II

- Bag of Words Creator
- Term to String
- GroupBy
- Row Filter
- Reference Row Filter



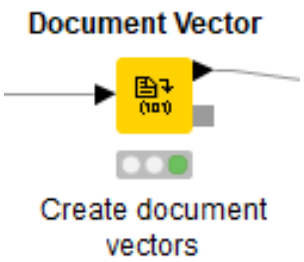
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Document Vector

- Transforms bag of words into document vectors
  - Creates bit or numerical vectors

Bag of words with frequency column

Term	Document	Orig Document	TF rel
Italian[]	"Italian With Pizza...	"Italian With Pizza...	0.027
With[]	"Italian With Pizza...	"Italian With Pizza...	0.014
Pizzazz[]	"Italian With Pizza...	"Italian With Pizza...	0.014
I[]	"Italian With Pizza...	"Italian With Pizza...	0.041
'm[]	"Italian With Pizza...	"Italian With Pizza...	0.014
a[]	"Italian With Pizza...	"Italian With Pizza...	0.027
particular[]	"Italian With Pizza...	"Italian With Pizza...	0.014
eater[]	"Italian With Pizza...	"Italian With Pizza...	0.014

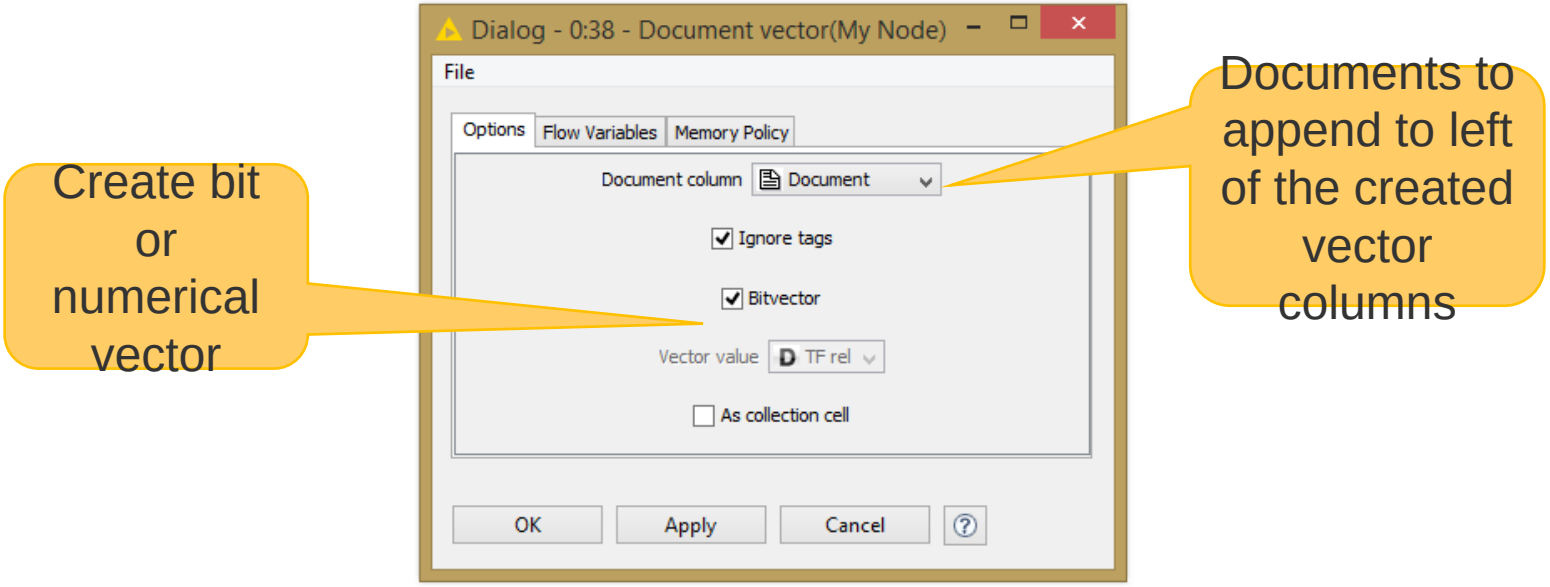


Document vector

Document	dim	sum	wife
""Dim Sum No Mor...	1	1	1
""Hole in the wall,...	1	0	0
""Order of the Ca...	0	0	0
"16th birthday su...	0	0	0
"25 years betwee...	0	0	0
"3 dishes, 3 times...	0	0	1

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Document Vector: Configuration



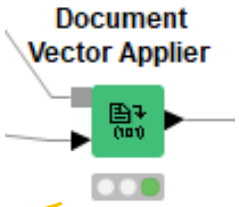
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Document Vector Applier

- Transforms bag of words into document vectors
  - Creates feature space of reference document vectors
  - Creates bit or numerical vectors

Document vector

Term	Document	Orig Document	TF rel
Italian[]	"Italian With Pizza...	"Italian With Pizza...	0.027
With[]	"Italian With Pizza...	"Italian With Pizza...	0.014
Pizzazz[]	"Italian With Pizza...	"Italian With Pizza...	0.014
I[]	"Italian With Pizza...	"Italian With Pizza...	0.041
'm[]	"Italian With Pizza...	"Italian With Pizza...	0.014
a[]	"Italian With Pizza...	"Italian With Pizza...	0.027
particular[]	"Italian With Pizza...	"Italian With Pizza...	0.014
eater[]	"Italian With Pizza...	"Italian With Pizza...	0.014

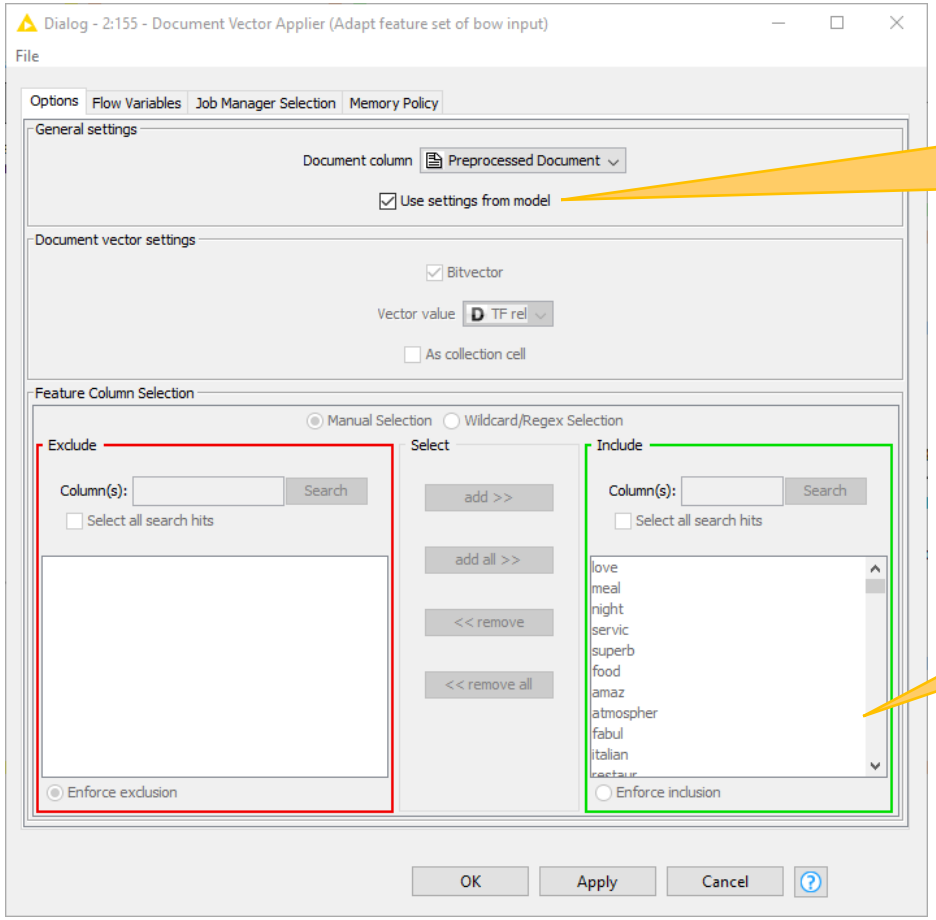


Document	dim	sum	wife
""Dim Sum No Mor...	1	1	1
""Hole in the wall,...	1	0	0
""Order of the Ca...	0	0	0
"16th birthday su...	0	0	0
"25 years betwee...	0	0	0
"3 dishes, 3 times...	0	0	1

Reference document vectors

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Document Vector Applier: Configuration



Use settings from model input

Include and exclude lists of features of the reference vectors

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Document Vector Hashing

- Transforms documents into document vectors
  - Vector indices of terms are determined by term hashing
  - Creates bit or numerical vectors
  - Is streamable

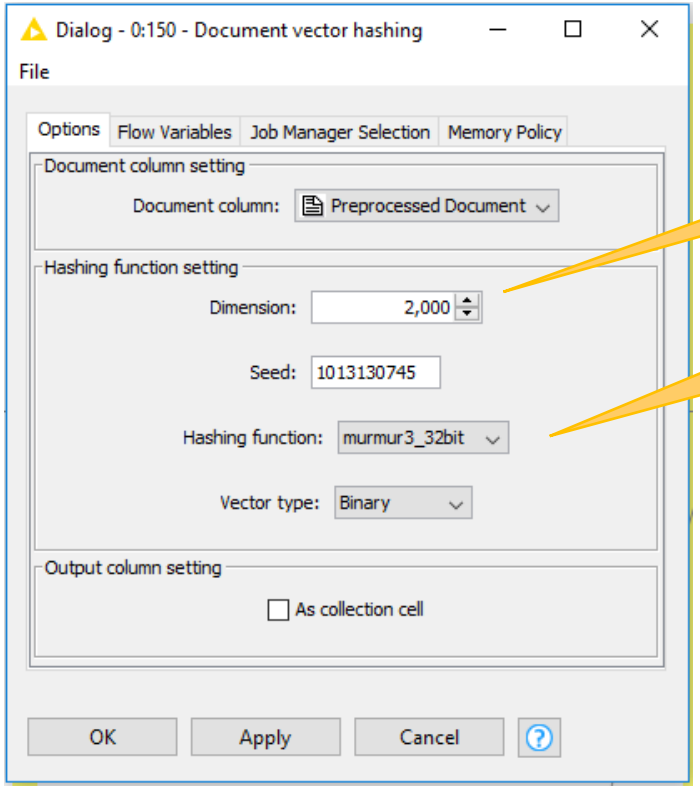
Row ID	Document
Row...	"Italian With Pizzazz"
Row...	"Who Doesn't like Italian?"
Row...	"Great Italian Food Served by Italian People"
Row...	"Fantastic Italian"
Row...	"Outstanding Italian!"

Hashed document vector

Row ID	Document	Preproc...	Col0	Col1	Col2	Col3
Row0_1_Row...	"Dreaming of this piz...	"dream pizza"	0	0	0	0
Row0_2_Row...	"amazing pizza"	"amaz pizza"	0	0	0	0
Row0_3_Row...	"Wow! Simply aweso...	"wow simpli ...	0	0	0	0
Row0_4_Row...	"Great Pizza...."	"pizza"	0	0	0	0
Row0_5_Row...	"wow"	"wow"	0	0	0	0
Row0_6_Row...	"Best pizza we ever ...	"pizza"	0	0	0	0
Row0_7_Row...	"Incredible flavor, fa...	"incred flavo...	0	0	0	0

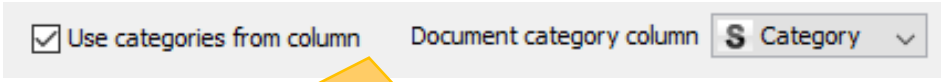
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Document Vector Hashing: Configuration



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Document Data Extractor

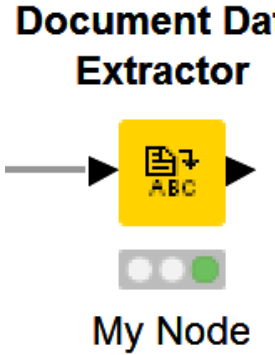


- Extracts document fields as strings
  - Title, text, categories, …

Reminder: we stored restaurant type into Category field in String to Document conversion

Document column

Row ID	Document
Row…	"Italian With Pizzazz"
Row…	"Who Doesn't like Italian?"
Row…	"Great Italian Food Served by Italian People"
Row…	"Fantastic Italian"
Row…	"Outstanding Italian!"



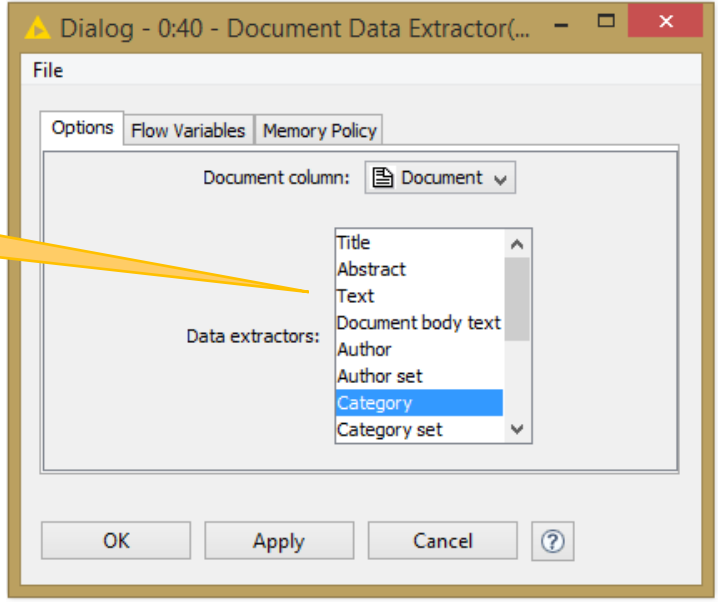
Extracted field as string column

Document	Category
"Italian With Pizzazz"	Italian
"Who Doesn't like Italian?"	Italian
"Great Italian Food Served by Italian People"	Italian
"Fantastic Italian"	Italian
"Outstanding Italian!"	Italian

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Document Data Extractor: Configuration

Fields to extract



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Frequencies

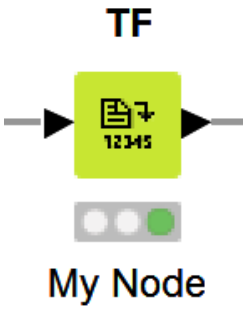
- Frequencies are based on the number of occurrences of terms
  - Locally (in documents): term frequency (TF) absolute or relative
  - Globally (in corpus): inverse document frequency (IDF)
- Frequencies can also be used for term filtering

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Frequency Nodes

- Typically characterized by:
  - Green color
  - 1 input port, 1 output port
  - Require bag of words

Term	Document	Orig Document
Italian[]	"Italian With Pizza...	"Italian With Pizza...
With[]	"Italian With Pizza...	"Italian With Pizza...
Pizzazz[]	"Italian With Pizza...	"Italian With Pizza...
I[]	"Italian With Pizza...	"Italian With Pizza...
'm[]	"Italian With Pizza...	"Italian With Pizza...
a[]	"Italian With Pizza...	"Italian With Pizza...
particular[]	"Italian With Pizza...	"Italian With Pizza...
eater[]	"Italian With Pizza...	"Italian With Pizza...



Append column with relative TF values

Term	Document	Orig Document	TF rel
Italian[]	"Italian With Pizza...	"Italian With Pizza...	0.027
With[]	"Italian With Pizza...	"Italian With Pizza...	0.014
Pizzazz[]	"Italian With Pizza...	"Italian With Pizza...	0.014
I[]	"Italian With Pizza...	"Italian With Pizza...	0.041
'm[]	"Italian With Pizza...	"Italian With Pizza...	0.014
a[]	"Italian With Pizza...	"Italian With Pizza...	0.027
particular[]	"Italian With Pizza...	"Italian With Pizza...	0.014
eater[]	"Italian With Pizza...	"Italian With Pizza...	0.014

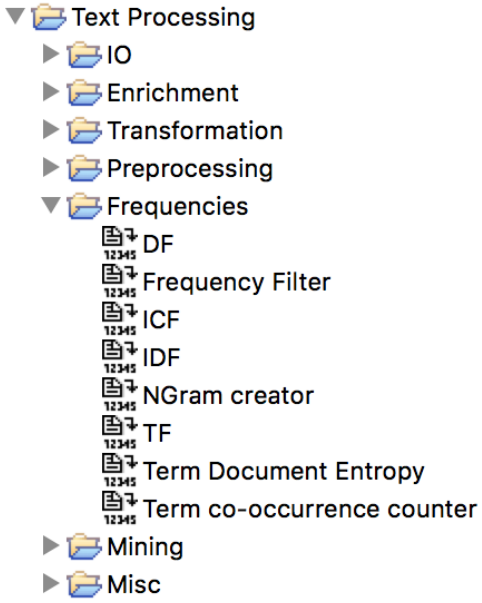
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Frequency Nodes

- Node Repository:

  Other Data Types/Text Processing/Frequencies

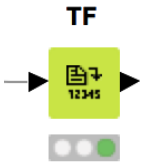
- Available Frequency Nodes
  - TF
  - IDF
  - NGram creator
  - …



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: TF

- Computes the relative or absolute term frequency (TF) of each term within a document



Terms and documents output table - 3:119 - TF

File

Table "default" - Rows: 9135 | Spec - Columns: 5 | Properties | Flow Variables

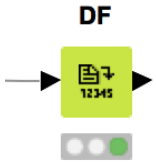
Row ID	Term	Document	Orig Document	Term ...	TF rel
Row0	italian[JJ(PO...	"italian pizza...	"Italian With Pizza...	italian	0.1
Row3	tri[VBN(POS)]	"italian pizza...	"Italian With Pizza...	tri	0.05
Row4	review[NNS...	"italian pizza...	"Italian With Pizza...	review	0.05
Row5	disappoint[...	"italian pizza...	"Italian With Pizza...	disappoint	0.05
Row6	food[NN(P...	"italian pizza...	"Italian With Pizza...	food	0.05
Row7	melt[VBN(P...	"italian pizza...	"Italian With Pizza...	melt	0.05
Row8	mouth[NN(...	"italian pizza...	"Italian With Pizza...	mouth	0.05
Row9	wow[VB(POS)]	"italian pizza...	"Italian With Pizza...	wow	0.05
Row11	italian[NNP(...	"italian pizza...	"Italian With Pizza...	italian	0.1
Row12	waiter[NNS(...	"italian pizza...	"Italian With Pizza...	waiter	0.05
Row17	help[JJ(POS)]	"italian pizza...	"Italian With Pizza...	help	0.05
Row18	popular[JJ(...	"italian pizza...	"Italian With Pizza...	popular	0.05
Row19	lot[NNS(POS)]	"italian pizza...	"Italian With Pizza...	lot	0.05
Row20	italian[NNP(...	"italian"	"Who Doesn't like I...	italian	0.105
Row21	food[NN(P...	"italian"	"Who Doesn't like I...	food	0.053
Row22	wrong[JJ(PO...	"italian"	"Who Doesn't like I...	wrong	0.053
Row24	review[NN(...	"italian"	"Who Doesn't like I...	review	0.053

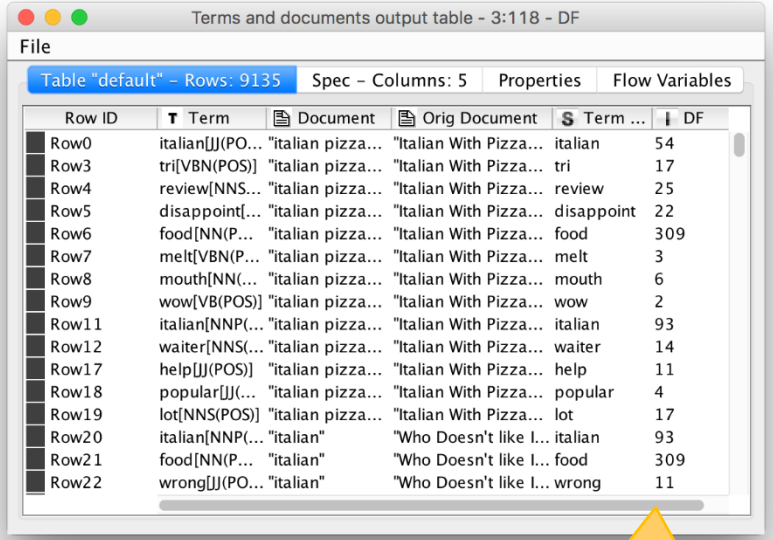
Appended column with TF values

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: DF

- Computes the number of documents that contain each term



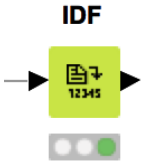


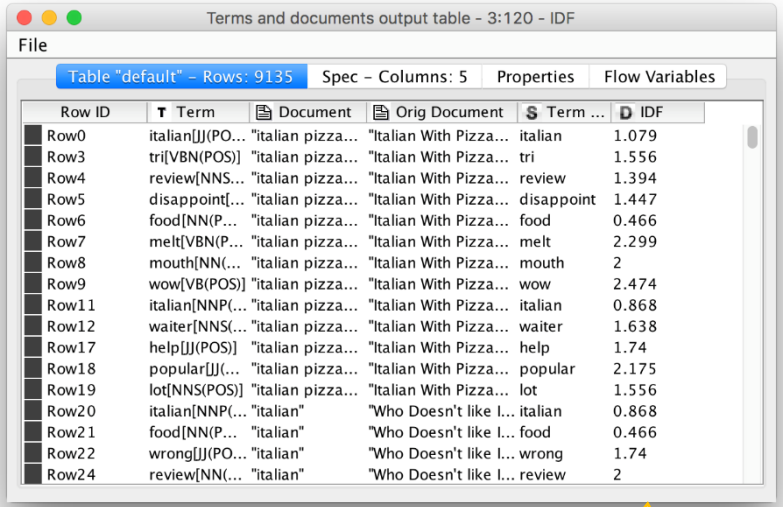
Appended column with DF values

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: IDF

- Computes three variants of inverse document frequency (IDF) for each term within the documents
  - Smooth, normalized, and probabilistic



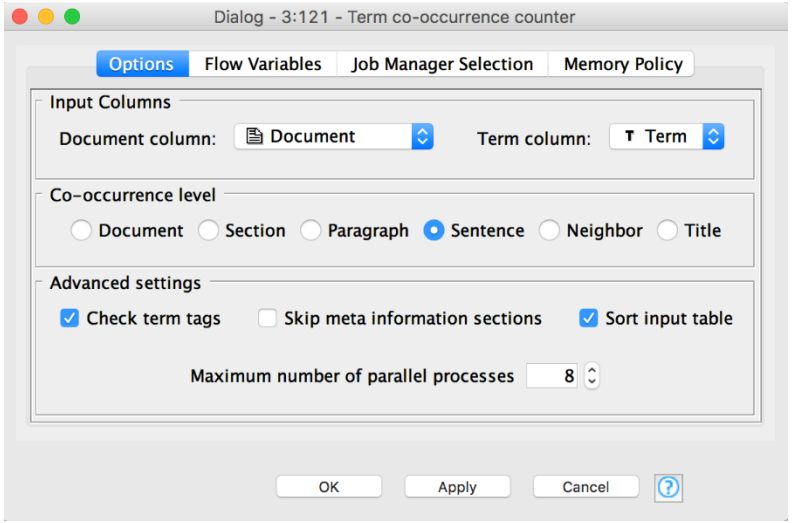


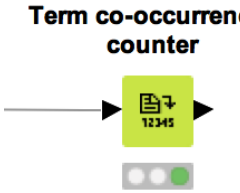
Appended column with IDF values

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Term Co-Occurence Counter

- Counts the number of pairwise co-occurences of terms in bag of words within selected parts of document (e.g. sentence, paragraph, title)

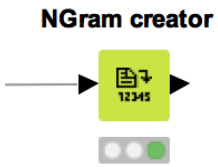


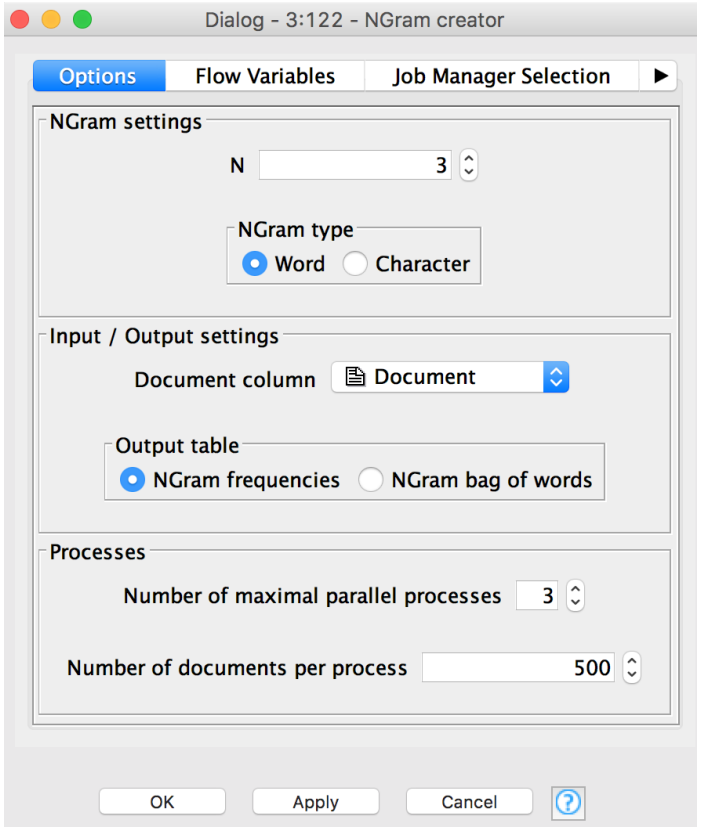


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Ngram Creator

- Creates ngrams from documents of input table and counts their frequencies
- Both word and character ngrams are possible





These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Exercise

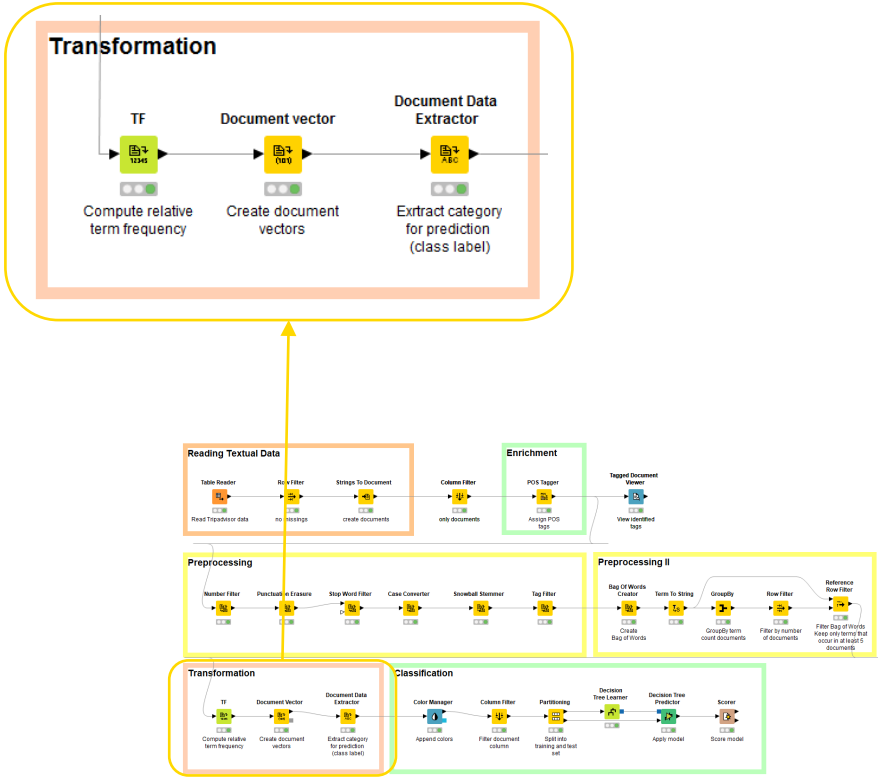
- Start with “Exercise: Transformation”
  - Compute relative term frequencies
  - Create document vectors
  - Extract class label / category

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Solution

Transformation

- TF
- Document Vector
- Document Data Extractor



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Classification

Classification

- Assigning pre-defined labels to documents
  - Categorization
  - Sentiment analysis
  - Topic assignment
- Supervised learning

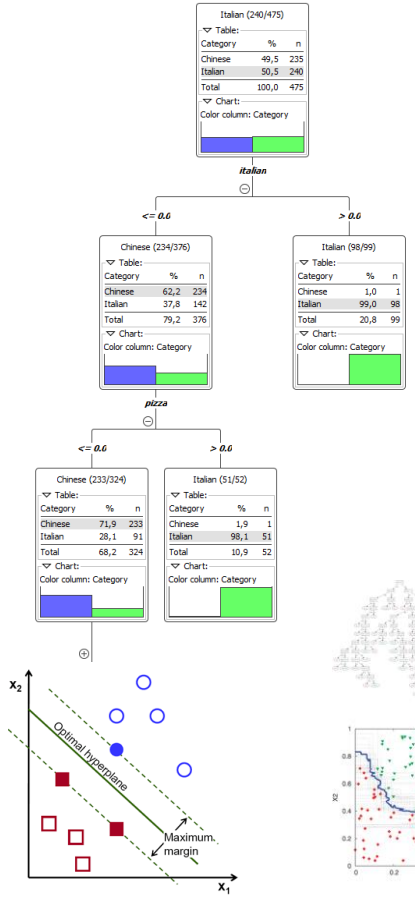
- In the last section we transformed textual documents into a numerical representation (document vectors).
- We can use standard KNIME nodes to classify / analyze these vectors.

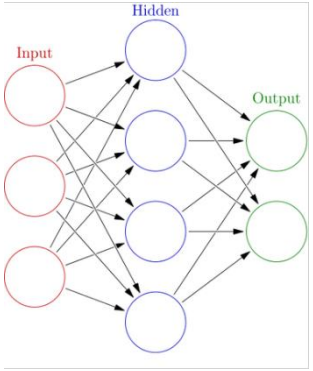
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

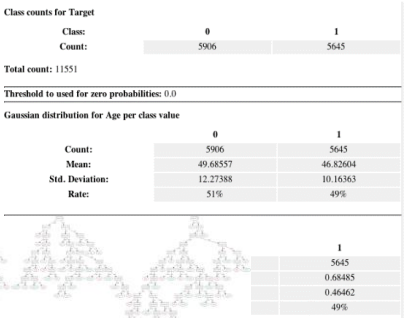
Classification

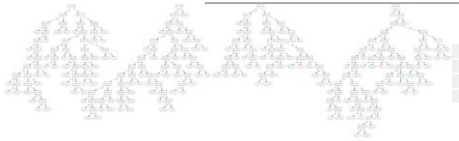
Methods:

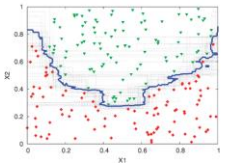
- Decision Trees
- Neural Networks
- Naïve Bayes
- Logistic Regression
- Support Vector Machine
- Tree Ensembles

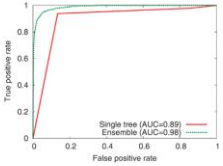






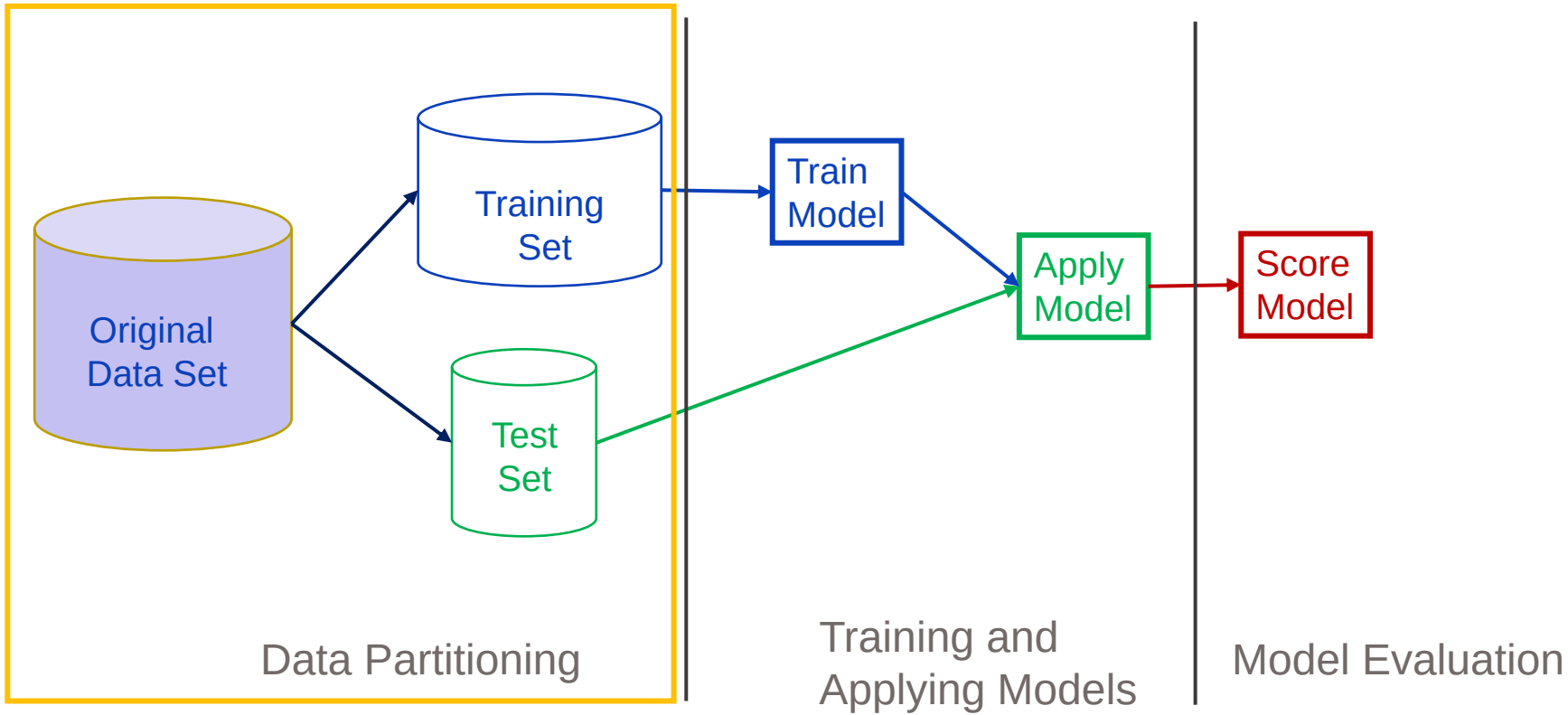






These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

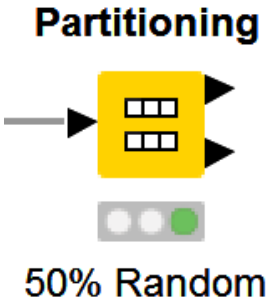
Predictive Modeling Overview

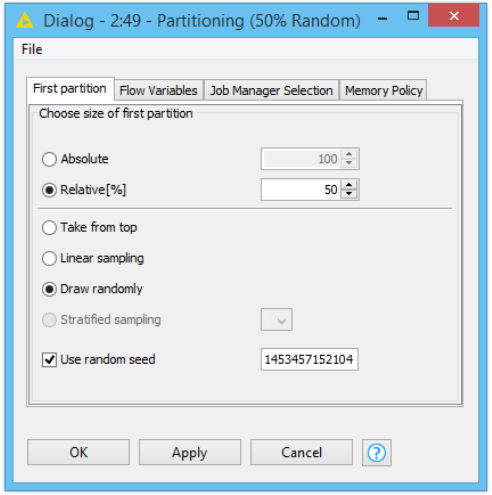


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Partitioning

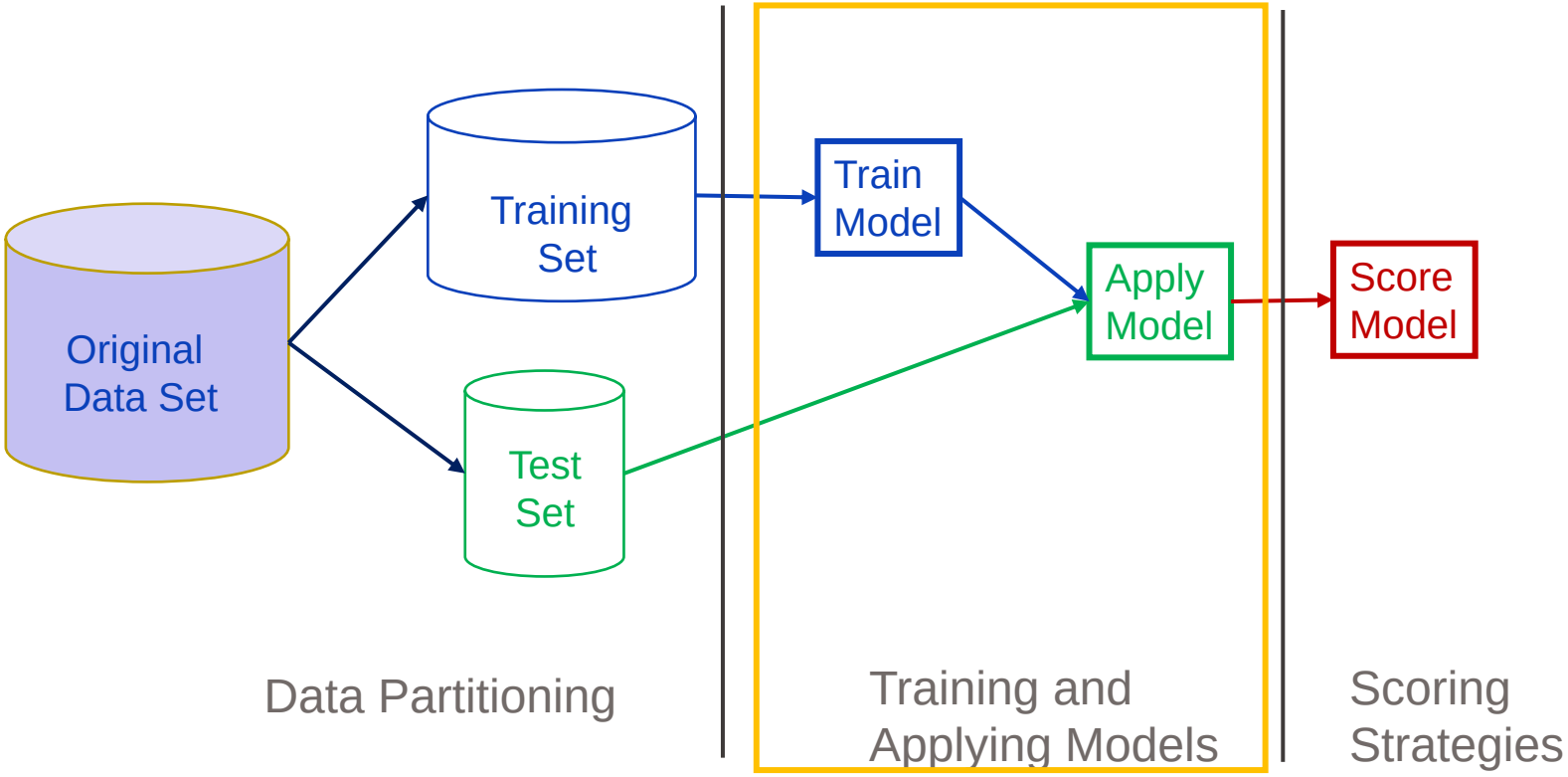
- Use it to split data into training and evaluation sets
- Partition by count (e.g. 10 rows) or fraction (e.g. 10%)
- Sample by a variety of methods; random, linear, stratified





These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

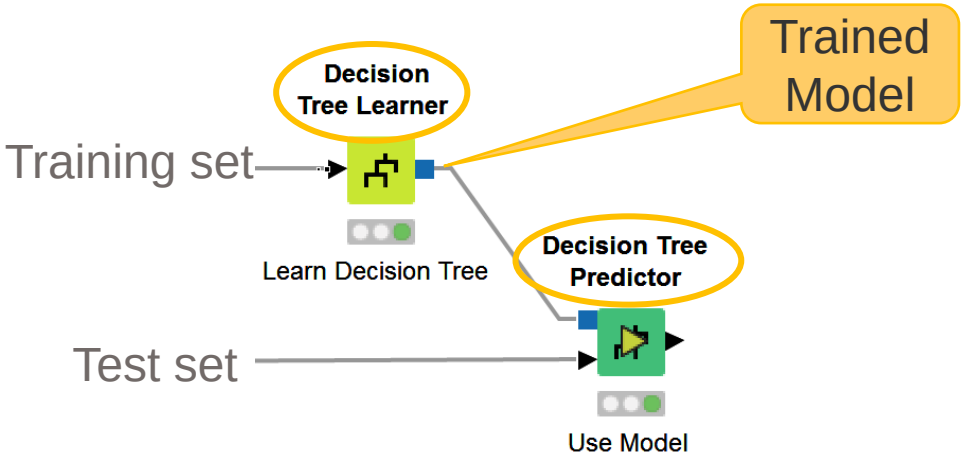
Predictive Modeling Overview



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The Learner-Predictor Motif

- All data mining models use a Learner-Predictor motif.
- The Learner node trains the model with its input data.
- The Predictor node applies the model to a different subset of data.



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Decision Tree

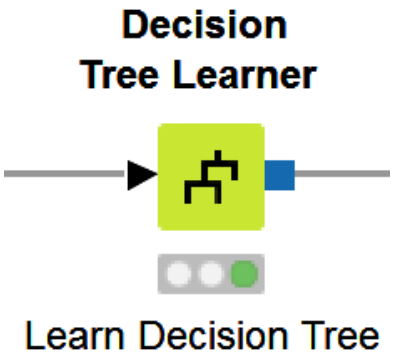
J.R. Quinlan, “C4.5 Programs for machine learning”

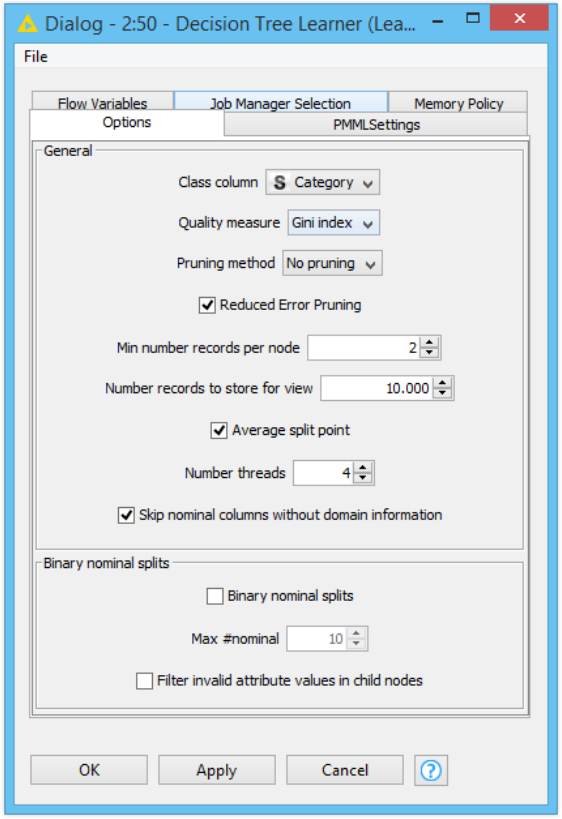
J. Shafer, R. Agrawal, M. Mehta, “SPRINT: A Scalable Parallel Classifier for Data Mining”

- C4.5 builds a tree from a set of training data using the concept of information entropy.
- At each node of the tree, the attribute of the data with the highest normalized information gain (difference in entropy) is chosen to split the data.
- The C4.5 algorithm then recourses on the smaller sub lists.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

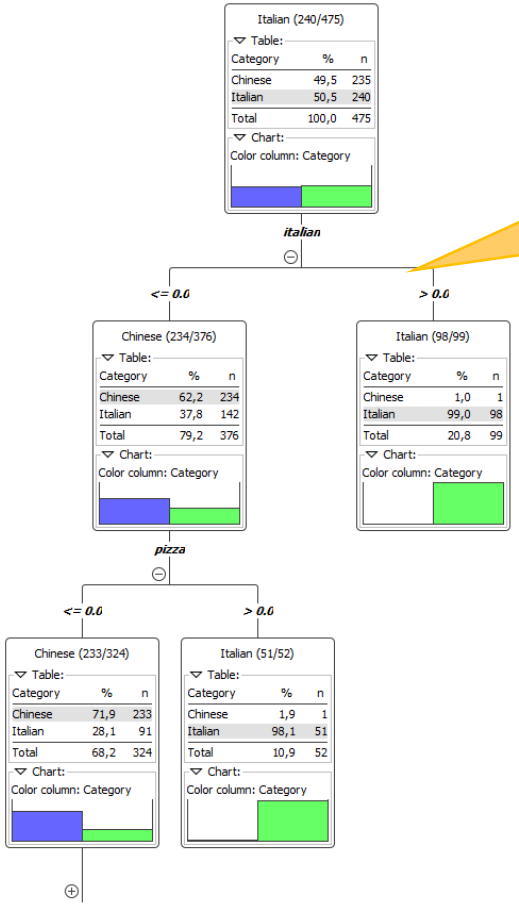
New Node: Decision Tree Learner





These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Decision Tree: View

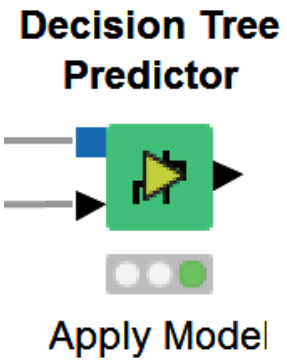


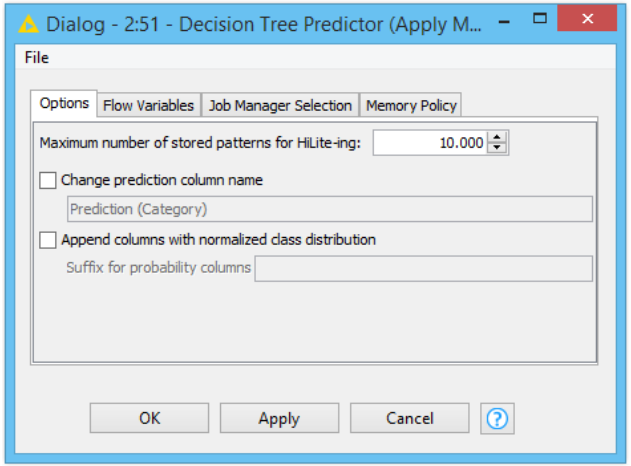
If the word “Italian” occurs in a review, the restaurant is very likely an Italian restaurant.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Decision Tree Predictor

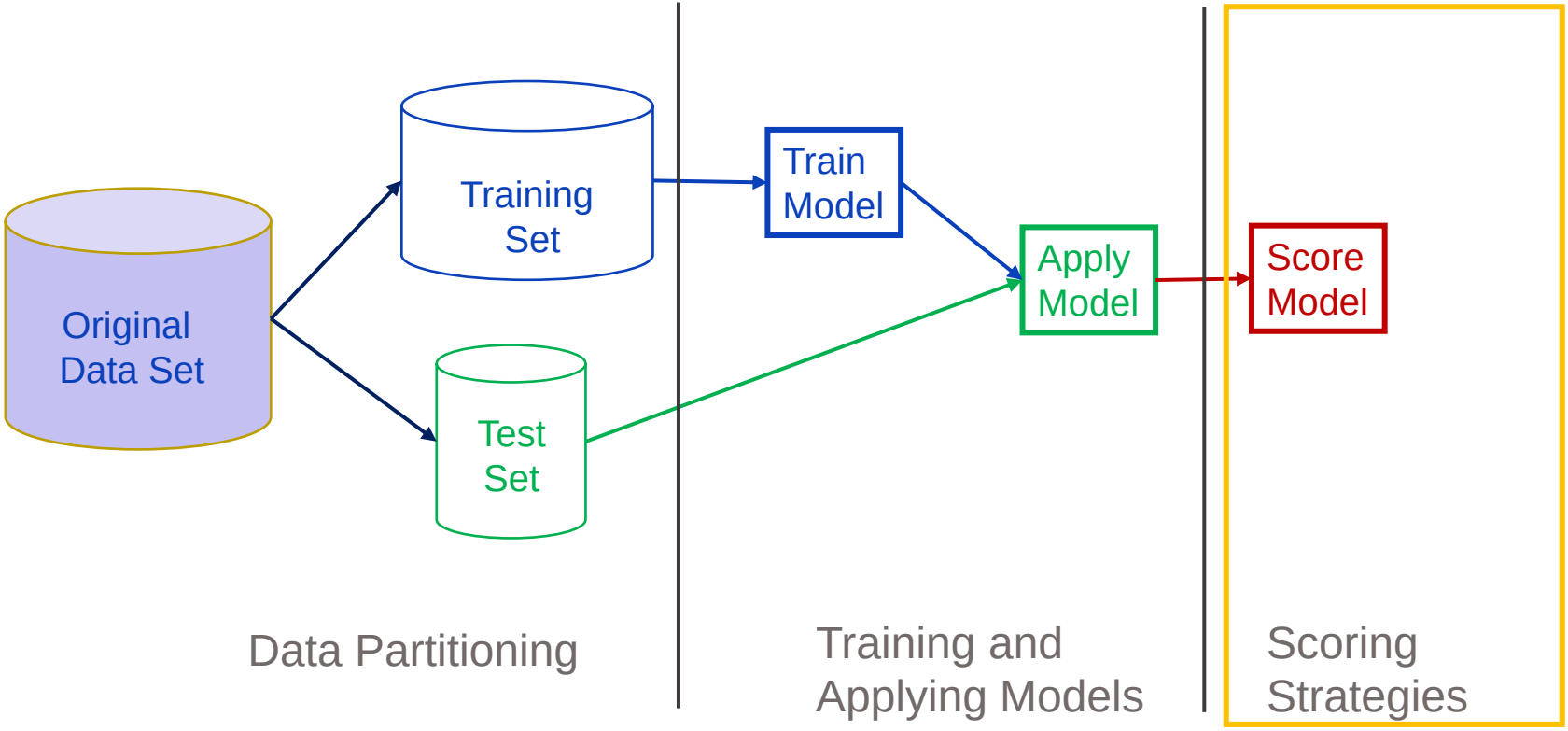
- Consumes a Decision Tree model and new data to classify
- Check the box to append class probabilities





These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

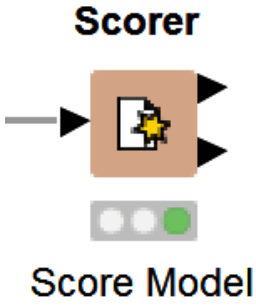
Predictive Modeling Overview

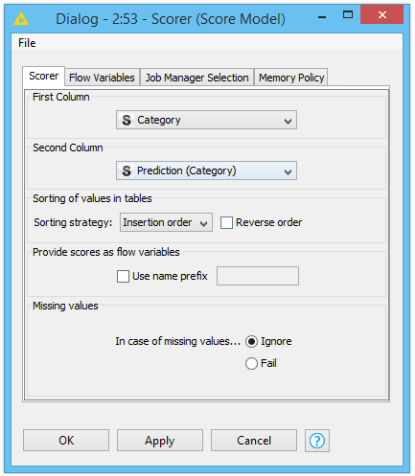


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

New Node: Scorer

- Compare predicted results to known truth to evaluate model quality
- Confusion matrix shows the distribution of model errors
- An accuracy statistics table provides additional info

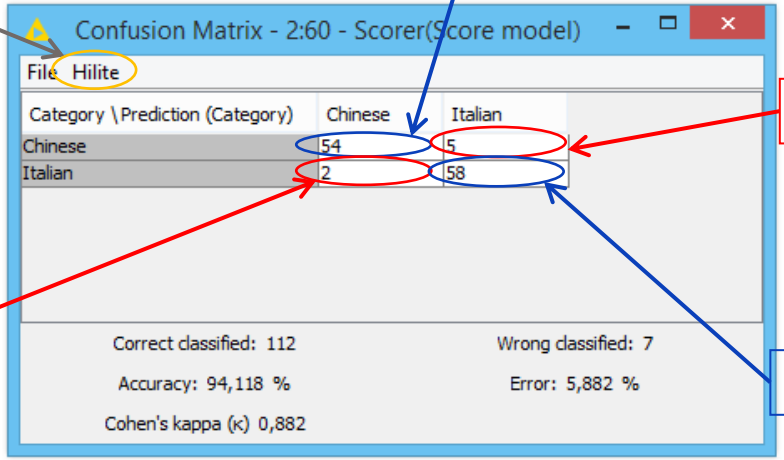




These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Scorer: Confusion Matrix

This is the difference between the confusion matrix data table and the confusion matrix view



True Positives

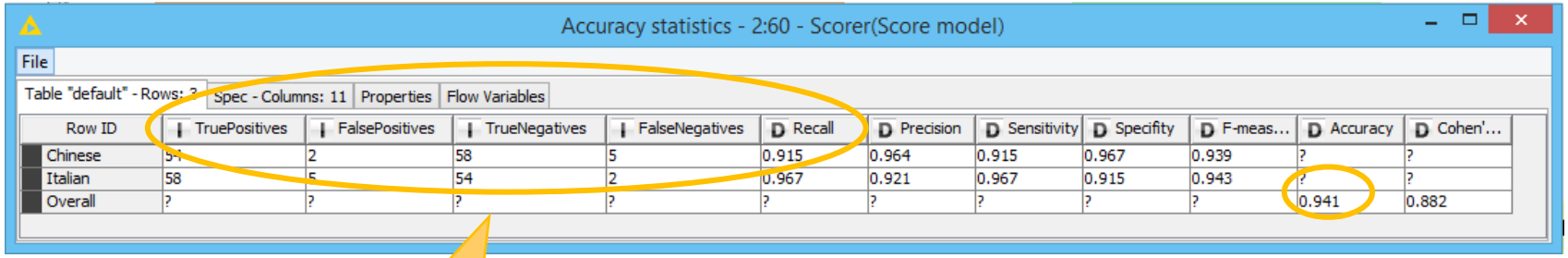
False Negatives

False Positives

True Negatives

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Scorer: Accuracy Measures



Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
Chinese	54	2	58	5	0.915	0.964	0.915	0.967	0.939	?	?
Italian	58	5	54	2	0.967	0.921	0.967	0.915	0.943	?	?
Overall	?	?	?	?	?	?	?	?	?	0.941	0.882

From the confusion matrix

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Exercise

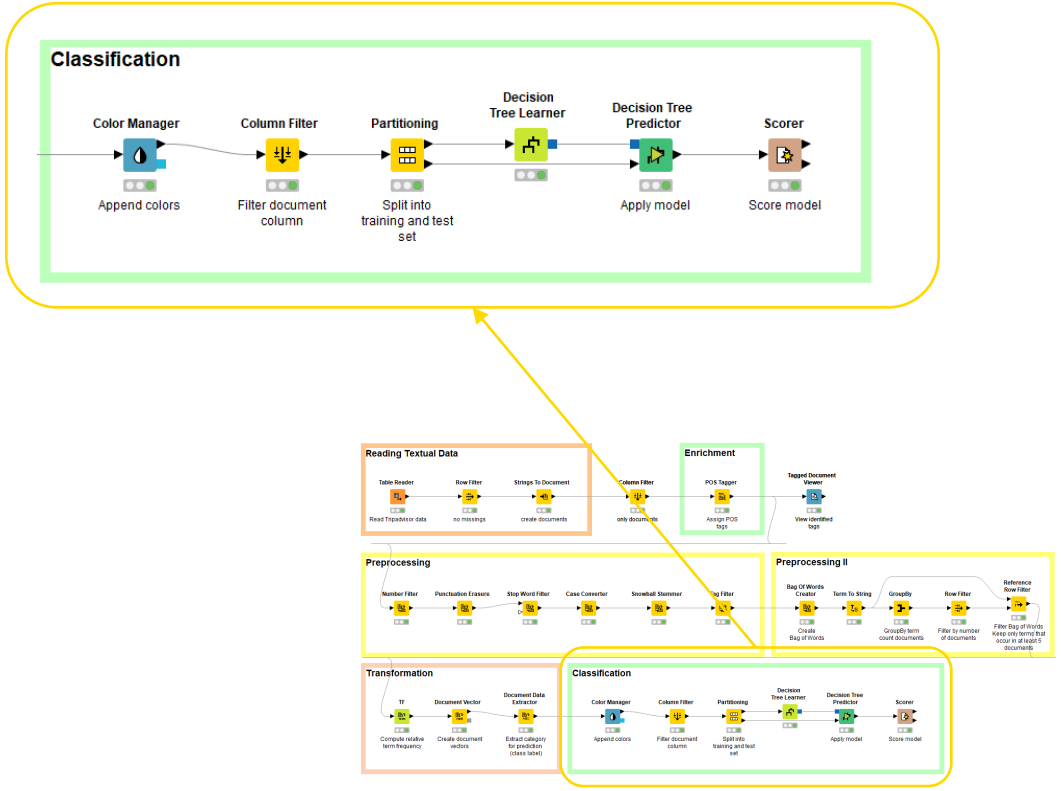
- Start with “Exercise: Classification”
  - Append color information based on class labels
  - Split data into training and test set
  - Train decision tree classifier on training set
  - Apply trained model on test set
  - Score model

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Section Solution

Classification

- Color Manager
- Column Filter
- Partitioning
- Decision Tree Learner
- Decision Tree Predictor
- Scorer



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Overview Text Mining Use Cases

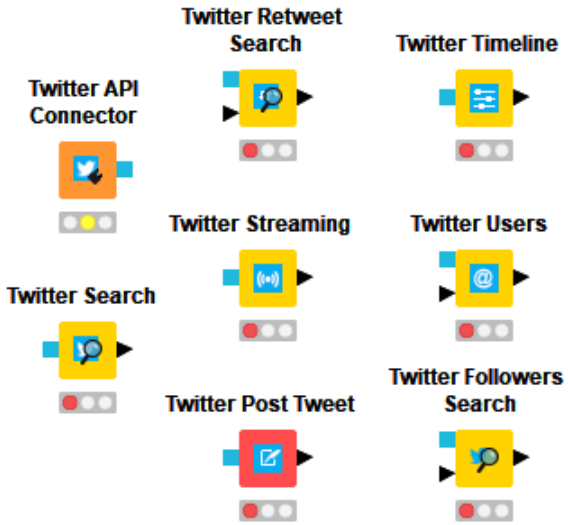
Use Cases

1. Mining Social Media Activity from Twitter
2. Sentiment Analysis of Social Media Activity
3. Network Analysis of User Activity on Social Media
4. Entity Recognition in Documents
5. Topic Detection in Documents with LDA
6. Guided Labeling: Train Document Classifiers from Unlabeled Data

1. Mining Social Media Activity from Twitter

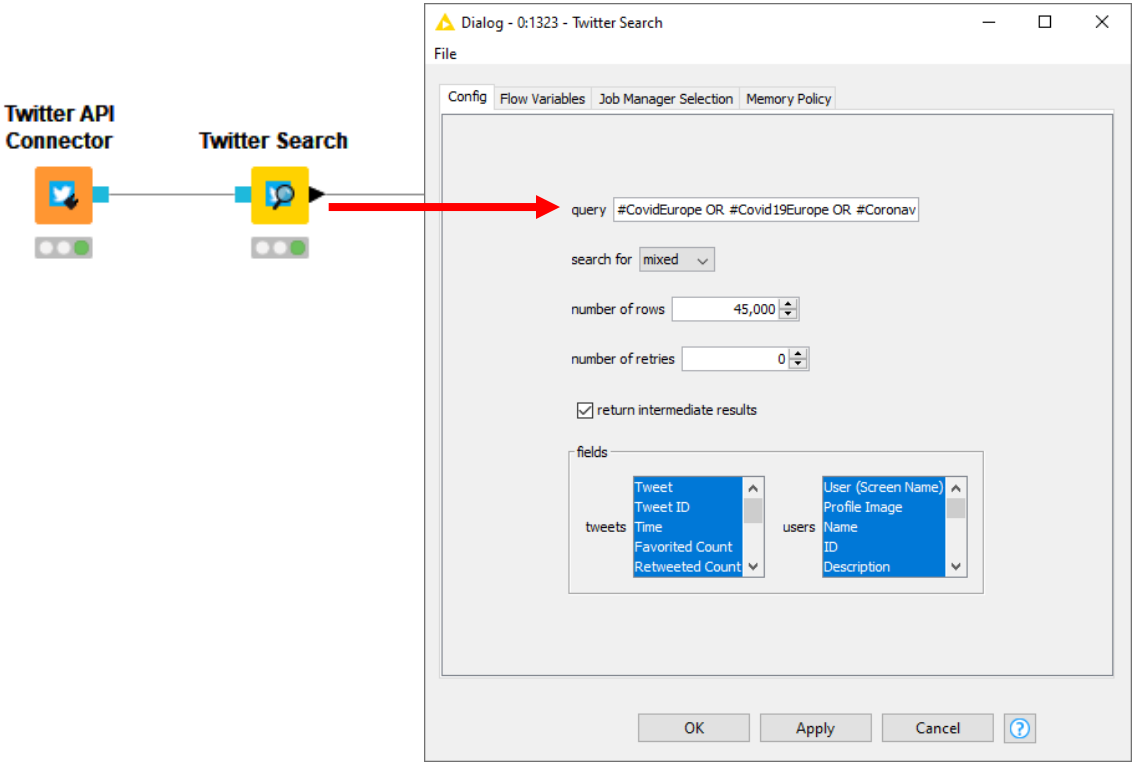
The Twitter API Integration

- To import tweets data directly into KNIME via official Twitter API
- Different nodes for different tasks
- Twitter API Connectors needs keys from a Twitter app



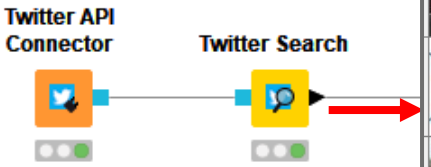
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

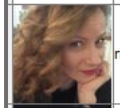
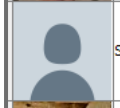
Use Case: Download #COVID19Europe Tweets



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

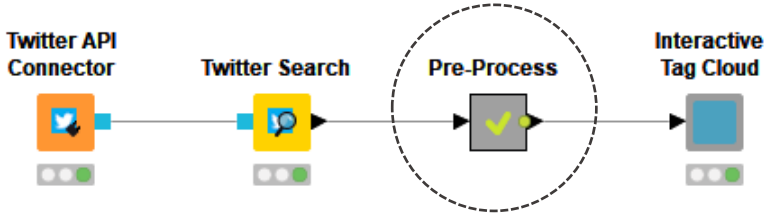
Use Case: Download #COVID19Europe Tweets

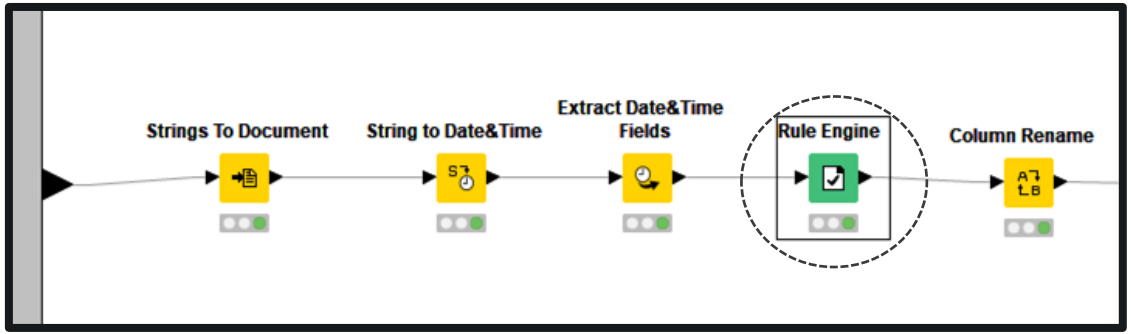


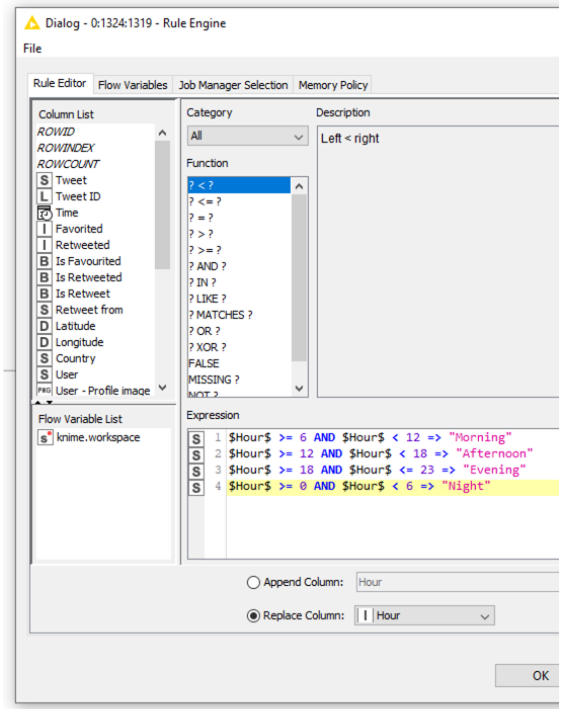
User - ...	User	Tweet	Tweet ID	Time	
	Slabodkin	#COVID19 drives #EC to change rules on #notifiedbody designations https://t.co/E4xIy1zeBX #n...	1263141884643495...	2020-05-20 18:17:47	0
	mykonostick...	Οι χώρες της Ευρωπαϊκής Ένωσης πρέπει τώρα να επικεντρωθούν στις επενδύσεις στη δημόσια υγ...	1263096288985395...	2020-05-20 15:16:37	0
	riikka_kevo	Tässä koko pöytäkirja ko. kokouksesta. Tarkkuudella: kuka jäsenistä sanoi ja kommentoi mitä Euro... https://t.co/IsdEEmt9Kx	1263060976284336...	2020-05-20 12:56:17	0
	ernesturtasun	RT @RasmusAndresen: Happy to discuss the #covid crisis, the situation in Spain and a common eu...	1262808131127717...	2020-05-19 20:11:34	1
	RasmusAndr...	Happy to discuss the #covid crisis, the situation in Spain and a common european answer with my ...	1262805031851614...	2020-05-19 19:59:16	1
	Manos81A	Γίνονται και θαύματα? Ανάγκα και Γερμανοί πείθονται. Μην τους ματιάσουμε. #κορωνοιος #κορονοιος #covid19EUROPE #coronavirusEurope #COVID19Pandemic #Greece #c...	1262747384871714...	2020-05-19 16:10:11	0
	StokNathalie	150.138 #Corona Tote in #Europa - Stand 7. Mai Europa ist der von der #CoronaPandemie am härtesten getroffene Kontinent: Vor allem #Italien... #SARSCoV2 #ChinaVirus #CoronavirusDE #coronavirus	1262705115682742...	2020-05-19 13:22:14	0
	Aravalipoesia	RT @GAIA_Insights: The Language of Impactful Leaders, Join our next GAIA Gives free webinar o... Register here: https://t.co/yClMjG4esq #gaiainsights #together #webinar #resources #COVID19Europe #leadership https://t.co/xsr5OF...	1262676274943795...	2020-05-19 11:27:37	1

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Use Case: Download #COVID19Europe Tweets







These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Use Cas

Twitter API Connector

Class Name	Class Abbreviation
Afternoon	*Afternoon*
Evening	*Evening*
Morning	*Morning*
Night	*Night*

Colored with Model Labels	User - Profile image
No matching records found	

Showing 0 to 0 of 0 entries (filtered from 990 total entries)

Previous Next

Reset Apply Close

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

2. Sentiment Analysis of Social Media Activity

Sentiment Analysis – An Example

Samsung

Samsung Galaxy S7 Edge G935A 32GB Unlocked - Gold Platinum

★★★★☆ 125 customer reviews | 606 answered questions

★★★★★ **Beautiful phone from a wonderful seller!**

By on May 29, 2017

Color: Gold | **Verified Purchase**

This practically new beautiful phone well exceeded my expectations!

★☆☆☆☆ **One Star**

By on August 3, 2016

Color: Black Onyx | **Verified Purchase**

Very bad experience

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Sentiment Analysis – Some Use Cases

- ***Movies***: is this review positive or negative?
- ***Products***: what do people think about a new smartphone?
- ***Politics***: what do people think about a specific candidate or political issue?
- ***Prediction***: predict election outcomes or market trends from sentiment.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Today's Use Case

- Dataset: subset of the IMDb (Internet Movie Database) Large Movie Review Dataset v1.0 with 2000 documents(*).
  - 1000 documents from the positive group
  - 1000 documents from the negative group

- Goal: To assign the correct sentiment label to each document.

(*) For details about the data set see http://ai.stanford.edu/~amaas/data/sentiment/
Data citation: Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)

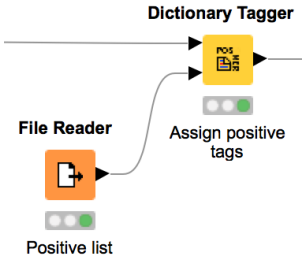
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Sentiment Analysis

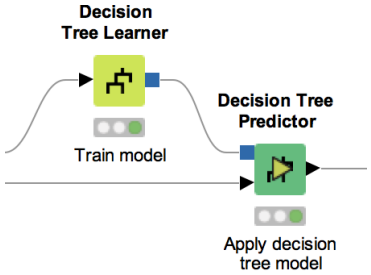
Task: Determine the expressed opinion in a document/text, e.g. positive, negative

Sentiment Analysis = Opinion Mining = Emotion AI

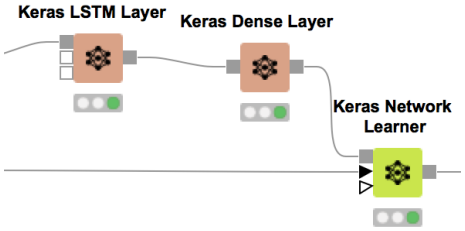
Lexicon Based



Machine Learning



Deep Learning



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Approach 1: Lexicon Based

Idea: Rule-based classification with Dictionary Tagger

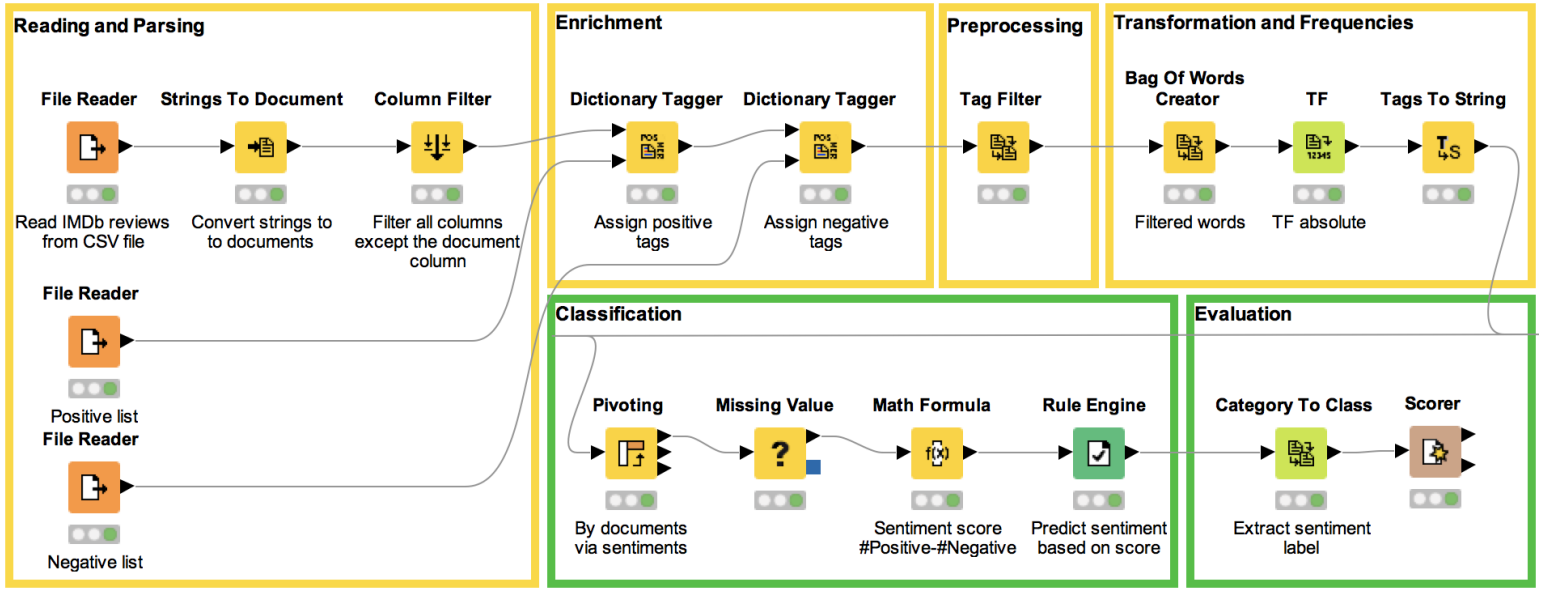
1. Use custom dictionary to tag positive and negative words
2. Count number of positive and negative words per document
3. Assign classes depending on the number of positive and negative words

$$sentiment\ score = \frac{\#\ positive\ words - \#\ negative\ words}{\#total\ words}$$

Advantage: No labels needed

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Rule-based Classification with Dictionary Tagger



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Approach 2: Machine Learning

Idea: Train a model to make predictions on unseen data

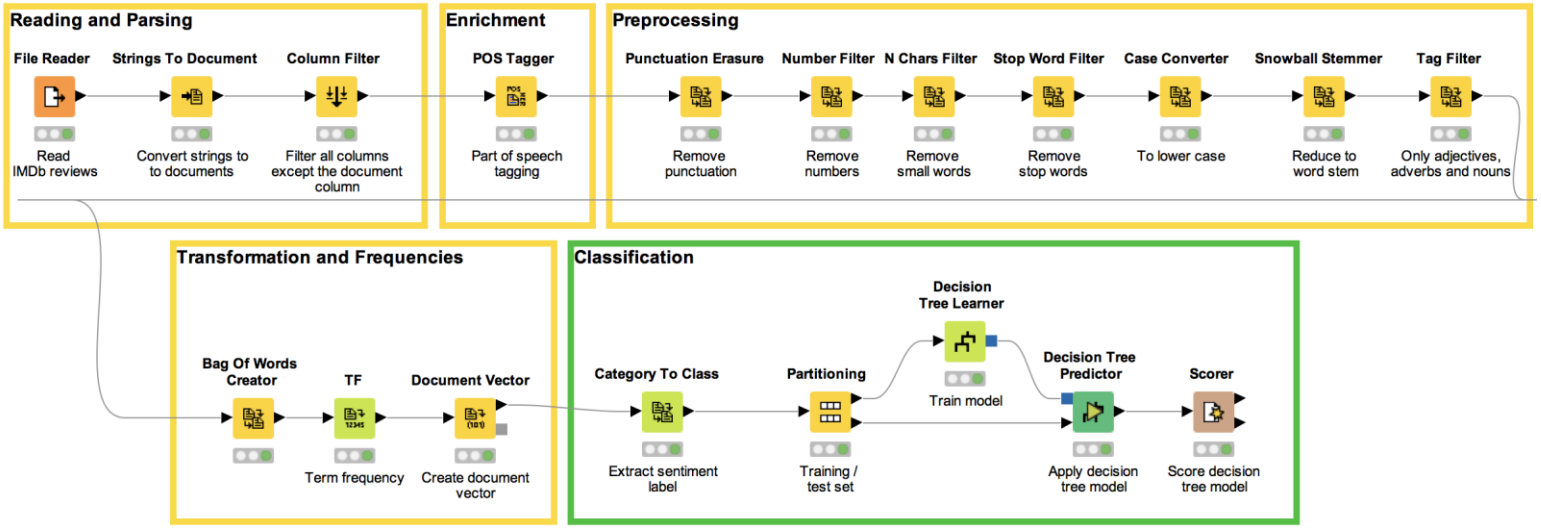
1. Start the analysis with a labeled dataset
2. Extract a feature space from the documents, e.g. only keywords
3. Train a supervised model, e.g. decision tree, logistic regression

Advantage: Better performance

Vohra, S. M., and J. B. Teraiya. "A comparative study of sentiment analysis techniques." Journal JIKRCE 2.2 (2013): 313-317.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Approach 2: Machine Learning



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Approach 3: Deep Learning

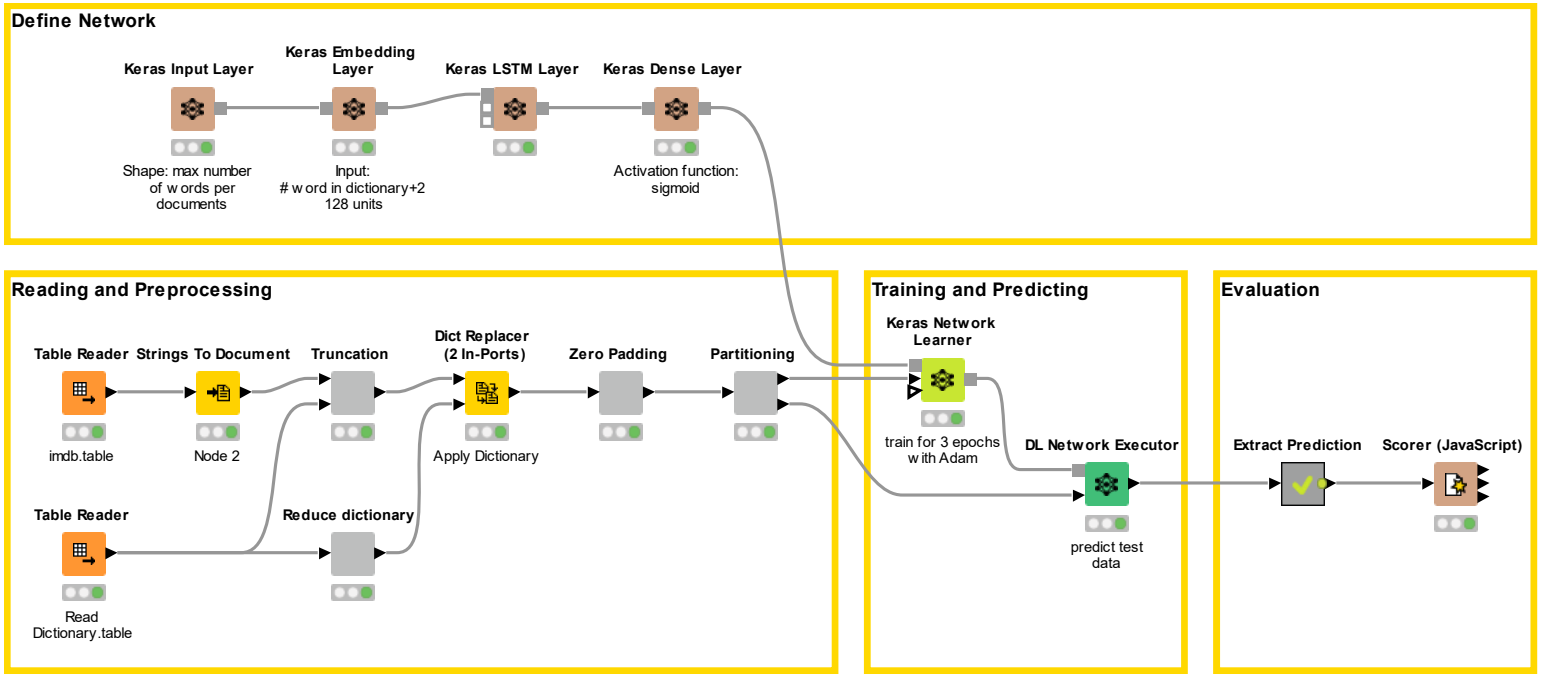
Idea: Train a network to make predictions on unseen data

1. Start the analysis with a labeled dataset
2. Bring the documents into a standard numeric shape
   - Hot encoding vs. embeddings
   - Zero padding and truncation
3. Define the network structure. Commonly used layers for text are embedding and LSTM layers
4. Train and apply the network

Advantage: Can lead to better performance on big data sets

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

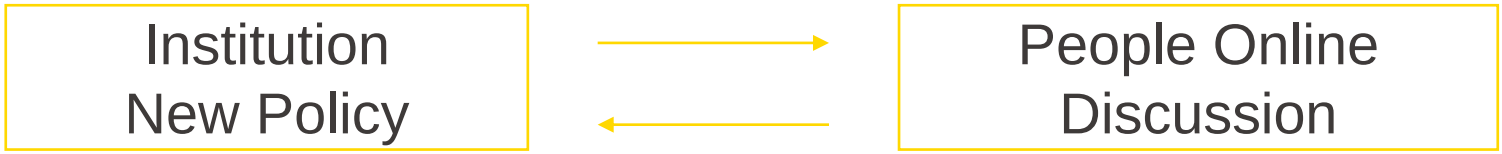
Approach 3: Deep Learning



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

3. Network Analysis of User Activity on Social Media

The Data Science Problem



- Can you tell us what people say about our new policy?
- Can you tell us how many support the policy?
- How many are against it?
- Of those, can you tell us who is an influencer?

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The Data from Slashdot Forum

- Slashdot was a public forum built in 1997 and hosting several discussions: from software to philosophy, from science fiction to politics.
- Politics was the biggest discussion group

- So, politics is what we analyzed to find out:
  - What users were thinking about a political issue
  - Who was pro and who was con
  - Who was an influencer

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The Techniques

Text Analytics

1. Forum posts as *Documents.*
2. Simple *Sentiment Analysis.* counting pos. vs neg. words (*Lexicon Based*).
3. Scoring Sentiment on *Documents* (posts) and then overall on *Authors* (users).
4. Visualize in Tag Cloud

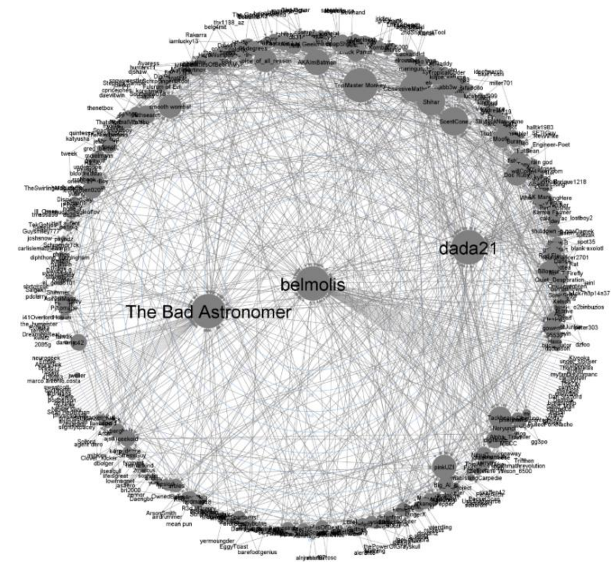
Network Mining

1. Forum posts as network connections between users using the Forum threads.
2. Connection in network weighted on number of posts between users.
3. Find in the network influencers (*Hubs).*
4. Visualize in Network View

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The Techniques

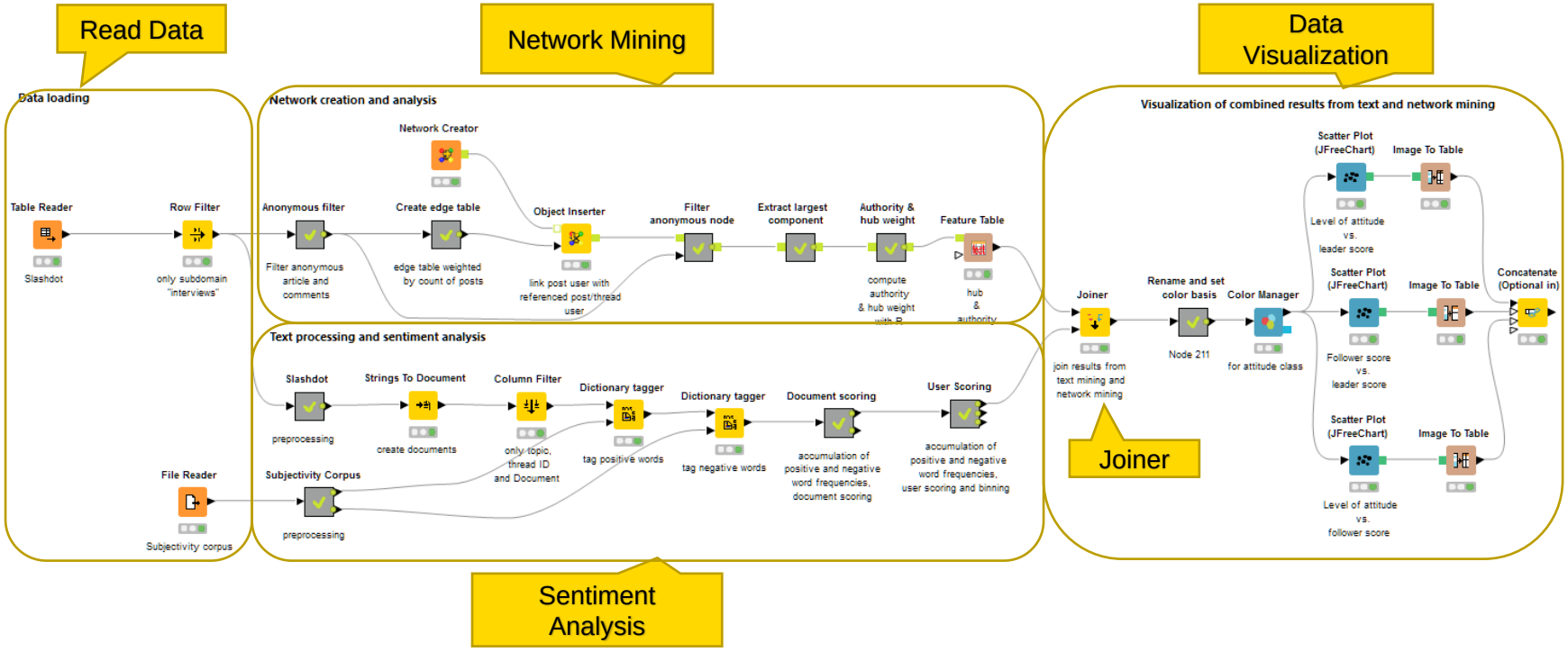
[Text Analytics](#)



[Network Mining](#)

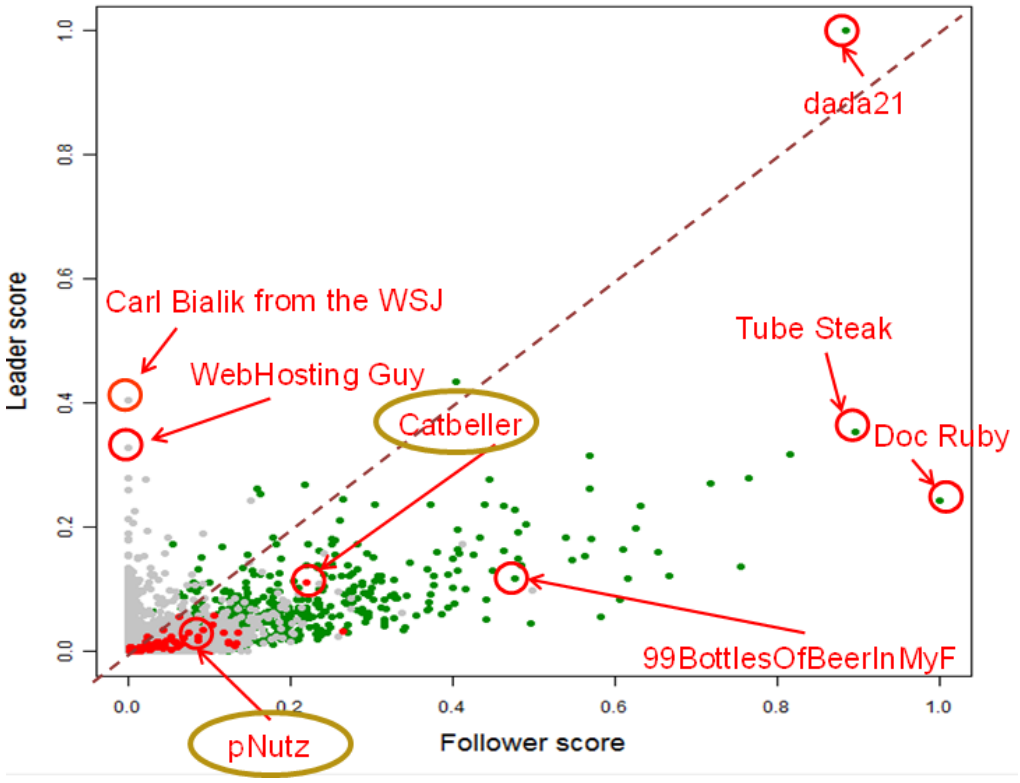
These slides are a derivative of [KNIME Course Material](#) of [KNIME AG](#) used under [CC BY 4.0](#)

Text Analytics and Network Mining: Workflow



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Text Analytics and Network Mining: Results



When looking on activity on social media is better to understand the argumentations of the most influential complainers, rather than the loudest ones.

In Favor

Neutral

Against

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

4. Entity Recognition in Documents

Text Tagging

Text tagging is the process of adding a tag, that is a language element descriptor, to unstructured data.

Tagged Document Viewer

You can display tags in docs with this node:

1	Chronic kidney disease (CKD) affects ~40% of patients with diabetes. DISEASE

Why do we need tags on documents?
How can we automate this with thousands of documents?

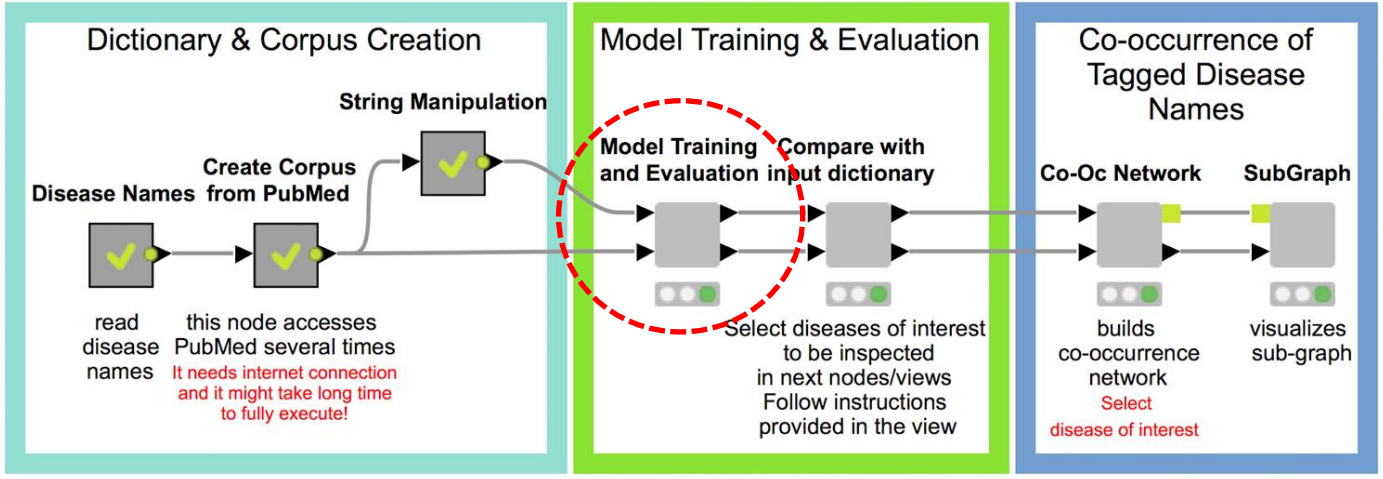
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Data Science Problem in Healthcare

- Medical text records from hospitals can be analyzed for aggregated and anonymous statistics
- Tag type of interest: disease names (e.g. “diabetes”)
- Showing those medical records to third party experts for manual tagging is often not possible
- This problem can be reproduced when using biomedical publications from PubMed.

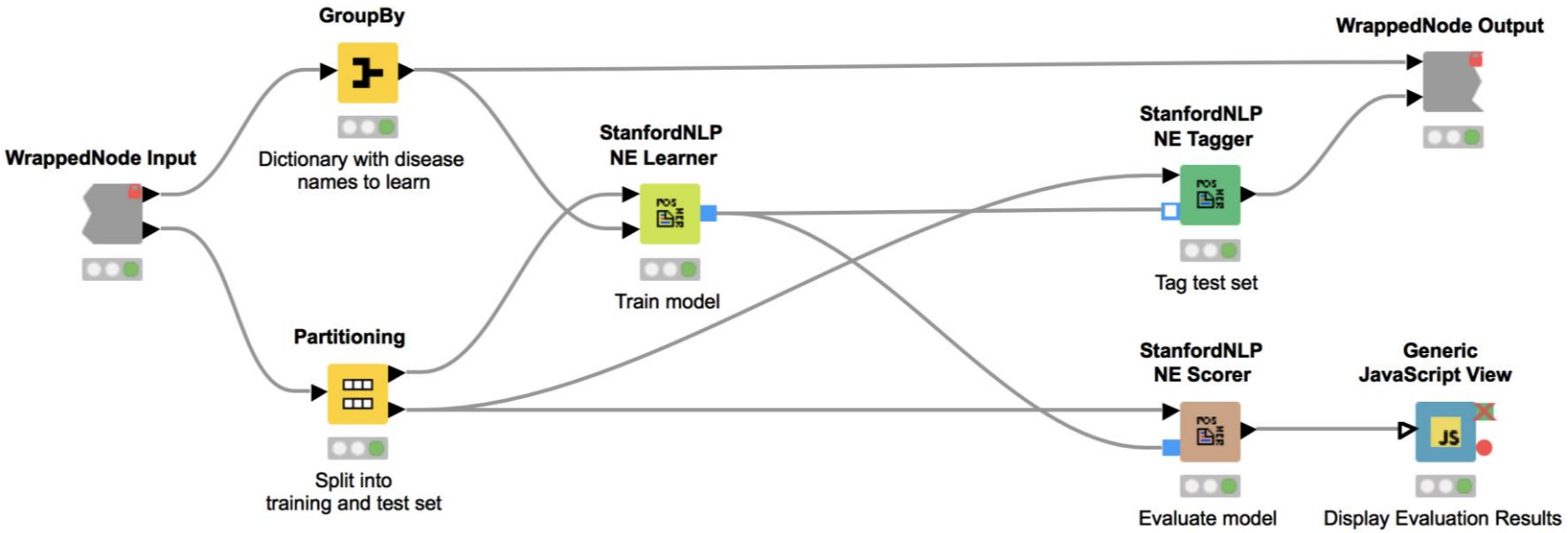
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The Workflow



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The Workflow

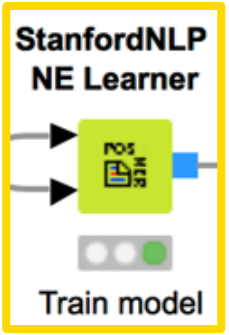


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Train NER Model – StanfordNLP NE Learner

- Training by using the **StanfordNLP NE Learner**
  - ⓾ Top port: The (training) documents
  - ⓾ Bottom port: Dictionary of disease names

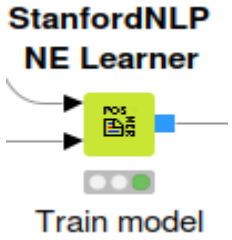
- Internally annotates the documents based on the dictionary

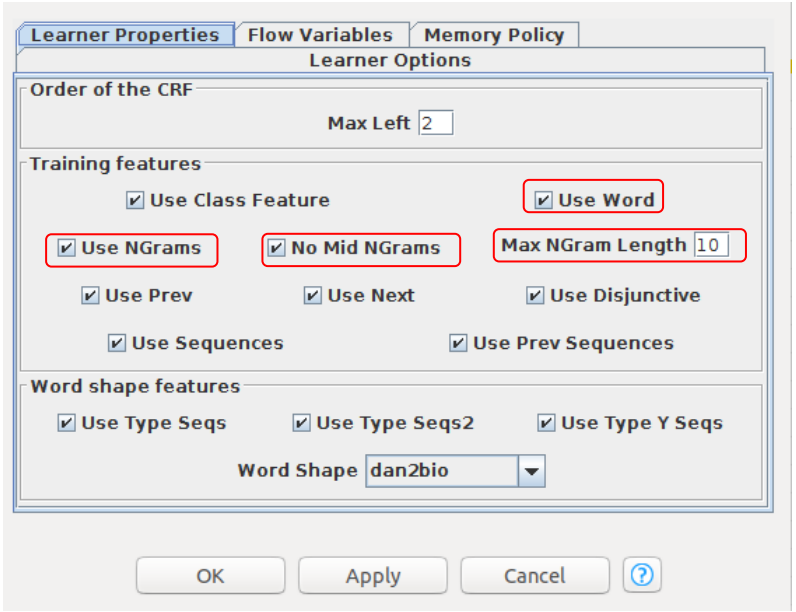


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

StanfordNLP NE Learner - Dialog

- **Use Word**: Whole entity is used as feature
- **Use NGrams**: Substrings of a word are used as feature
- **No Mid NGrams**: Only use substrings that include the beginning or ending of a word
- **Max NGram Length**: Maximum length of a substring





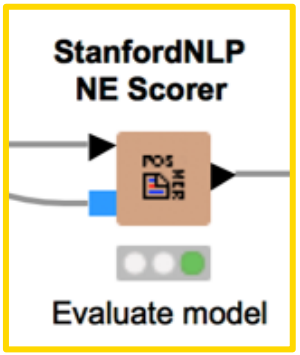
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

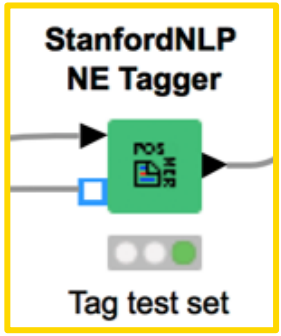
Evaluating the model – StanfordNLP NE Scorer

- Evaluating the model with **StanfordNLP NE Scorer**
  - Top : the (test) documents // Bottom : the model
- Calculating statistics such as precision and recall

Row ID	D Precision	D Recall	D F1	I TP	I FP	I FN
Row0	0.983	0.99	0.987	179668	3110	1773

- Deploy with the **StanfordNLP NE Tagger**
  - Model won't perform well to detect new entities
  - Top : the (new) documents // Bottom : the model





These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

5. Topic Detection in Documents with LDA

Background Story

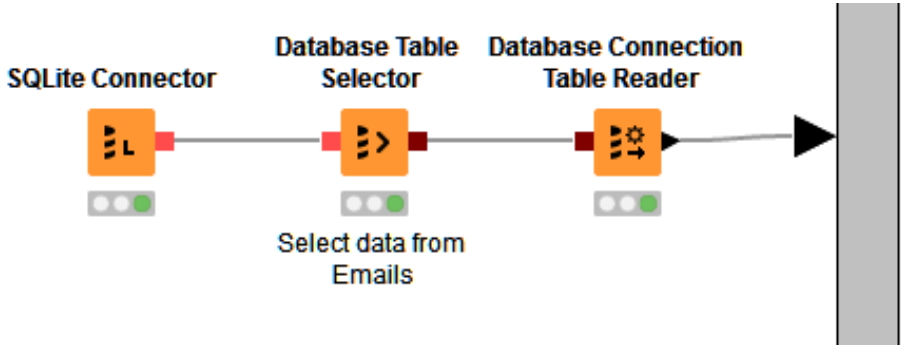
- Controversy regarding former Secretary Hillary Clinton's use of a non-government, privately maintained email server
- On 2015, Hillary Clinton released 7,945 emails in response to a U.S. Department of State Freedom of Information Act (FOIA) request
- The original emails are available on the State Department's Website

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Acquiring and Accessing the Data

- Data available at kaggle.com/kaggle/hillary-clinton-emails

- The data have been downloaded and accessed as SQLite database



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Acquiring and Accessing the Data

MetadataDateReleased	MetadataPdfLink	MetadataCa...	Metadat...	Extract...	Extract...	ExtractedFrom	ExtractedCc	ExtractedDateSent	Extract...	Extract...	Extract...	Extracted...	ExtractedBodyText	RawText
2015-05-22T04:00:00+00:00	DOCUMENTS/HRC_Email...	F-2015-04841	HRC_Email_296	FW: Wow		Sullivan, Jacob J <...		Wednesday, September ...	F-2015-04841	C05739545	05/13/2015	RELEASE IN FULL		UNCLASSIFIED U.S. Department of State Case No. F-2015-04841 Doc No. C05739545 Date: 05/13/2015 STATE DEPT. - PRODUCED TO HOUSE SELECT BENGHAZI COMM.
2015-05-22T04:00:00+00:00	DOCUMENTS/HRC_Email...	F-2015-04841	HRC_Email_296						F-2015-04841	C05739546	05/13/2015	RELEASE IN PART	B6 Thursday, March 3, 2011 9:... H: Latest How Syria is aidin... hrc memo syria aiding libya ... March 3, 2011 For: Hillary	UNCLASSIFIED U.S. Department of State Case No. F-2015-04841 Doc No. C05739546 Date: 05/13/2015 STATE DEPT. - PRODUCED TO HOUSE SELECT BENGHAZI COMM.
2015-05-22T04:00:00+00:00	DOCUMENTS/HRC_Email...	F-2015-04841	HRC_Email_296	Re: Chris St...	B6	Mills, Cheryl D <Mill...	Abedin, Huma	Wednesday, September ...	F-2015-04841	C05739547	05/14/2015	RELEASE IN PART	Thx	UNCLASSIFIED U.S. Department of State Case No. F-2015-04841 Doc No. C05739547 Date: 05/14/2015 STATE DEPT. - PRODUCED TO HOUSE SELECT BENGHAZI COMM.
2015-05-22T04:00:00+00:00	DOCUMENTS/HRC_Email...	F-2015-04841	HRC_Email_296	FW: Cairo C...		Mills, Cheryl D <Mill...	Mitchell, Andrew B	Wednesday, September ...	F-2015-04841	C05739550	05/13/2015	RELEASE IN PART		UNCLASSIFIED U.S. Department of State Case No. F-2015-04841 Doc No. C05739550 Date: 05/13/2015 STATE DEPT. - PRODUCED TO HOUSE SELECT BENGHAZI COMM.
2015-05-22T04:00:00+00:00	DOCUMENTS/HRC_Email...	F-2015-04841	HRC_Email_296						F-2015-04841	C05739554	05/13/2015	RELEASE IN PART	H <hrod17@clintonemail.co... Friday, March 11, 2011 1:3... Huma Abedin Fw: H: Latest: How Syria is ... hrc memo syria aiding libya ... Pls print.	B6 UNCLASSIFIED U.S. Department of State Case No. F-2015-04841 Doc No. C05739554 Date: 05/13/2015
2015-05-22T04:00:00+00:00	DOCUMENTS/HRC_Email...	F-2015-04841	HRC_Email_296	Meet The Ri...				Wednesday, September ...	F-2015-04841	C05739559	05/13/2015	RELEASE IN PART	Pls print. -•-...-^ H < hrod17@clintonemailco... Wednesday, September 12... 'RussorV@state.gov' Fw: Meet The Right-Wing E...	B6 UNCLASSIFIED U.S. Department of State Case No. F-2015-04841 Doc No. C05739559 Date: 05/13/2015

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Browsing the Emails

Quick Search: Mills | Authors

#	Document Title	Authors
539	HAITIAN JUDGE SAYS AMERICAN MISSIONARIES WILL BE FREED THIS WEEK	Mills Cheryl D
540	HAITIAN POLITICAL SUPPORT/FRUSTRATION WITH US	Mills Cheryl D
541	HANDOVER IN SCA - MAY 15	Mills Cheryl
542	HANDOVER IN SCA - MAY 15	Mills Cheryl
543	HANK NAUGHTON & FOX NEWS	- Millscd
544	HAPPY BIRTHDAY MY DEAR	Mills Cheryl D
545	HAPPY CHRISTMAS - WE ALL GOT OUR WISH - THAT YOU BE CONFIRMED. XO CDM	Mills Cheryl D
546	HAPPY HOLIDAYS!	Mills Cheryl D
547	HAPPY HOLIDAYS!	Mills Cheryl D
548	HAPPY HOLIDAYS!	Mills Cheryl D
549	HAVE	
550	HAVE	
551	HAVE	
552	HEAD	
553	HEAD	
554	HEAD	
555	HEAD	
556	HEAR	
557	HELP	
558	HELP!	
559	HELP!	
560	HERE	
561	HERE	
562	HERE	
563	HERE	
564	HERE	
565	HERE	
566	HERE'S MY PERSONAL EMAIL	Mills Cheryl D
567	HERE'S MY PERSONAL EMAIL	Mills Cheryl D
568	HERE'S MY PERSONAL EMAIL	Mills Cheryl D
569	HEY	Mills Cheryl D

HAPPY BIRTHDAY MY DEAR

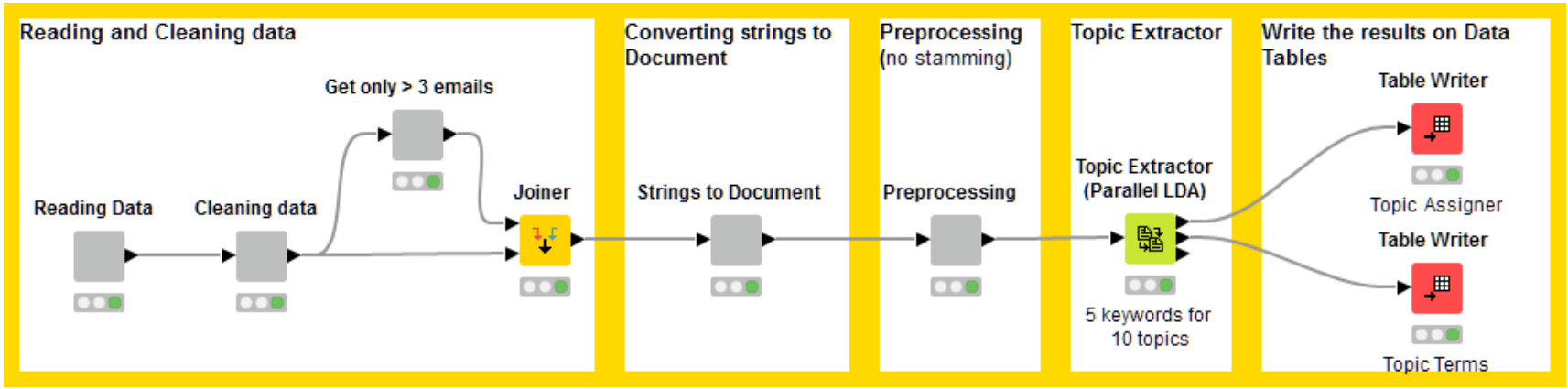
UNKNOWN

So glad you were born! Kiss your mom for me she did us right bringing you into the world. XO Cdm

Document Viewer

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

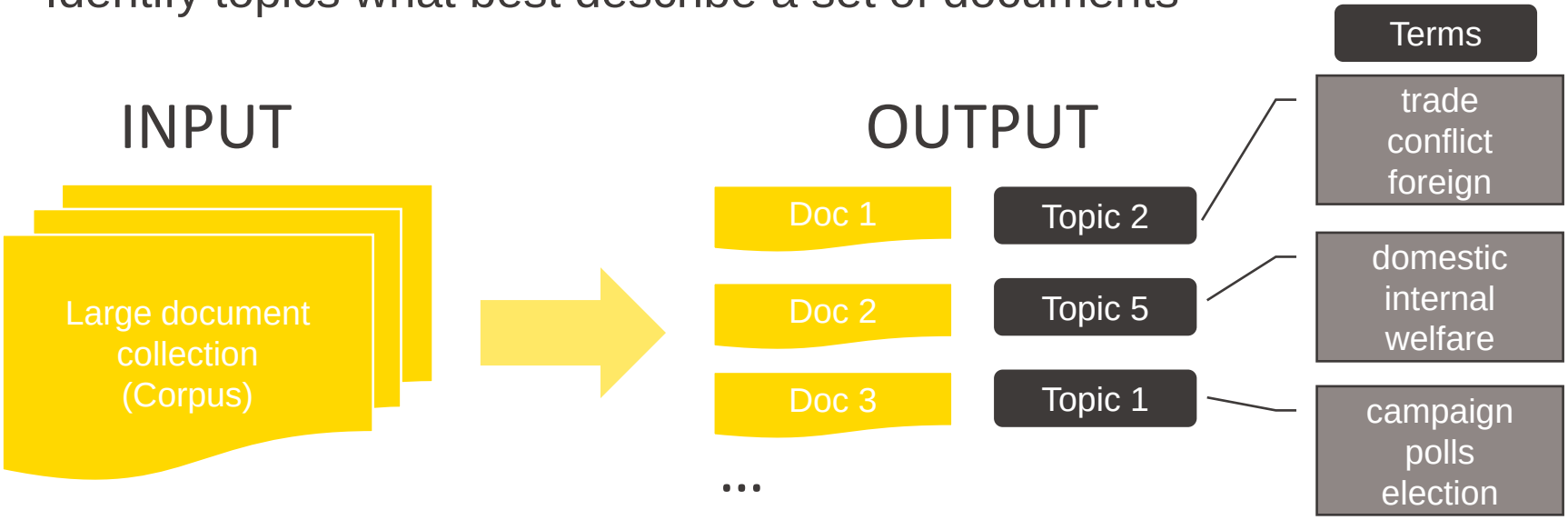
Training Model



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

What is Topic Modeling?

- Developed to organize large collections of documents and to **extract the hidden thematic structure**
- Identify topics what best describe a set of documents



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

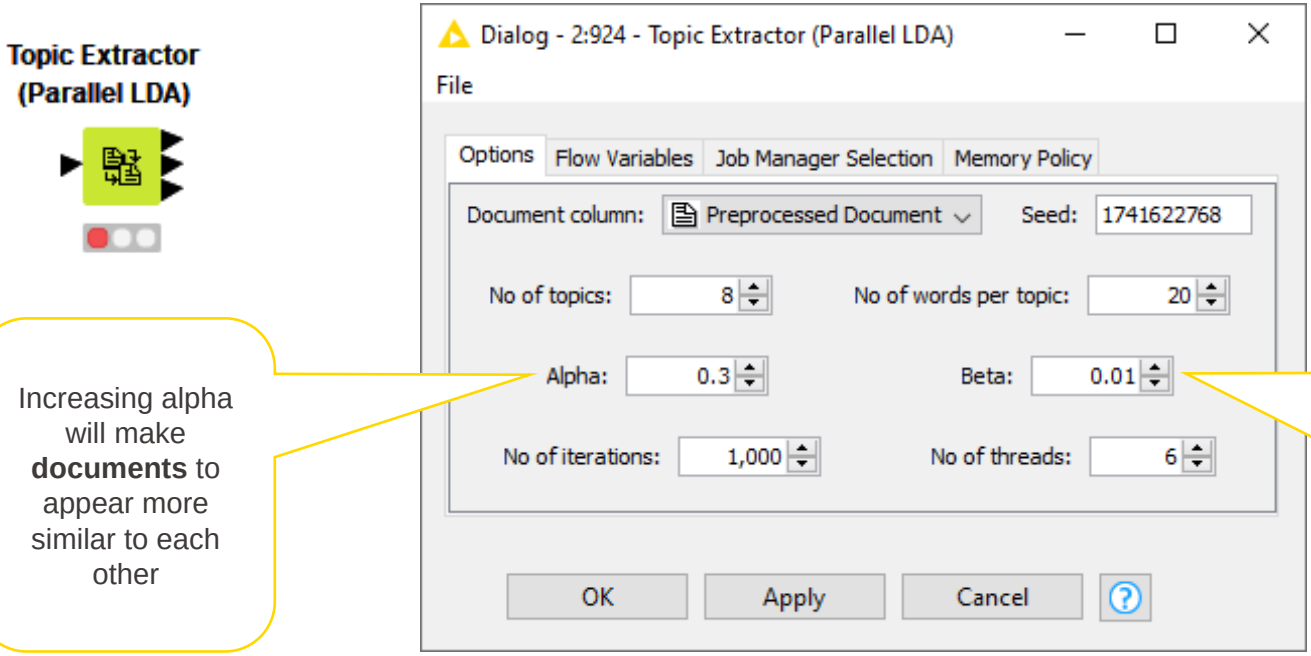
Latent Dirichlet Allocation (LDA)

- Automatically finds the top K topics with the most relevant N keywords discussed in a collection of unlabeled documents (considered unsupervised)
- It represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words
- Syntax or order of the words in document is not important (bag of words model)
- Document order is not important
- The same word can belong to different topics
- **The number of topics needs to be selected/known in advance**

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. **"Latent Dirichlet Allocation."** *Journal of Machine Learning Research* 3: 993–1022.

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Parallel LDA



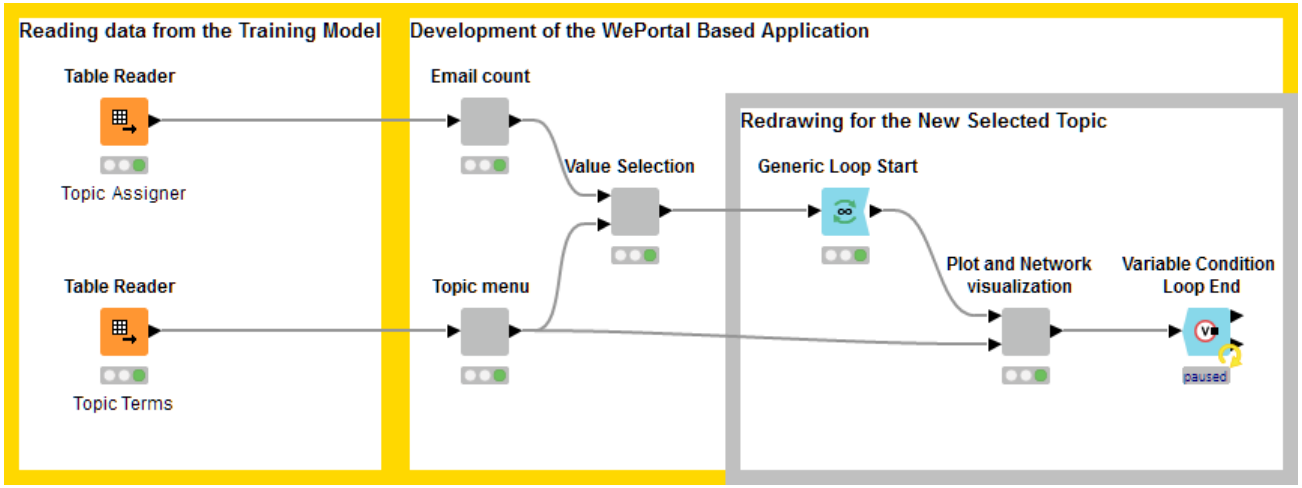
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Results: Topic and Terms

Topic id	Term
topic_0	[call, tomorrow, talk, fyi, speech]
topic_1	[office, secretary, meeting, department, arrive]
topic_2	[obama, president, house, clinton, white]
topic_3	[china, party, minister, election, labour]
topic_4	[percent, palau, favorable, unfavorable, obama]
topic_5	[benghazi, information, subject, department, doc]
topic_6	[gov, fyi, cheryl, sid, mills]
topic_7	[israel, israeli, afghanistan, military, iran]
topic_8	[party, tea, koch, political, movement]
topic_9	[women, world, people, united, security]

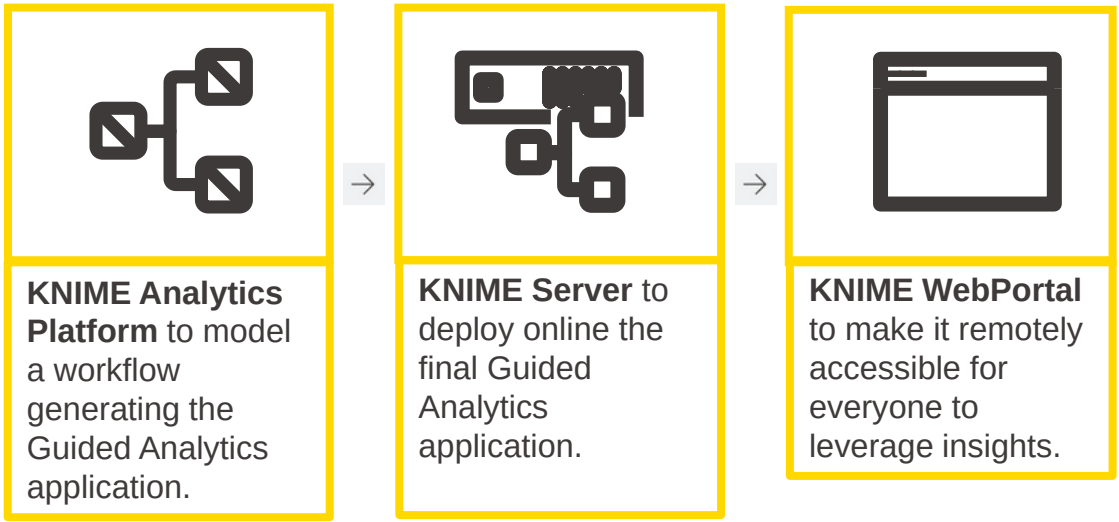
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Deployment Model



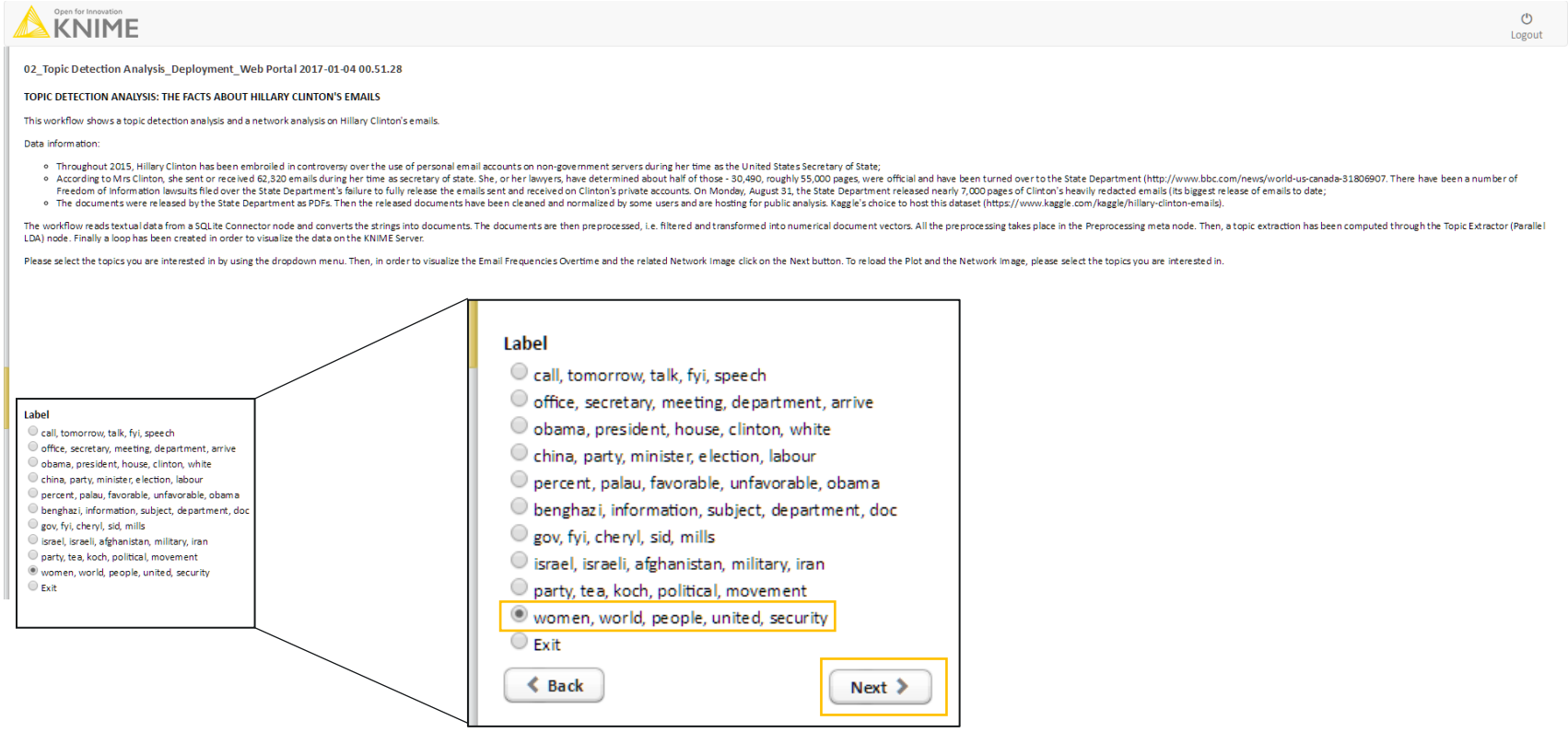
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

How KNIME Software works with Guided Analytics



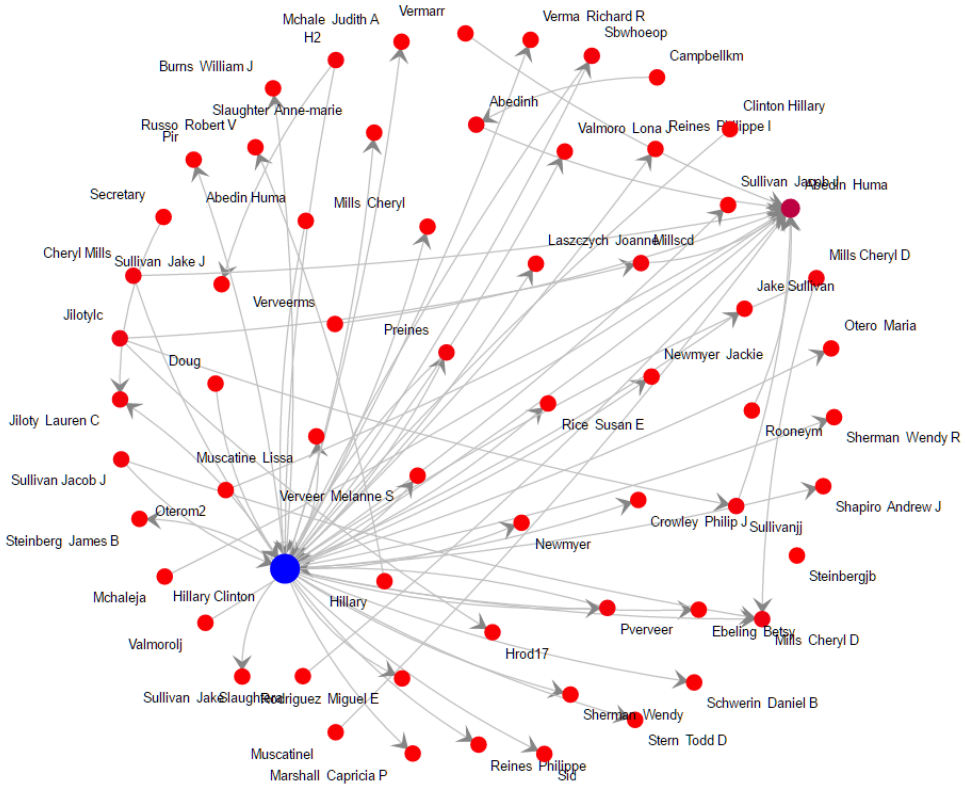
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

WebPortal – First Webpage



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

WebPortal – Second Webpage – Network Viewer



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

6. Guided Labeling:
Train Document Classifiers from Unlabeled Data

Our Data Science Problem

Train a supervised machine learning model..

- …on unlabeled documents
- ...with less labels than we would usually need
- ...provided by an expensive expert
- ...starting with no labels

Examples: *sentiment analysis, topic detection/modelling, automated document classification in precise categories..*

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Active Learning General Idea

Starting Situation:

- Lots of data points (input)
- No labels (output)

Active Learning Loop Initialization:

- Pick some unlabeled data points
- Ask oracle (expert or experiment) for labels
- Start Active Learning Loop

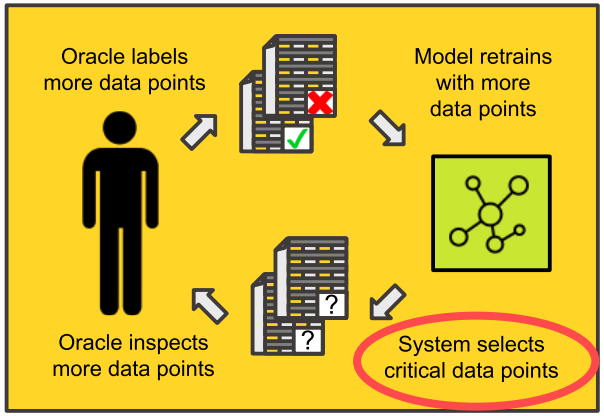
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Active Learning General Idea

Starting Situation:

- Lots of data points (input)
- No labels (output)

Human-in-the-Loop



Active Learning Loop:

- Learn from available labels
- Score all data points
- Select one or a few critical documents
- Ask oracle (expert or experiment) for labels
- Repeat

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Active Learning Sampling

intelligently

How do I select critical documents to be labeled?

- Label Density
  label first in dense but unexplored areas of the feat. space:
  documents that are new to the user but not unique
- Model Uncertainty
  label first where the model is uncertain:
  documents were the model is not sure of its prediction

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Active Learning Sampling

intelligently

How do I select critical data points to be labeled?

- Label Density : *Exploration*
- Model Uncertainty : *Exploitation*

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Exploration vs Exploitation

- Combines two ideas:
  - **explore** data points in very dense, unlabeled areas
  - **exploit** highly uncertain data points near class boundary

  Mixing label density with model uncertainty.

$$P(x_i) = (1 - \varepsilon)D(x_i) + \varepsilon R(x_i)$$

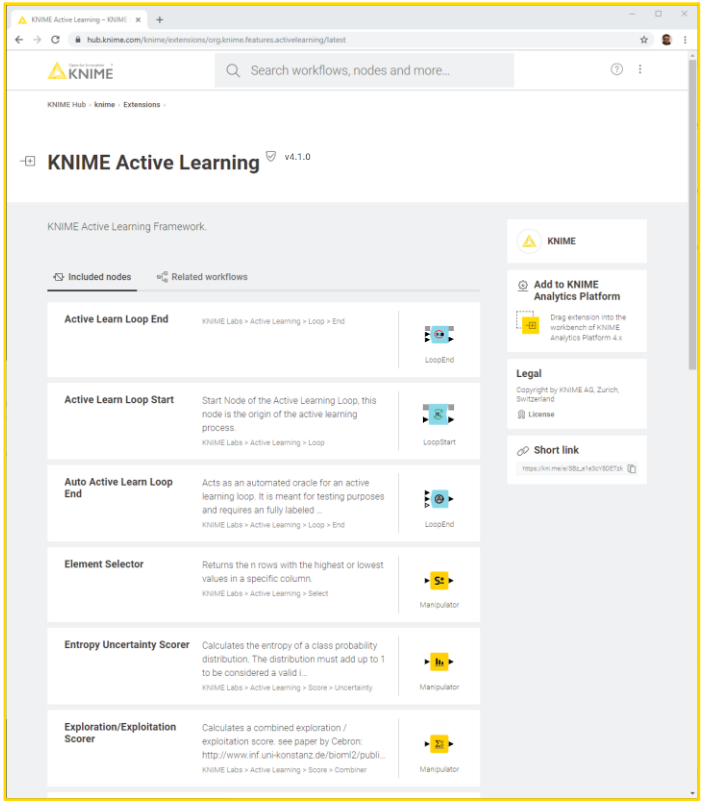
Estimate Potential

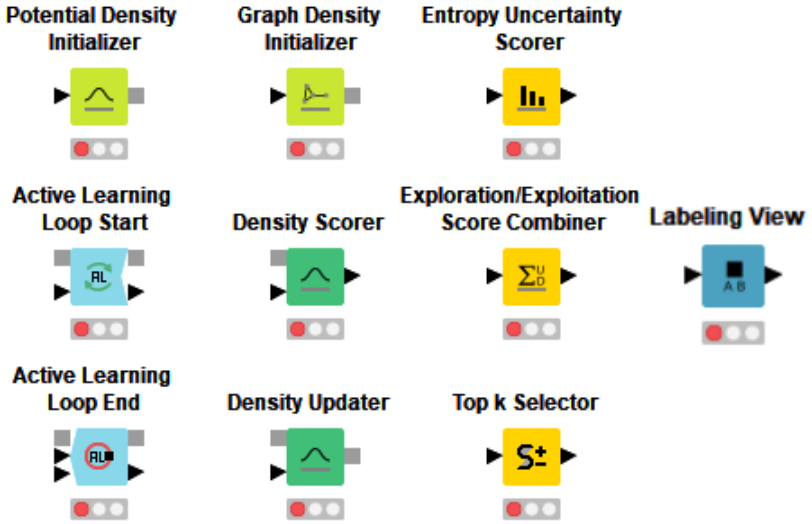
Active learning for object classification: from exploration to exploitation - Cebron et al, 2007

A Reinforced Active Learning Formulation for Object Class Recognition - Ebert et al, 2012

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Active Learning Extension with KNIME Software





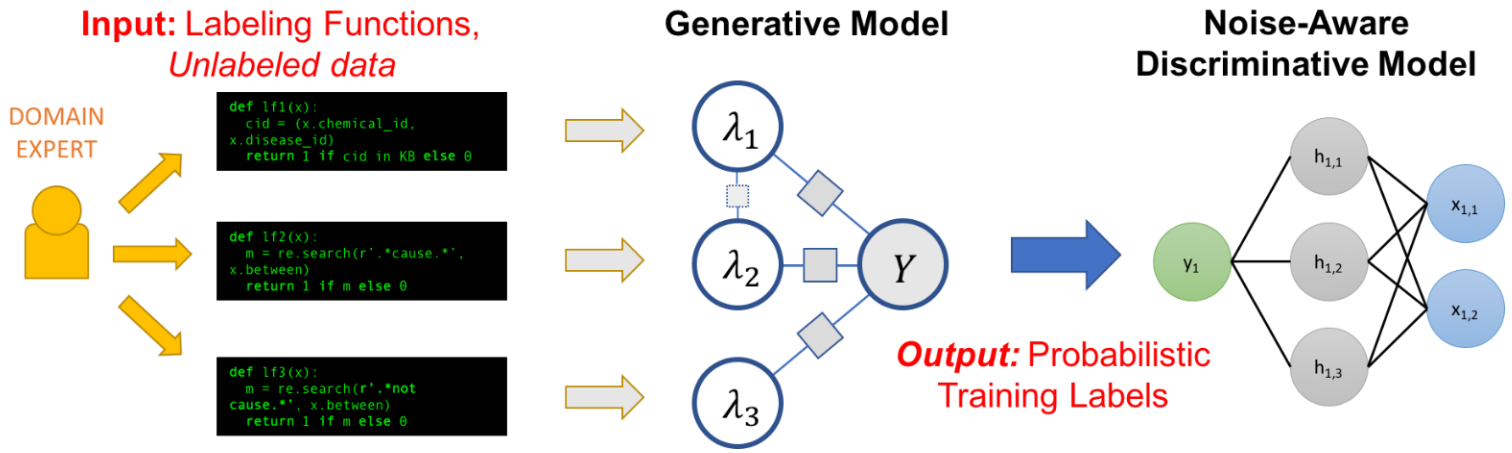
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Techniques to Train from Unlabeled Document Dataset

- Active Learning
  - Human-in-the-Loop Analytics
  - “Exploration vs Exploitation” Sampling
- Weak Supervision
  - Generative Model and Discriminative Model
  - Labeling Functions

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Weak Supervision



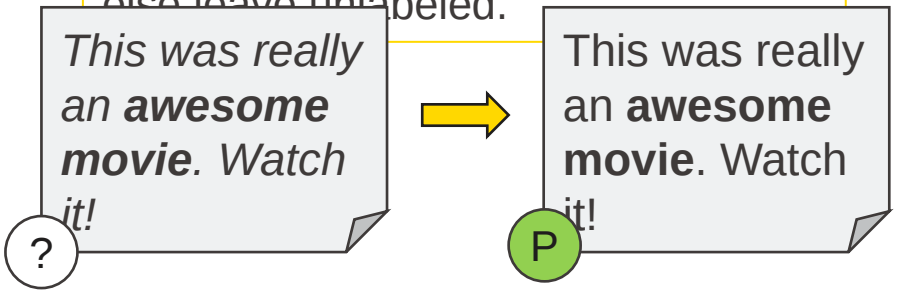
Weak Supervision: A New Programming Paradigm for Machine Learning
Alex Ratner et al, March 10, 2019, ai.stanford.edu/blog/weak-supervision/

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Weak Supervision: Labeling Function Example

Document Classification Example:

If “*awesome movie*” in movie review text
label as “**positive sentiment review**”,
else leave unlabeled.

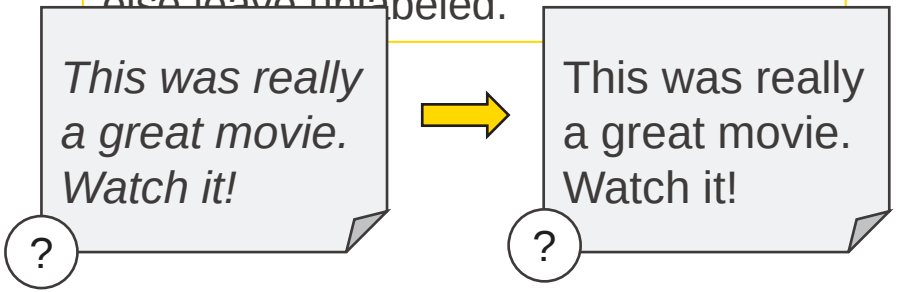


These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Weak Supervision: Labeling Function Example

Document Classification Example:

If “*awesome movie*” in movie review text
label as “**positive sentiment review**”,
else leave unlabeled.



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Weak Supervision: Labeling Function Example

Processed data - 0:429:0:427:440 - RowID

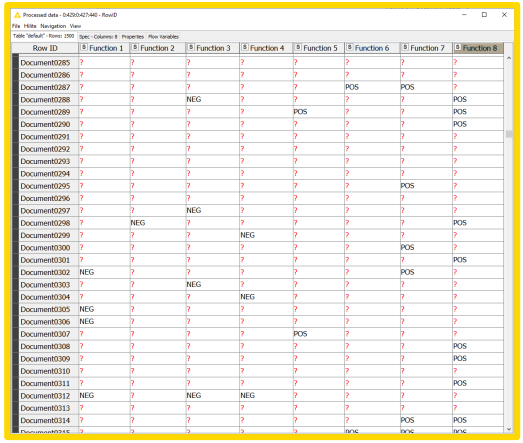
File Hilite Navigation View

Table "default" - Rows: 1500 | Spec - Columns: 8 | Properties | Flow Variables

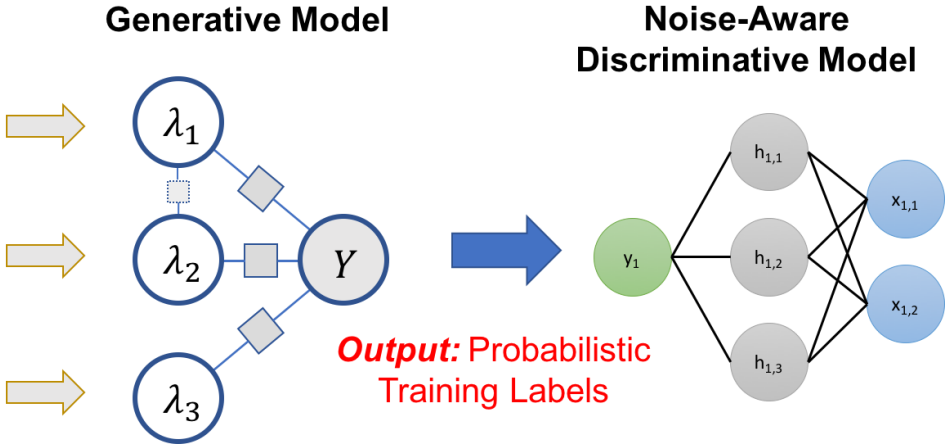
Row ID	Function 1	Function 2	Function 3	Function 4	Function 5	Function 6	Function 7	Function 8
Document0285	?	?	?	?	?	?	?	?
Document0286	?	?	?	?	?	?	?	?
Document0287	?	?	?	?	?	POS	POS	?
Document0288	?	?	NEG	?	?	?	?	POS
Document0289	?	?	?	?	POS	?	?	POS
Document0290	?	?	?	?	?	?	?	POS
Document0291	?	?	?	?	?	?	?	?
Document0292	?	?	?	?	?	?	?	?
Document0293	?	?	?	?	?	?	?	?
Document0294	?	?	?	?	?	?	?	?
Document0295	?	?	?	?	?	?	POS	?
Document0296	?	?	?	?	?	?	?	?
Document0297	?	?	NEG	?	?	?	?	?
Document0298	?	NEG	?	?	?	?	?	POS
Document0299	?	?	?	NEG	?	?	?	?
Document0300	?	?	?	?	?	?	POS	?
Document0301	?	?	?	?	?	?	?	POS
Document0302	NEG	?	?	?	?	?	POS	?
Document0303	?	?	NEG	?	?	?	?	?
Document0304	?	?	?	NEG	?	?	?	?
Document0305	NEG	?	?	?	?	?	?	?
Document0306	NEG	?	?	?	?	?	?	?
Document0307	?	?	?	?	POS	?	?	?
Document0308	?	?	?	?	?	?	?	POS
Document0309	?	?	?	?	?	?	?	POS
Document0310	?	?	?	?	?	?	?	?
Document0311	?	?	?	?	?	?	?	POS
Document0312	NEG	?	NEG	NEG	?	?	?	?
Document0313	?	?	?	?	?	?	?	?
Document0314	?	?	?	?	?	?	POS	POS
Document0315	?	?	?	?	?	POS	POS	POS

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Weak Supervision: Labeling Function Example



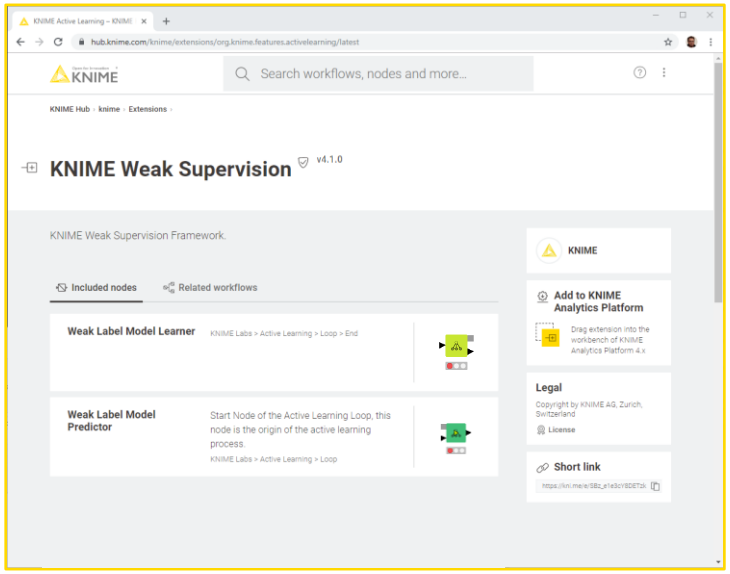
Row ID	Function 1	Function 2	Function 3	Function 4	Function 5	Function 6	Function 7	Function 8
Document0285	?	?	?	?	?	?	?	?
Document0286	?	?	?	?	?	?	?	?
Document0287	?	?	?	?	?	POS	POS	?
Document0288	?	?	NEG	?	?	?	?	POS
Document0289	?	?	?	?	POS	?	?	POS
Document0290	?	?	?	?	?	?	?	POS
Document0291	?	?	?	?	?	?	?	?
Document0292	?	?	?	?	?	?	?	?
Document0293	?	?	?	?	?	?	?	?
Document0294	?	?	?	?	?	?	?	?
Document0295	?	?	?	?	?	?	POS	?
Document0296	?	?	?	?	?	?	?	?
Document0297	?	?	NEG	?	?	?	?	?
Document0298	?	NEG	?	?	?	?	?	POS
Document0299	?	?	?	NEG	?	?	?	?
Document0300	?	?	?	?	?	?	POS	?
Document0301	?	?	?	?	?	?	?	POS
Document0302	NEG	?	?	?	?	?	POS	?
Document0303	?	?	NEG	?	?	?	?	?
Document0304	?	?	?	NEG	?	?	?	?
Document0305	NEG	?	?	?	?	?	?	?
Document0306	NEG	?	?	?	?	?	?	?
Document0307	?	?	?	?	POS	?	?	?
Document0308	?	?	?	?	?	?	?	POS
Document0309	?	?	?	?	?	?	?	POS
Document0310	?	?	?	?	?	?	?	?
Document0311	?	?	?	?	?	?	?	POS
Document0312	NEG	?	NEG	NEG	?	?	?	?
Document0313	?	?	?	?	?	?	?	?
Document0314	?	?	?	?	?	?	POS	POS

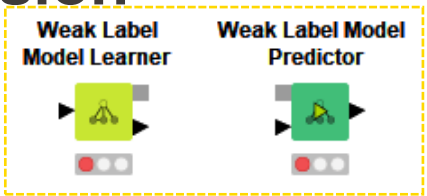


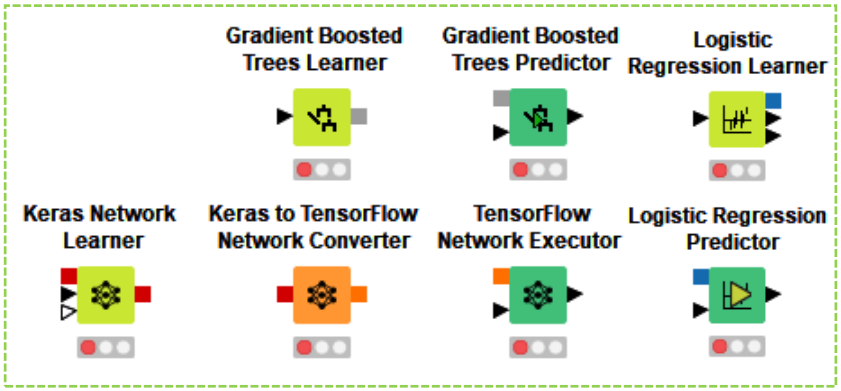
Weak Supervision: A New Programming Paradigm for Machine Learning
Alex Ratner et al, March 10, 2019, ai.stanford.edu/blog/weak-supervision/

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

KNIME Weak Supervision Extension

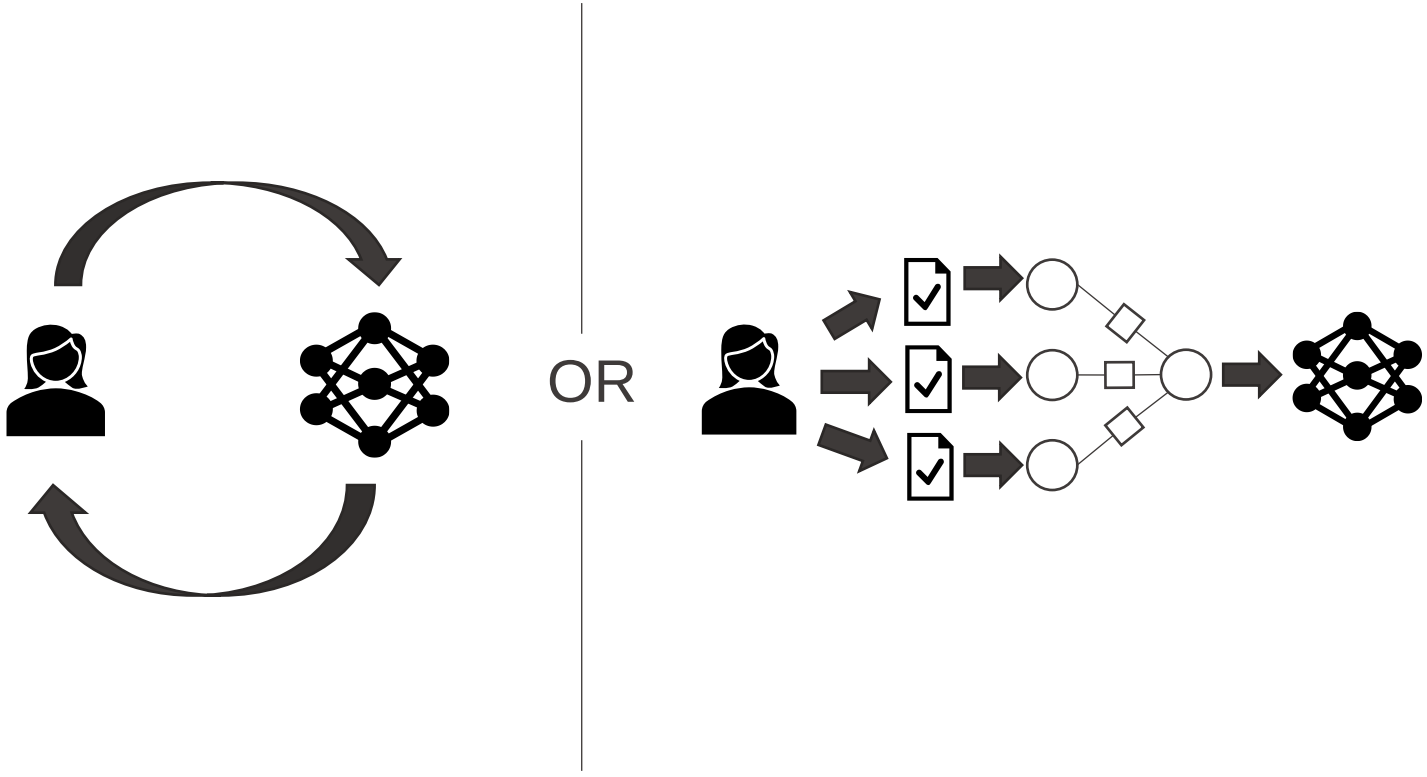






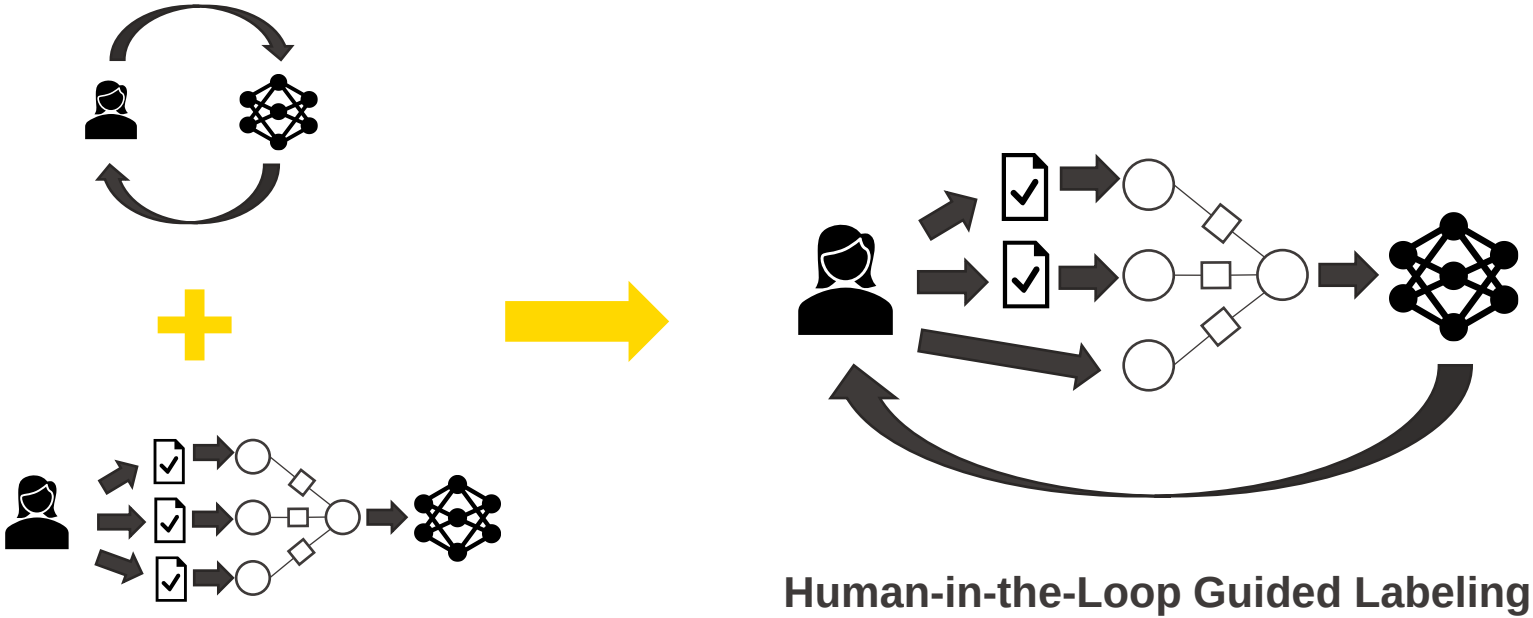
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Active Learning vs Weak Supervision



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

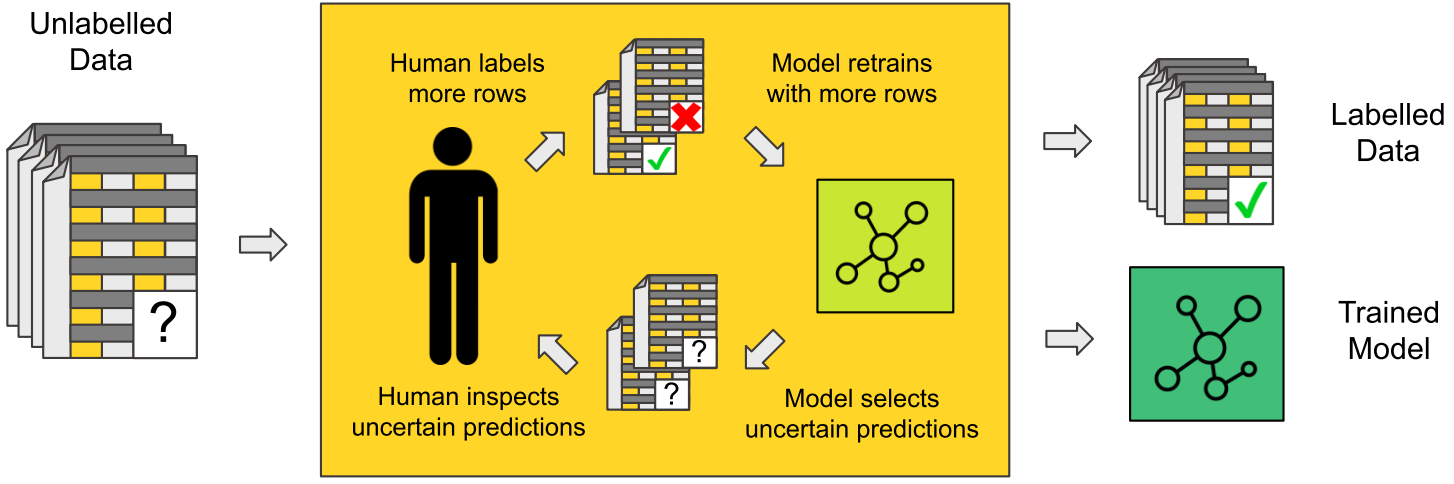
Active Learning + Weak Supervision



Human-in-the-Loop Guided Labeling: Single Interactive Web Based Application

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

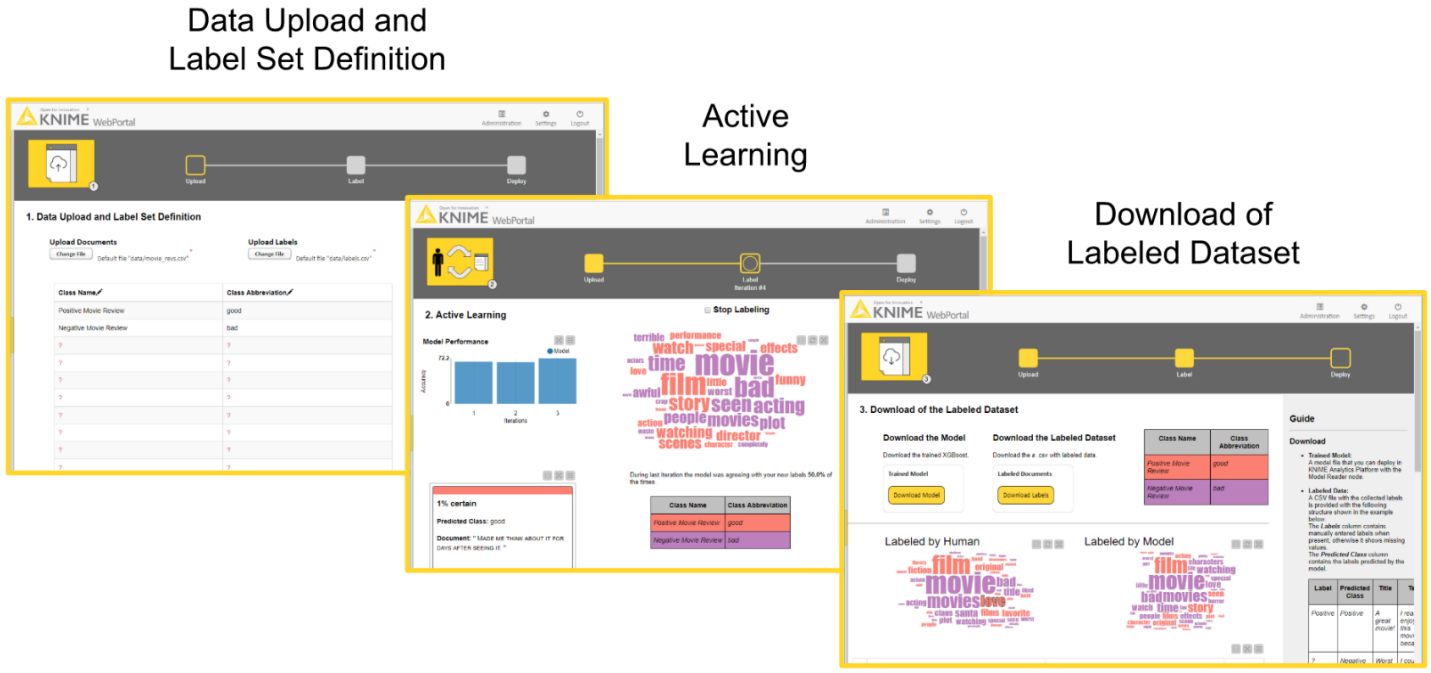
Guided Labeling for Document Classification



Labeling with Active Learning, Paolo Tamagnini and Rosaria Silipo, August 29, 2019, datasciencecentral.com/profiles/blogs/labeling-with-active-learning

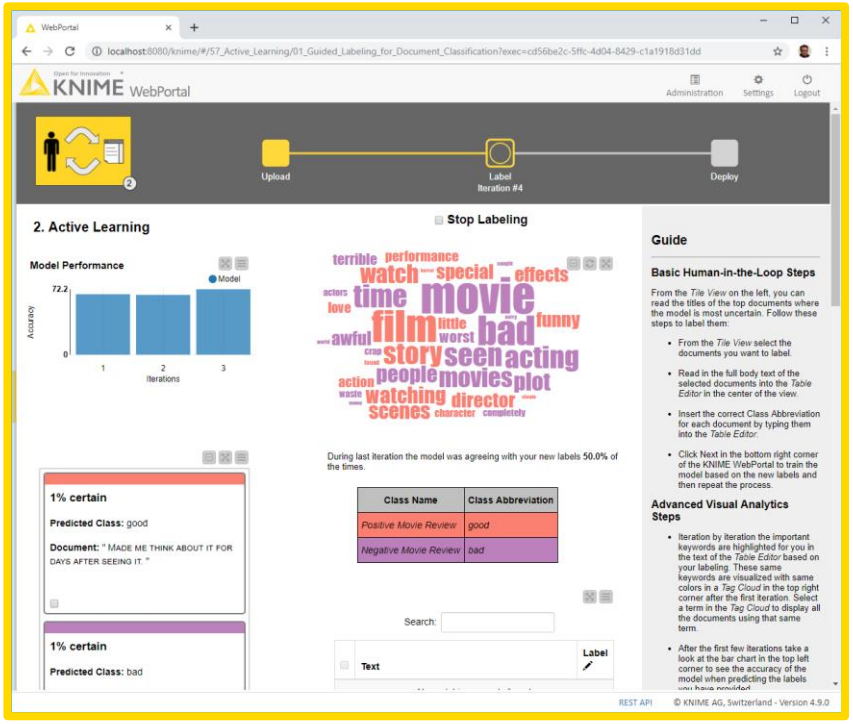
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Human-in-the-Loop Guided Labeling via KNIME WebPortal



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

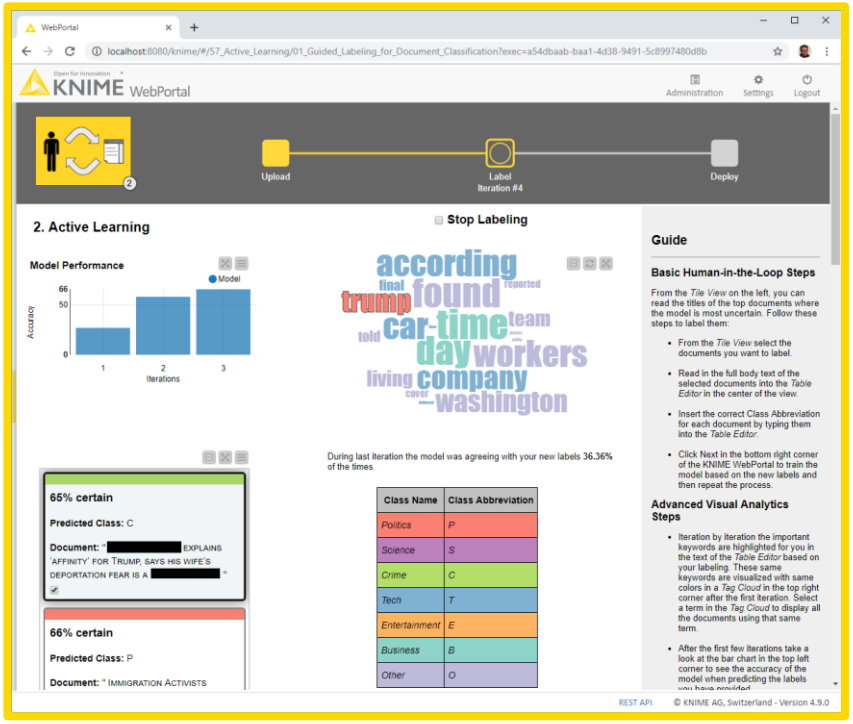
More Use Cases: Sentiment Analysis



- Documents : social media
- Labels are sentiment
- Binary classification
- Tag cloud shows sentiment

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

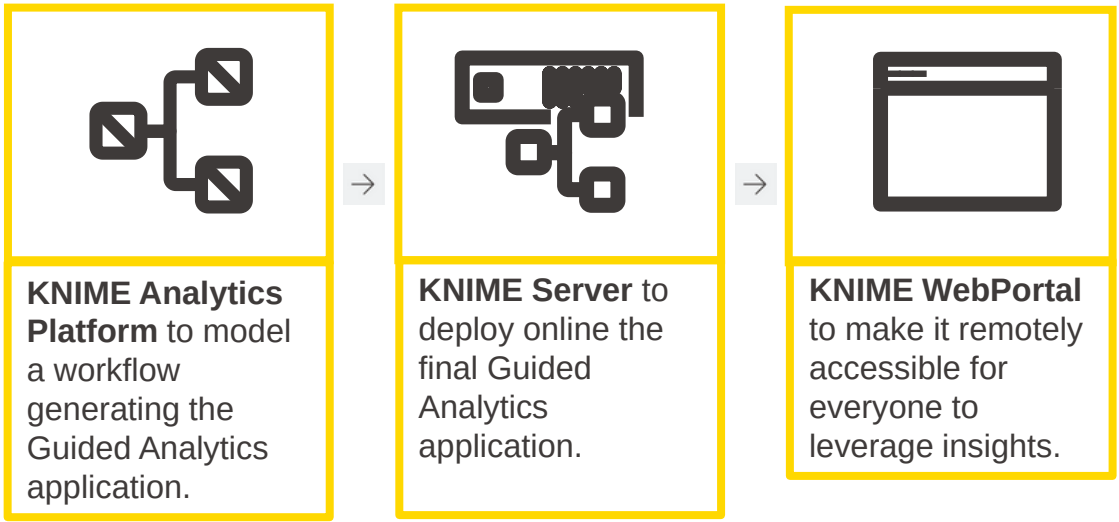
More Use Cases: Topic Detection



- Documents : news articles
- Labels are topics
- Multi-class classification
- Tag cloud shows low hanging fruits

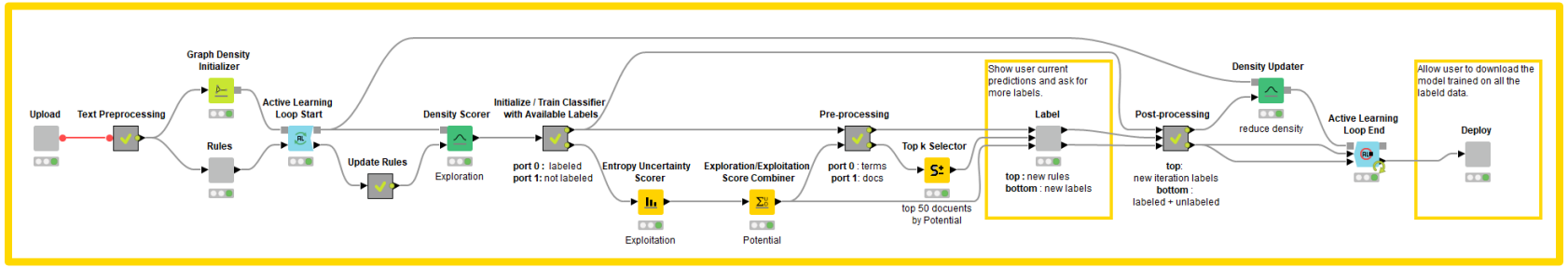
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

How KNIME Software works with Guided Analytics



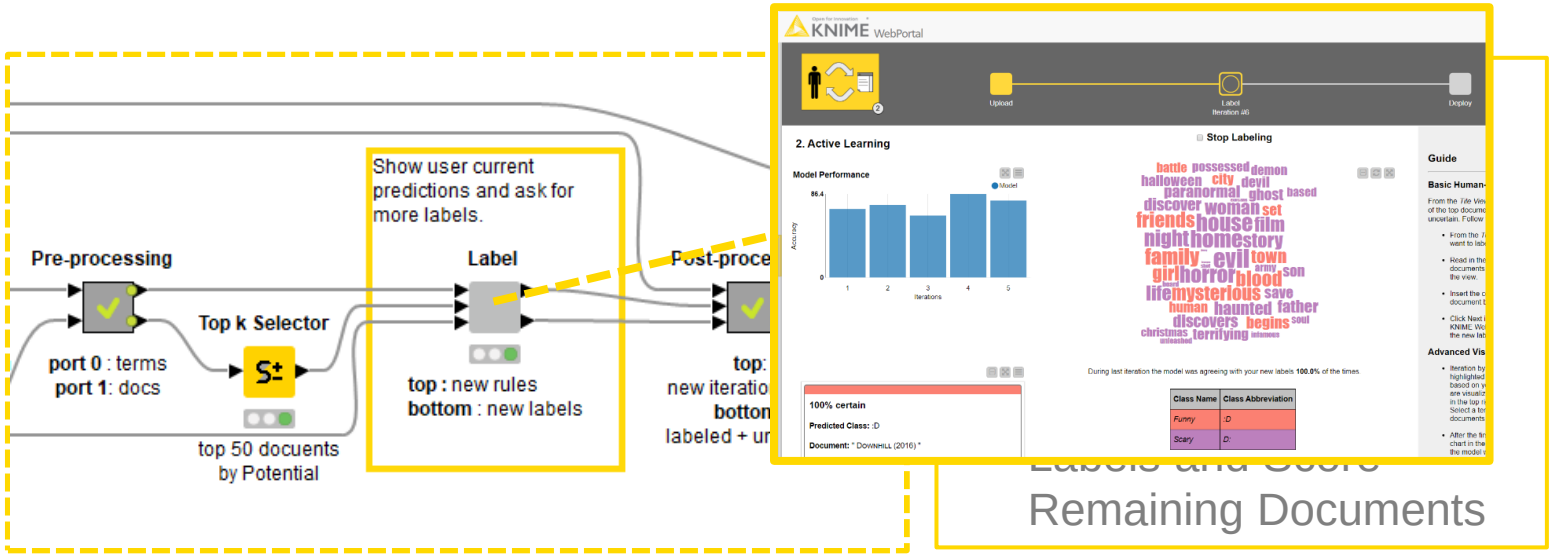
These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

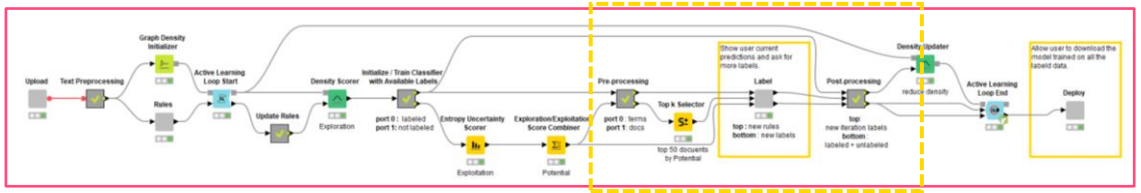
The Guided Labeling Workflow



These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

The Guided Labeling Workflow





These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

On the KNIME Hub

tinyurl.com/GuidedLabeling

- Go on **hub.knime.com**
- type in *Guided Labeling*
- Select the workflow
- Download and Open in KNIME (and leave a nice comment below)

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Stay Connected with KNIME

Blog: knime.com/blog

Forum:
forum.knime.com

KNIME Hub:
hub.knime.com

Follow us on social media:

Join us at a KNIME Meetup near you:

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

Data Connect Italy – Live from Bologna

Parte di **KNIME – 11 gruppi**

Italy KNIME Users

Milano, Italia

1.273 membri · Gruppo pubblico

Organizzato da **KNIME Event C.** e **altri 6**

Condividi:

meetup.com/it-IT/italy-knime-users/events/290474652/

Eventi in programma (2) Visualizza tutti

mer 8 feb 2023, 17:00 CET

Data Connect: Italia (Onsite/Hybrid in Bologna)

Fondazione ANT Italia Onlus

Partecipa alla seconda edizione del Data Connect: Italia a Bologna. Questo evento si terrà presso la Fondazione ANT Italia ONLUS.
Se partecipi all'evento online, registrati qui....

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0

From Words to Wisdom Book

Free Copy of “From Words to Wisdom” Book from KNIME Press

knime.com/knimepress

with code:
UF2-0223 *

Thank You!

(* also valid for any other book)

These slides are a derivative of KNIME Course Material of KNIME AG used under CC BY 4.0