Text Mining Webinar

The Textprocessing Extension

Rosaria Silipo and Kilian Thiel

Agenda

- Text Mining Goals and Usage
- Enrichment & Preprocessing
- Data Types & Structures
- Visualization
- Topic Detection
- Sentiment Analysis

Install TextProcessing Extension

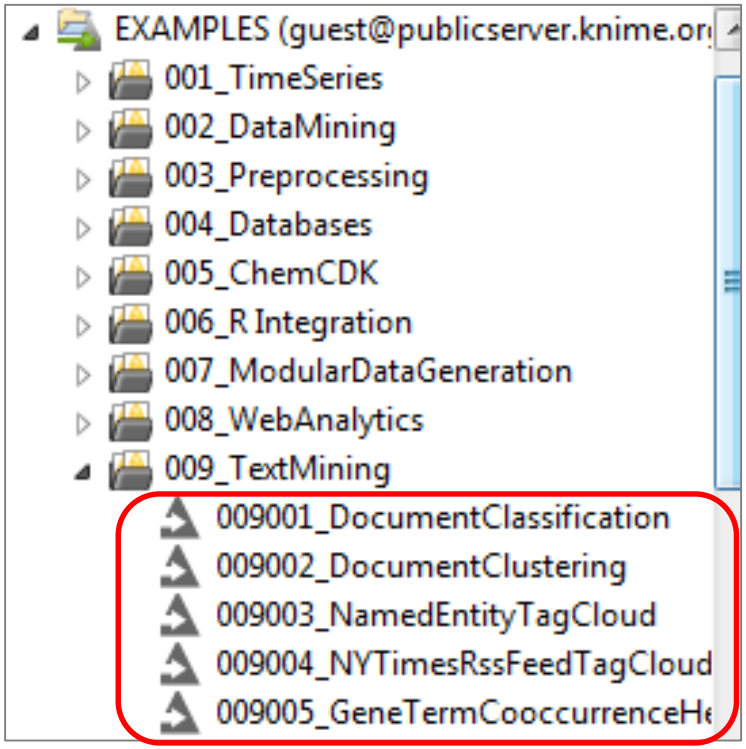
KNIME:

www.knime.org

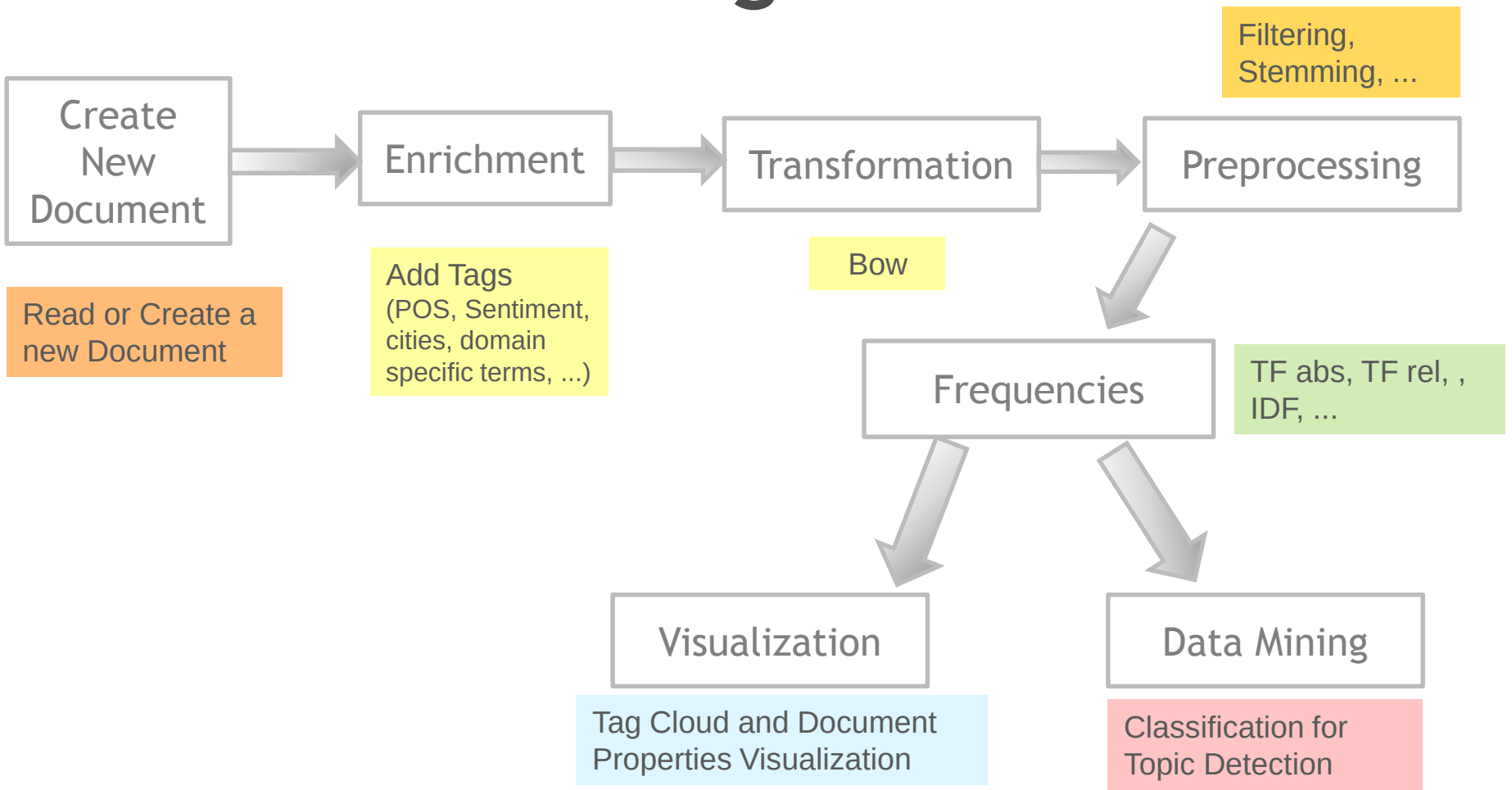
Install Textprocessing Extension
under KNIME Labs

Examples

Example Workflows available on the KNIME public server.



Text Mining Workflow

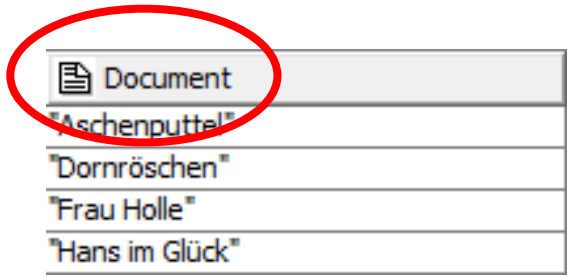


1 - Create a Document

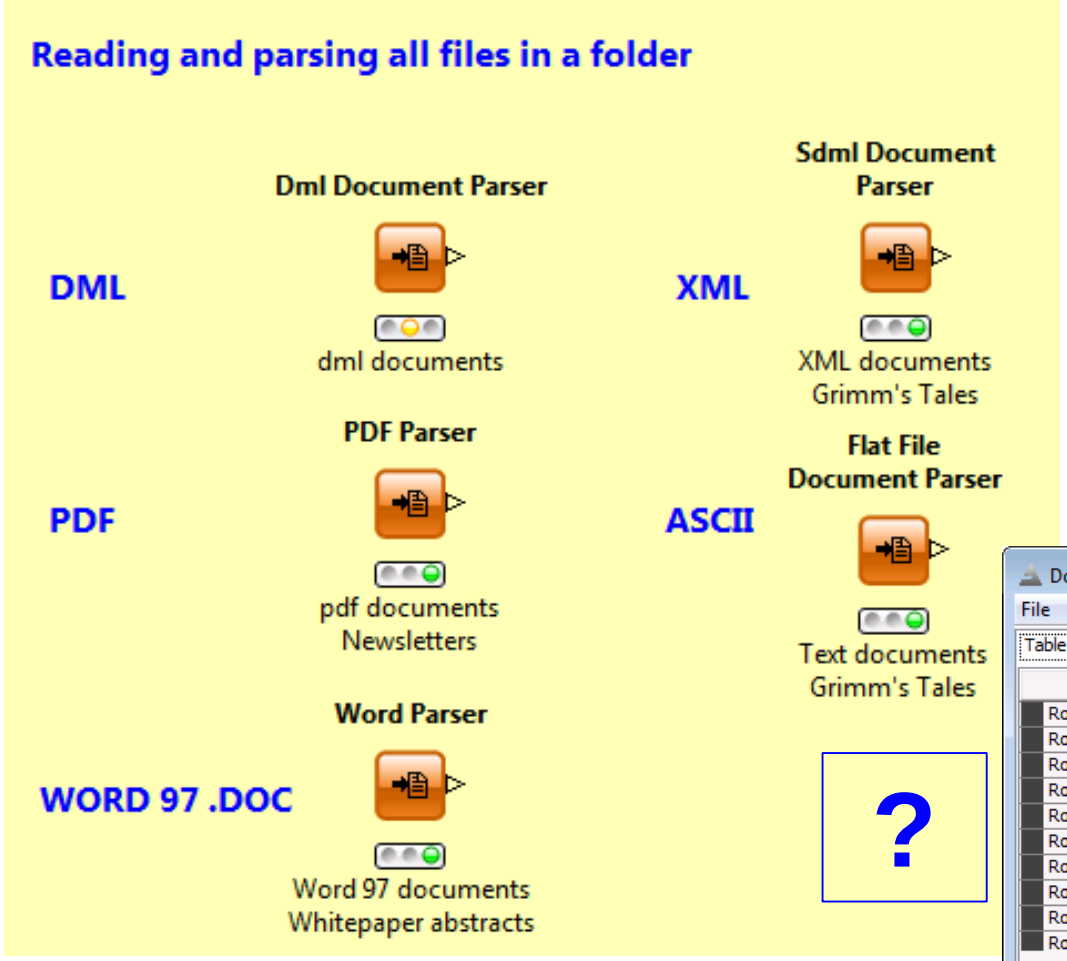
New Data Types

Document

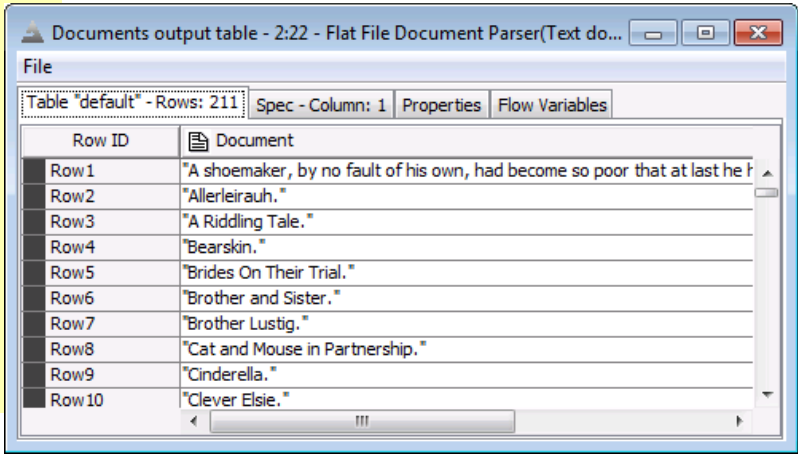
Encapsulates text, author, title, source, category, and type



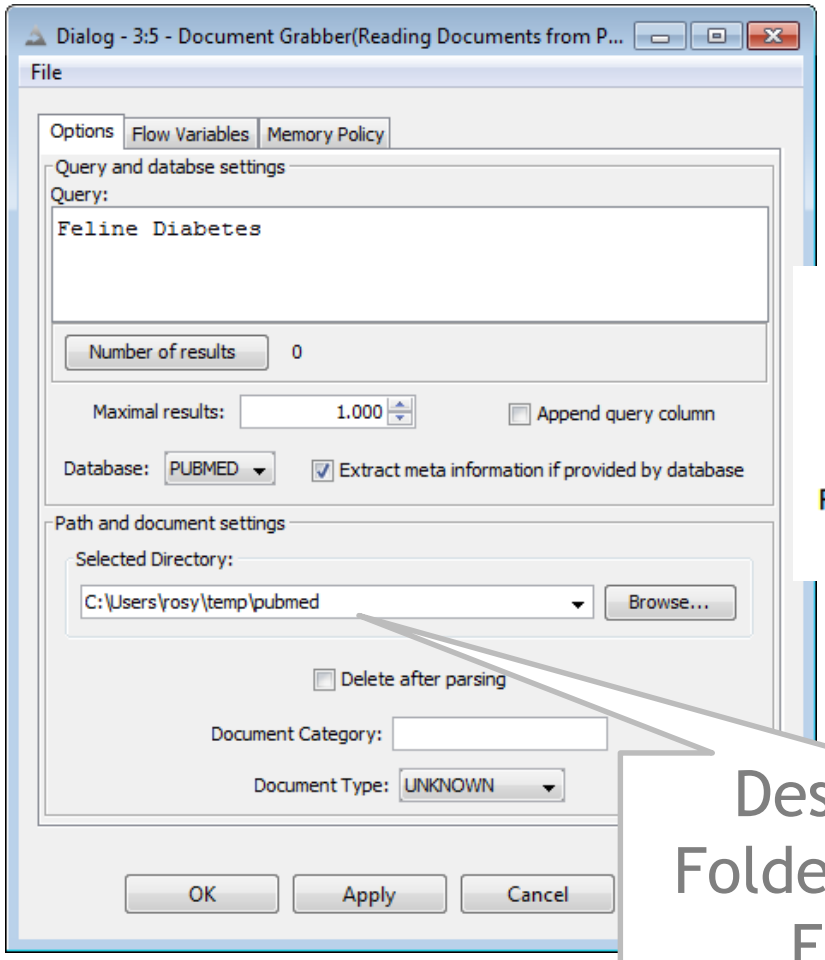
From a Folder

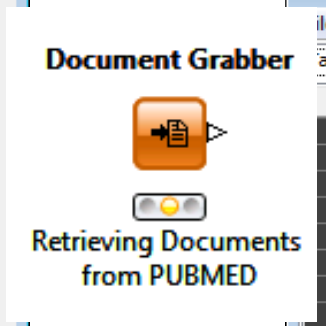


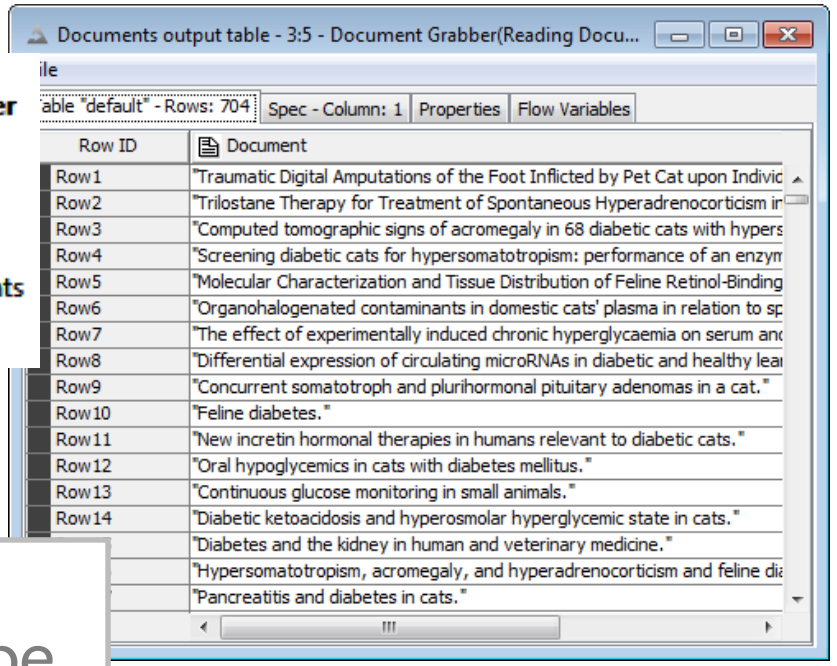
The output is a list of Documents



From PUBMED



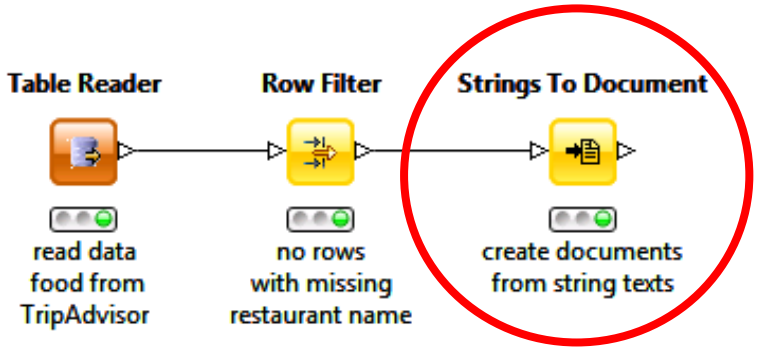




The output is a list of Documents

Destination Folder MUST be EMPTY!

Strings to Document

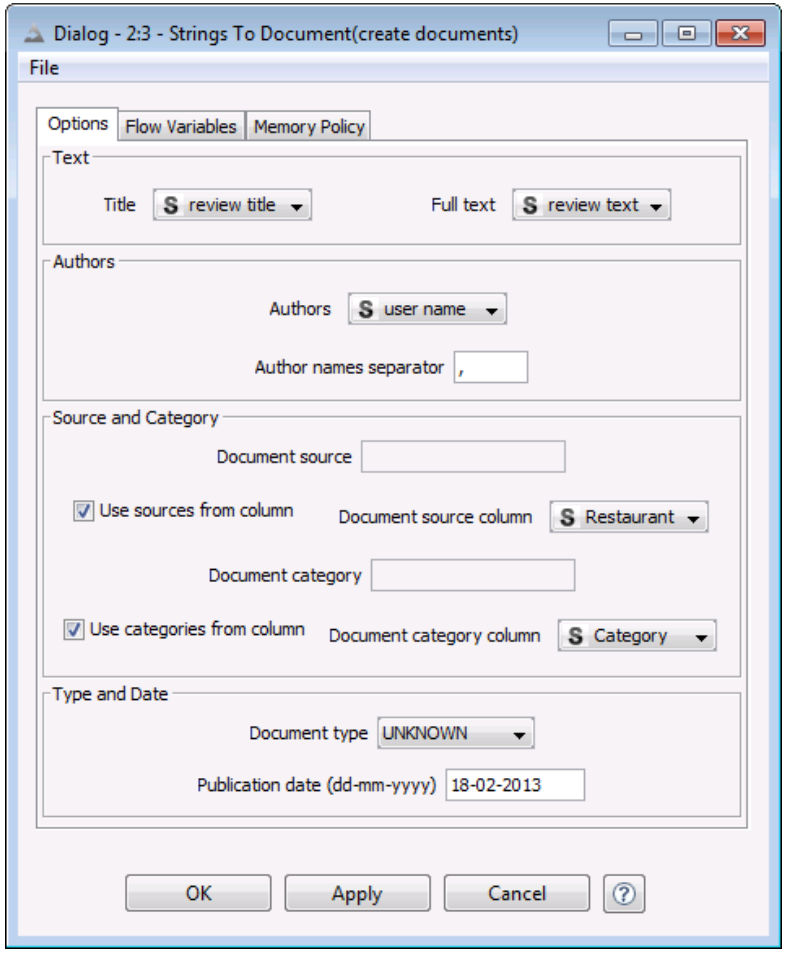


String and document output table - 2:3 - Strings To Do...

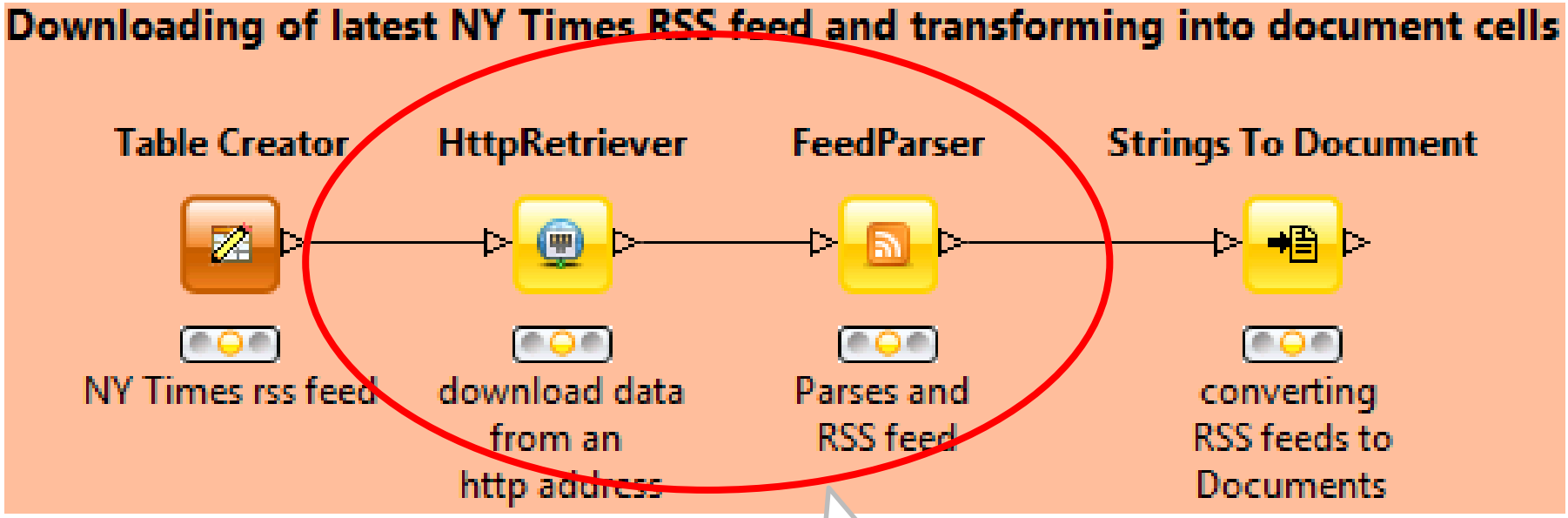
File

Table "default" - Rows: 440 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	Restaurant	Categor	Document
Row0_1_Row...	gon and m...	Asian	"Great food, interesting servi...
Row0_2_Row...	gon and m...	Asian	"Excellent Lunch Destination"
Row0_3_Row...	gon and m...	Asian	"Hidden treasure near KaDaWe"
Row0_4_Row...	gon and m...	Asian	"Excellent Food Very Reason...
Row0_5_Row...	gon and m...	Asian	"Good food, great prices!"



From RSS Feeds



http://feeds.nytimes.com/nyt/rss/World

Palladian Nodes

The Data Set

Reviews of Restaurants in Berlin from TripAdvisor

Self-downloaded with RSS Feeder

Read table - 0:2 - Table Reader(read data)

File

Table "default" - Rows: 449 | Spec - Columns: 7 | Properties | Flow Variables

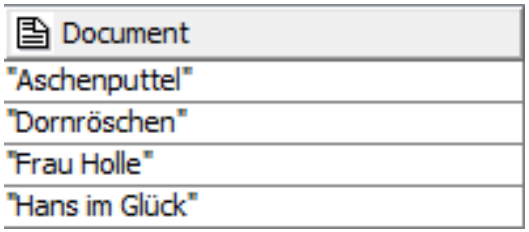
Row ID	user name	review title	review text	stars	Reviews	Restaurant	Category
Row0_1_Row...	1travellerBruss...	Great food, interesting ser...	this restaurant is ...	5	29	Saigon and m...	Asian
Row0_2_Row...	coverdriven	Excellent Lunch Destination	Very much enjoy...	4	2	Saigon and m...	Asian
Row0_3_Row...	Lula12783	Hidden treasure near KaDa...	We found this littl...	5	5	Saigon and m...	Asian
Row0_4_Row...	Deise_Boy08	Excellent Food Very Reaso...	Food was top cla...	5	12	Saigon and m...	Asian
Row0_5_Row...	Tanguyatea	Good food, great prices!	I went there bec...	4	6	Saigon and m...	Asian
Row0_6_Row...	OlgaNottingham	Nice food at a reasonable ...	I am no expert o...	4	12	Saigon and m...	Asian
Row0_7_Row...	Jack D	Good food and entertainin...	From reading oth...	4	3	Saigon and m...	Asian
Row0_8_Row...	Rick M	Very good	A very tasty Viet...	4	6	Saigon and m...	Asian

2 - Enrichment

New Data Types

Document

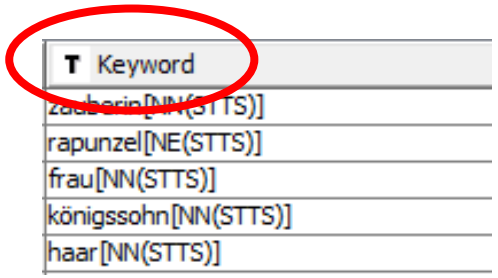
Encapsulates text, author, title, source, category, and type



Document
"Aschenputtel"
"Dornröschen"
"Frau Holle"
"Hans im Glück"

Term

Encapsulates a term



Keyword
zauberin[NN(STTS)]
rapunzel[NE(STTS)]
frau[NN(STTS)]
königssohn[NN(STTS)]
haar[NN(STTS)]

Enrichment (Tagging)

Enrichment nodes (mostly) change the granularity of terms.

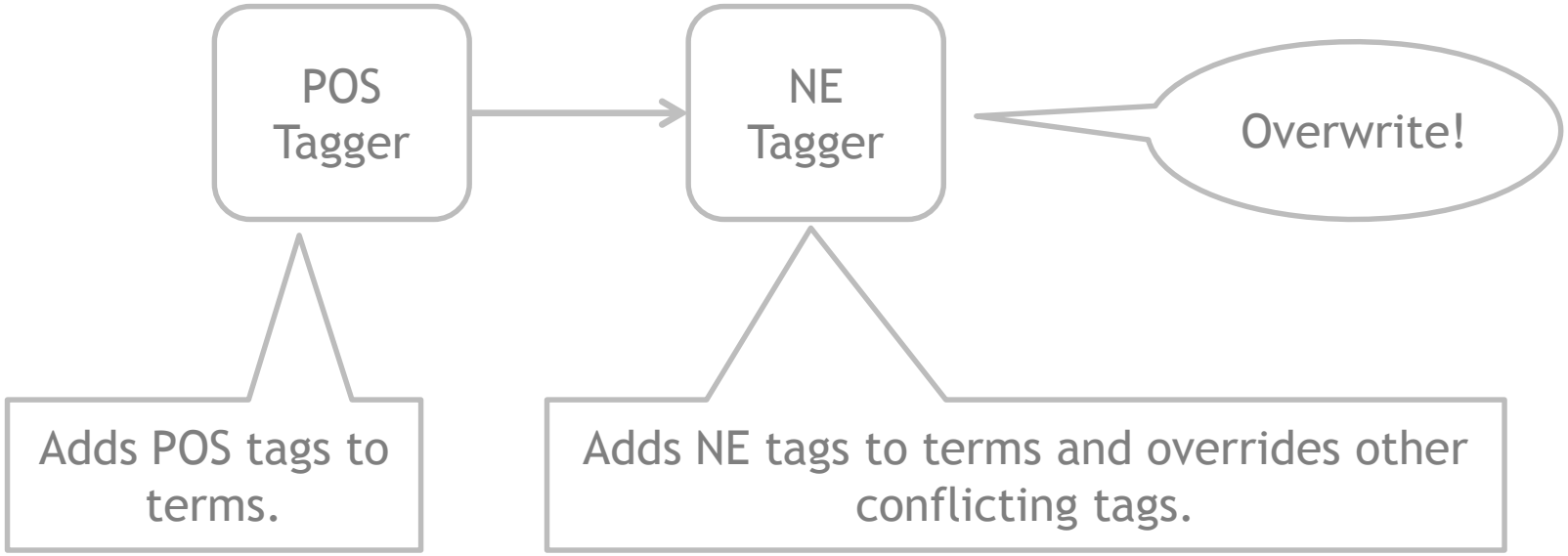
- Multiword detection, named entity recognition, part of speech definition, …
- To each detected entity (term) a tag is added, specifying its type.
- To avoid intersection of granularity the last node dominates.

Tagger Conflict Resolution

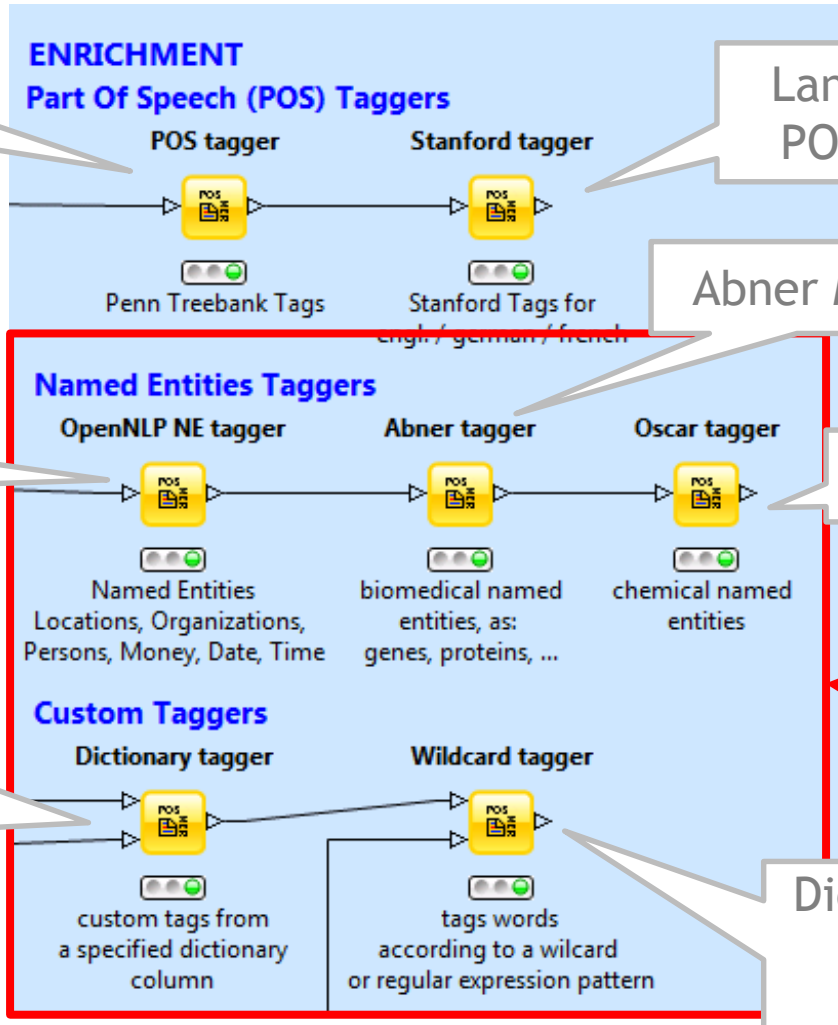
In case of intersections of granularity the last node overwrites.

Example: *"The gene interleukin 6 interacts ...."*

1. POS tagger: "The\DT gene\NN interleukin\NN 6\CD interacts \VBZ "
2. NE tagger: "The\DT gene\NN interleukin 6\GENE interacts \VBZ "



Tagging



No settings

Language + POS Model

Abner Model

Named Entity

No settings

Named Entity Tags can be set as unmodifiable

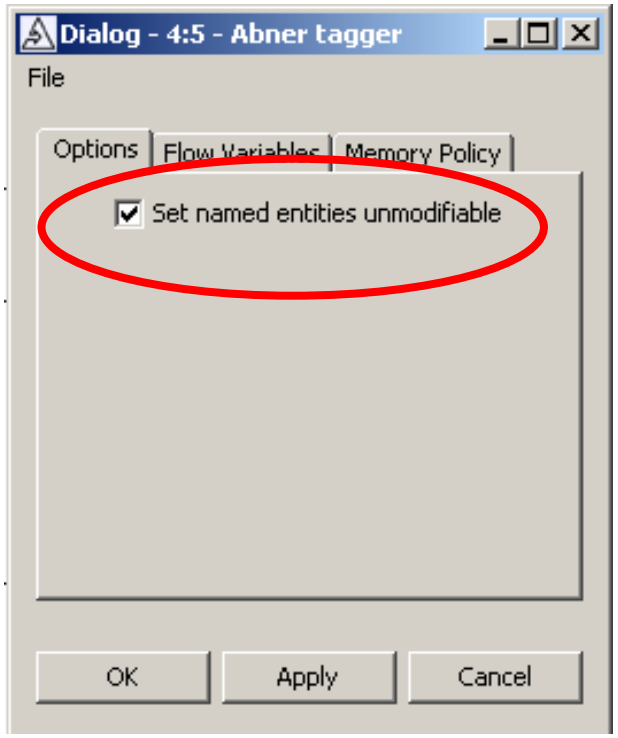
Dictionary Column
Tag Type
Tag Value

Dictionary Column
Tag Type
Tag Value
Matching Strategy

Unmodifiable Named Entity Tags

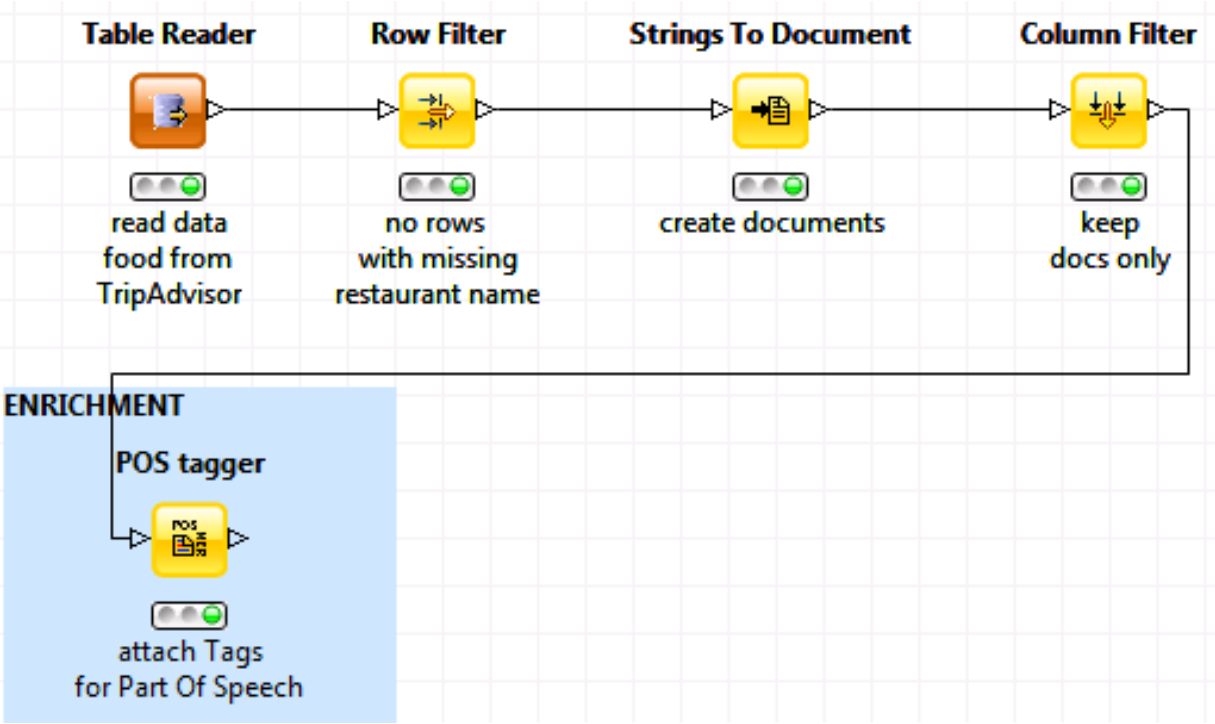
Named Entities Tags attached through enrichment nodes can be set as unmodifiable

Unmodifiable Tags are not affected by any preprocessing nodes (stemming, filtering, etc.)



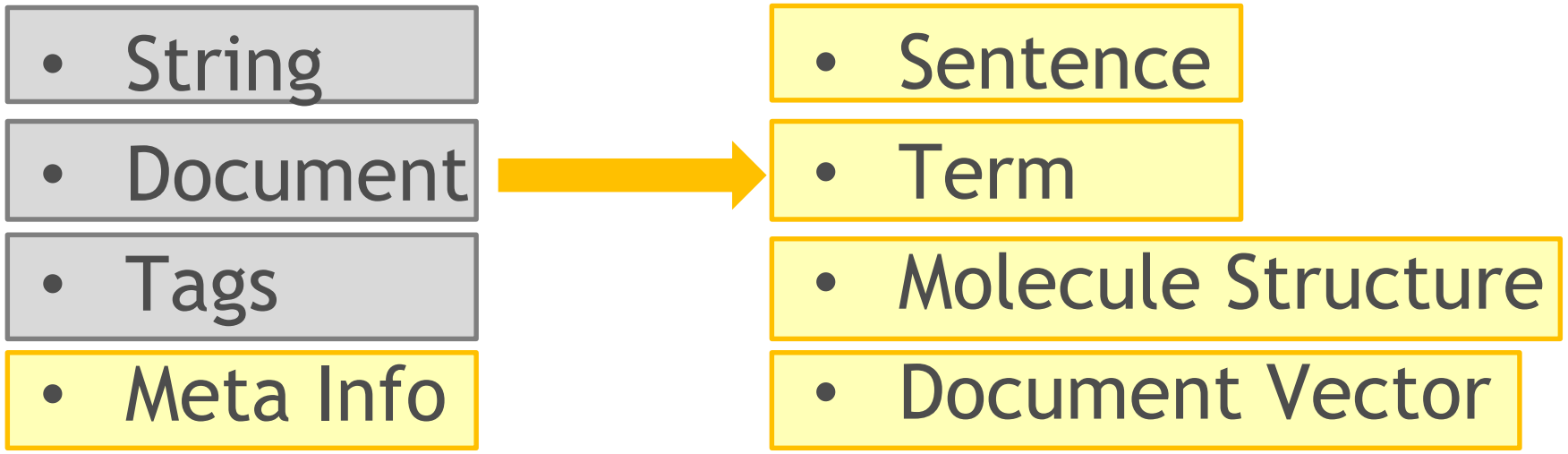
Workflow

Strings to Document Workflow + POS Tagger



3 - Transformation

Data Types and Features



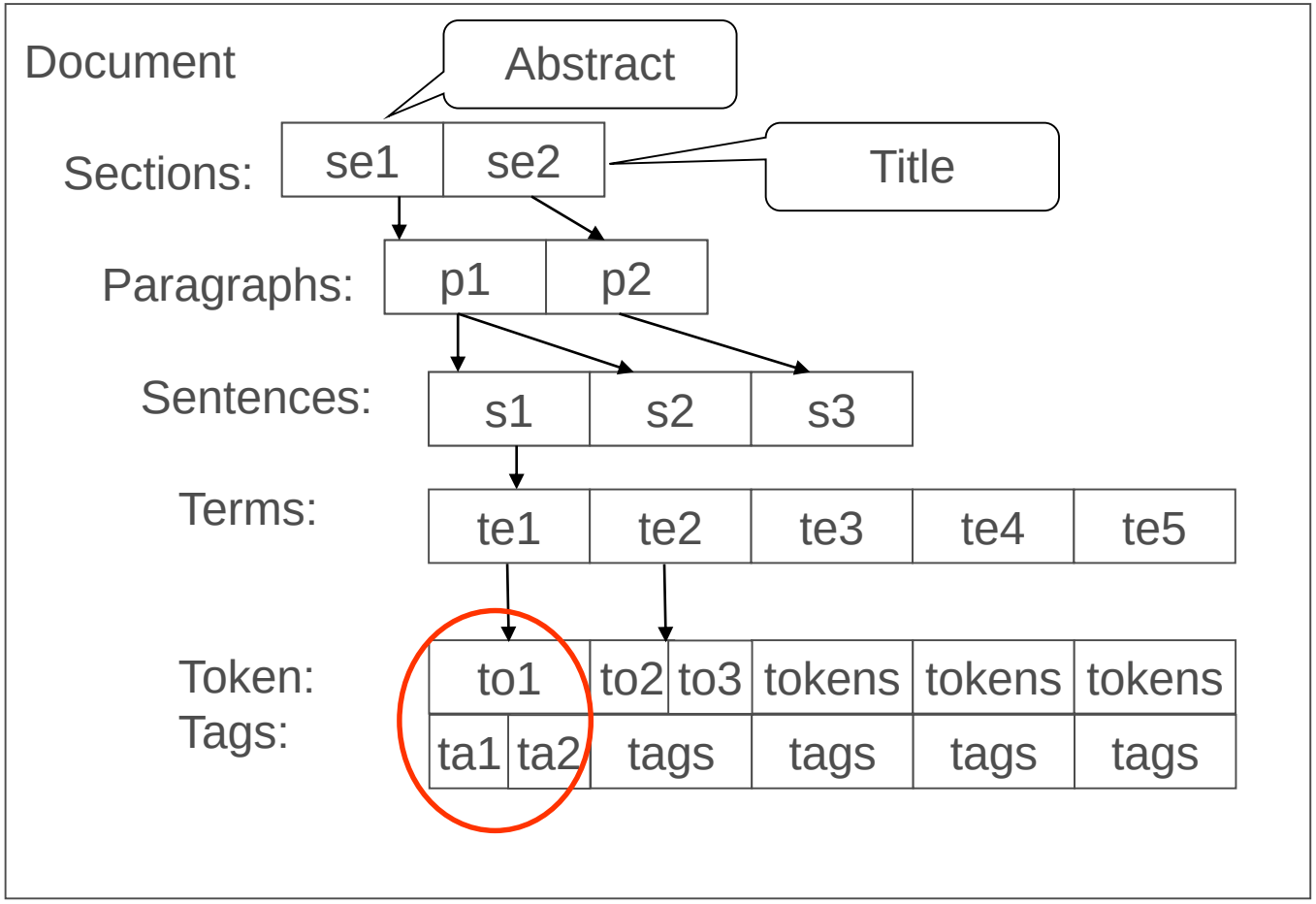
Parsing / Tokenization

Parser nodes parse documents by applying standard tokenization via OpenNLP tokenizer.

Each token is a term consisting of a single word.

Tags are applied to terms.

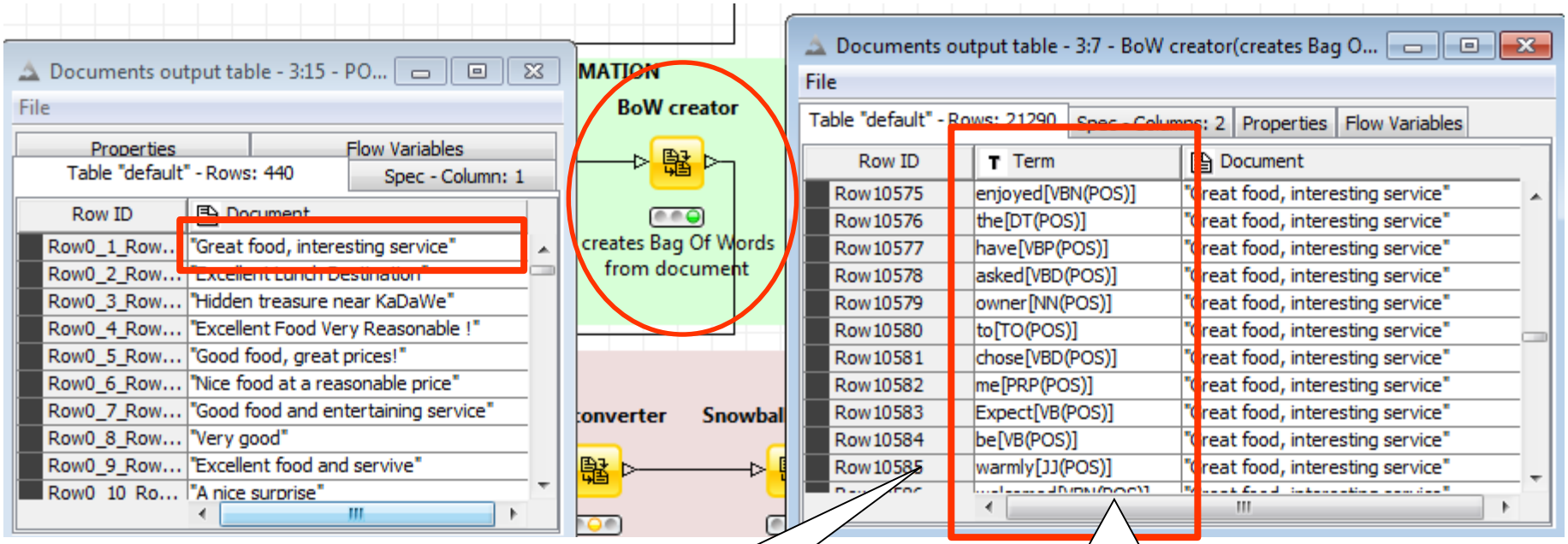
Parsing



Annotations label a section part in the document (like “Abstract”, “Title”, etc.)

Terms consist of tokens and tags

The Bag of Words

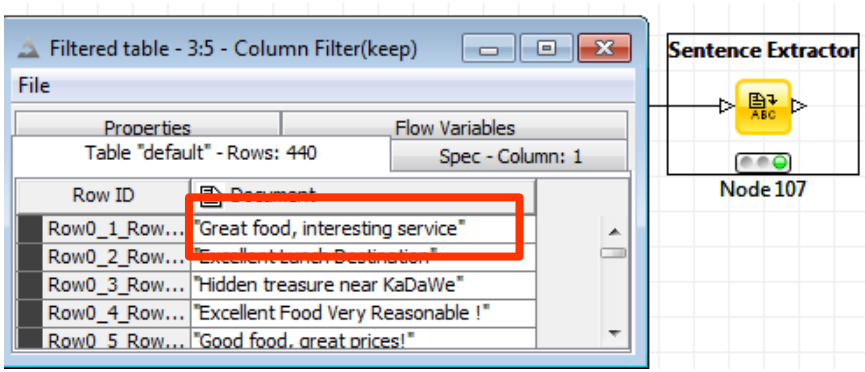


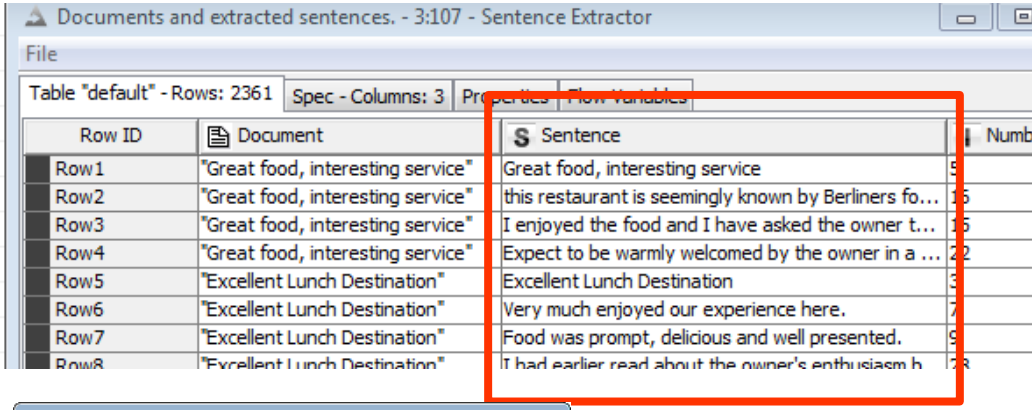
No Settings required

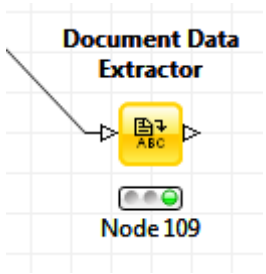
Each Term is extracted with Tags (NN, VB, …)

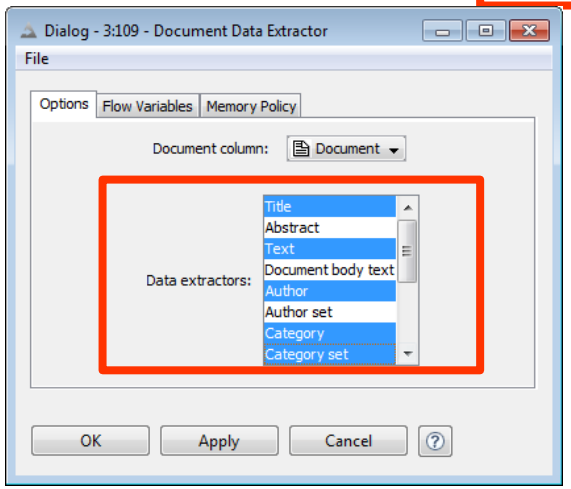
List (Bag) of Terms (Words) identified in Document

Data and Sentence Extractor

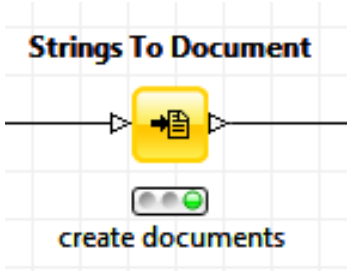


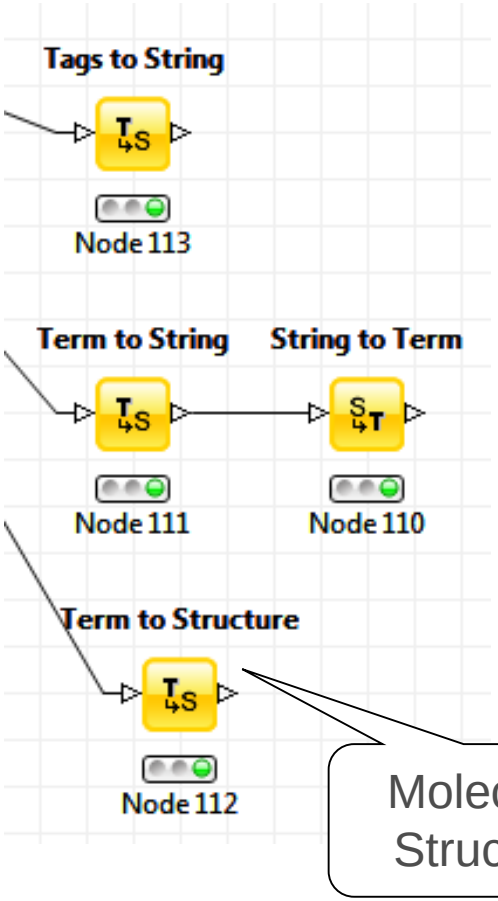






Conversions





Workflow

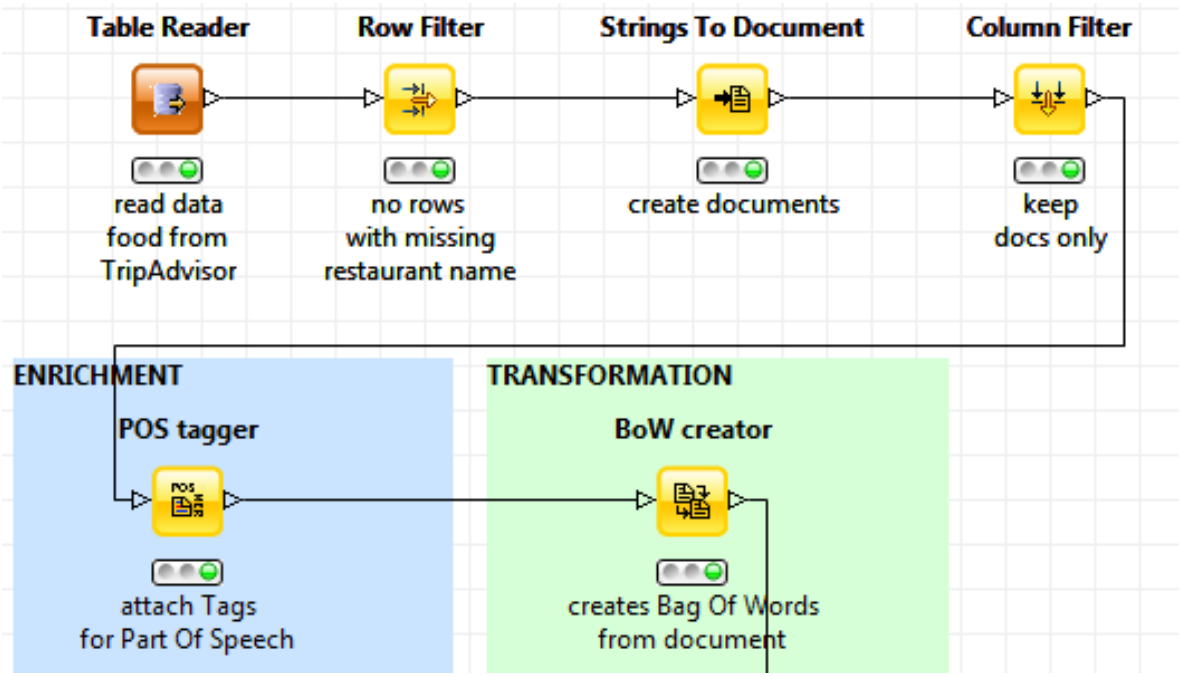


Table Reader
Row Filter
Strings To Document
Column Filter
read data food from TripAdvisor
no rows with missing restaurant name
create documents
keep docs only
ENRICHMENT
POS tagger
attach Tags for Part Of Speech
TRANSFORMATION
BoW creator
creates Bag Of Words from document

4 - Preprocessing

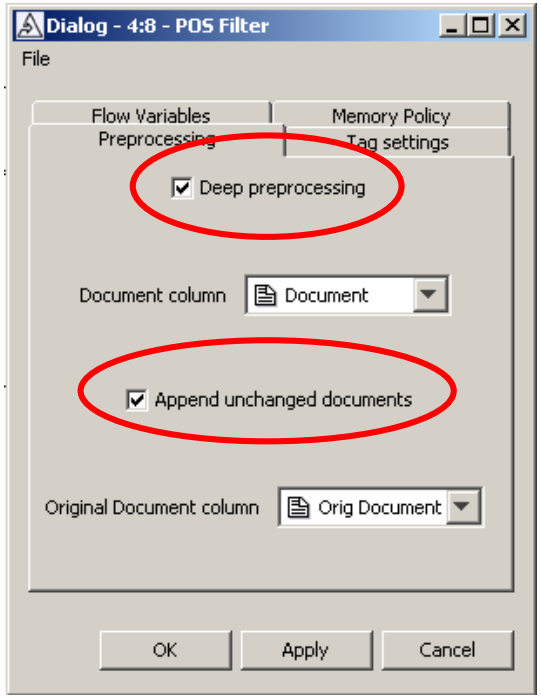
Preprocessing Mode

Normal

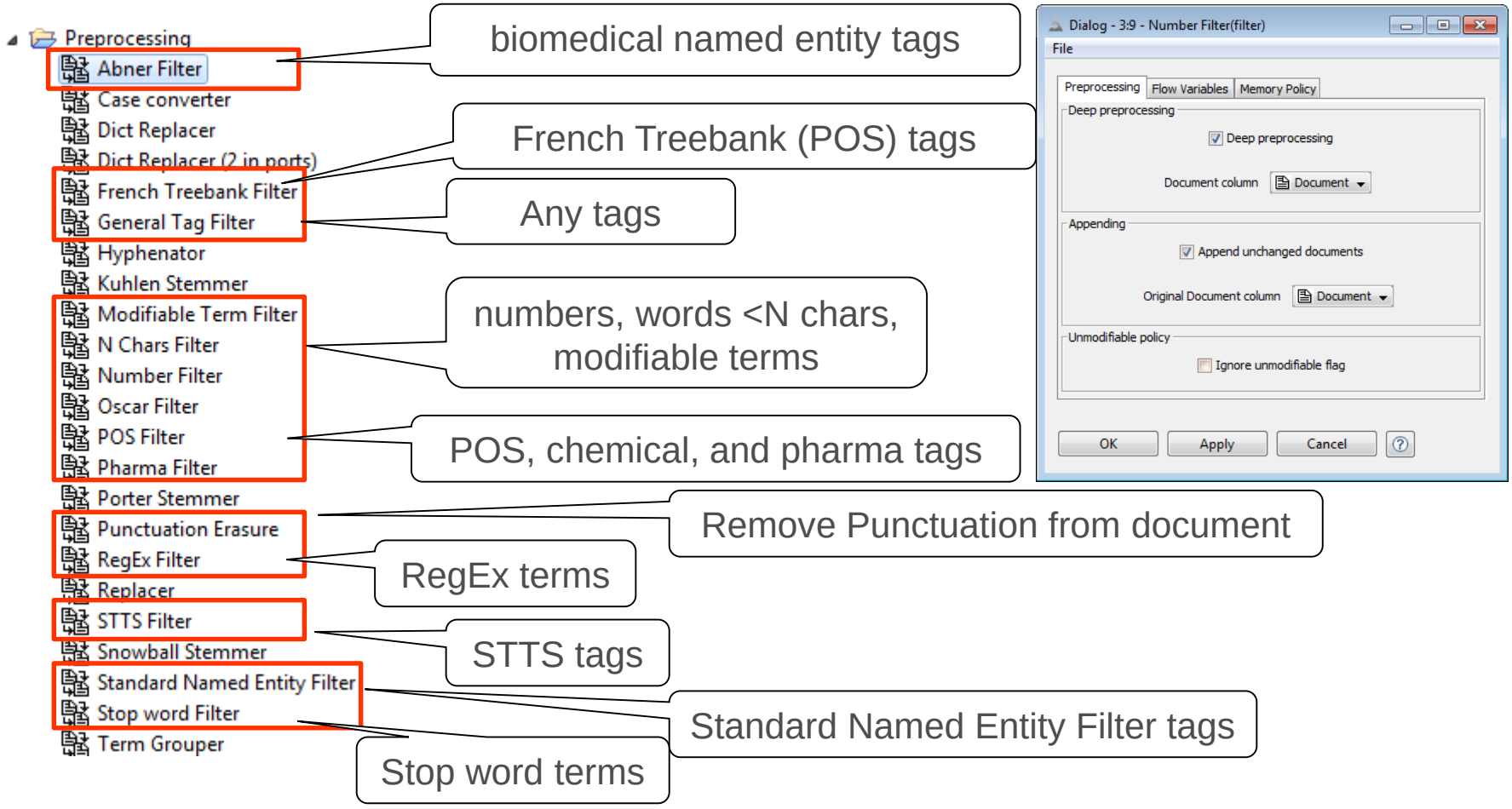
- Faster
- Preprocesses only terms of the term column
- Documents are not changed

Deep

- Slower
- Preprocesses terms of the term column
- Terms in documents are changed as well
- Unchanged documents can be appended



Filtering by Tags and Terms



Converting and Replacing

- Preprocessing
  - Abner Filter
  - Case converter
  - Dict Replacer
  - Dict Replacer (2 in ports)
  - French Treebank Filter
  - General Tag Filter
  - Hyphenator
  - Kuhlen Stemmer
  - Modifiable Term Filter
  - N Chars Filter
  - Number Filter
  - Oscar Filter
  - POS Filter
  - Pharma Filter
  - Porter Stemmer
  - Punctuation Erasure
  - RegEx Filter
  - Replacer
  - STTS Filter
  - Snowball Stemmer
  - Standard Named Entity Filter
  - Stop word Filter
  - Term Grouper

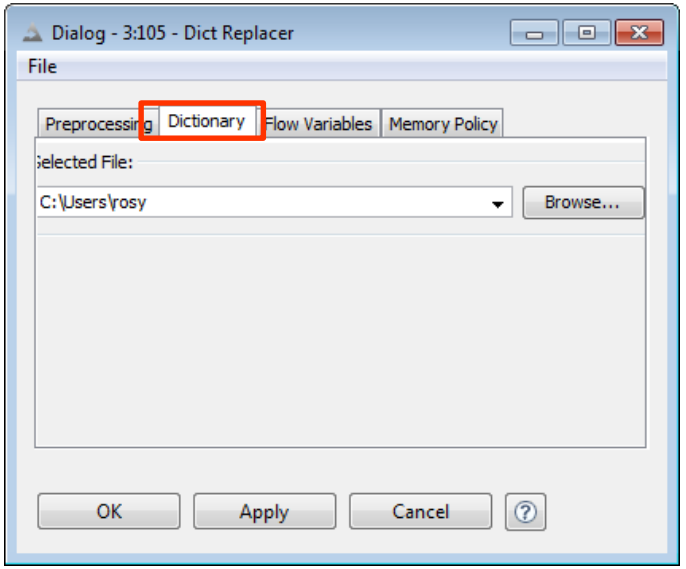
Term case conversion

Term replacement with dictionary word

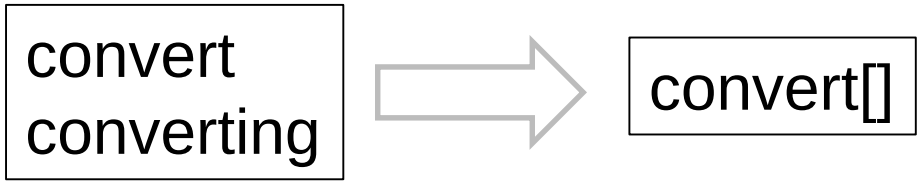
Introducing Hyphenation (Liang's algorithm)

RegEx based term replacer

Groups rows by term value



Stemming



- Kuhlen Stemmer (English only)
- Porter (English only)
- Snowball (English, German, French, ...)

Stemmed term replaces original term!

Workflow

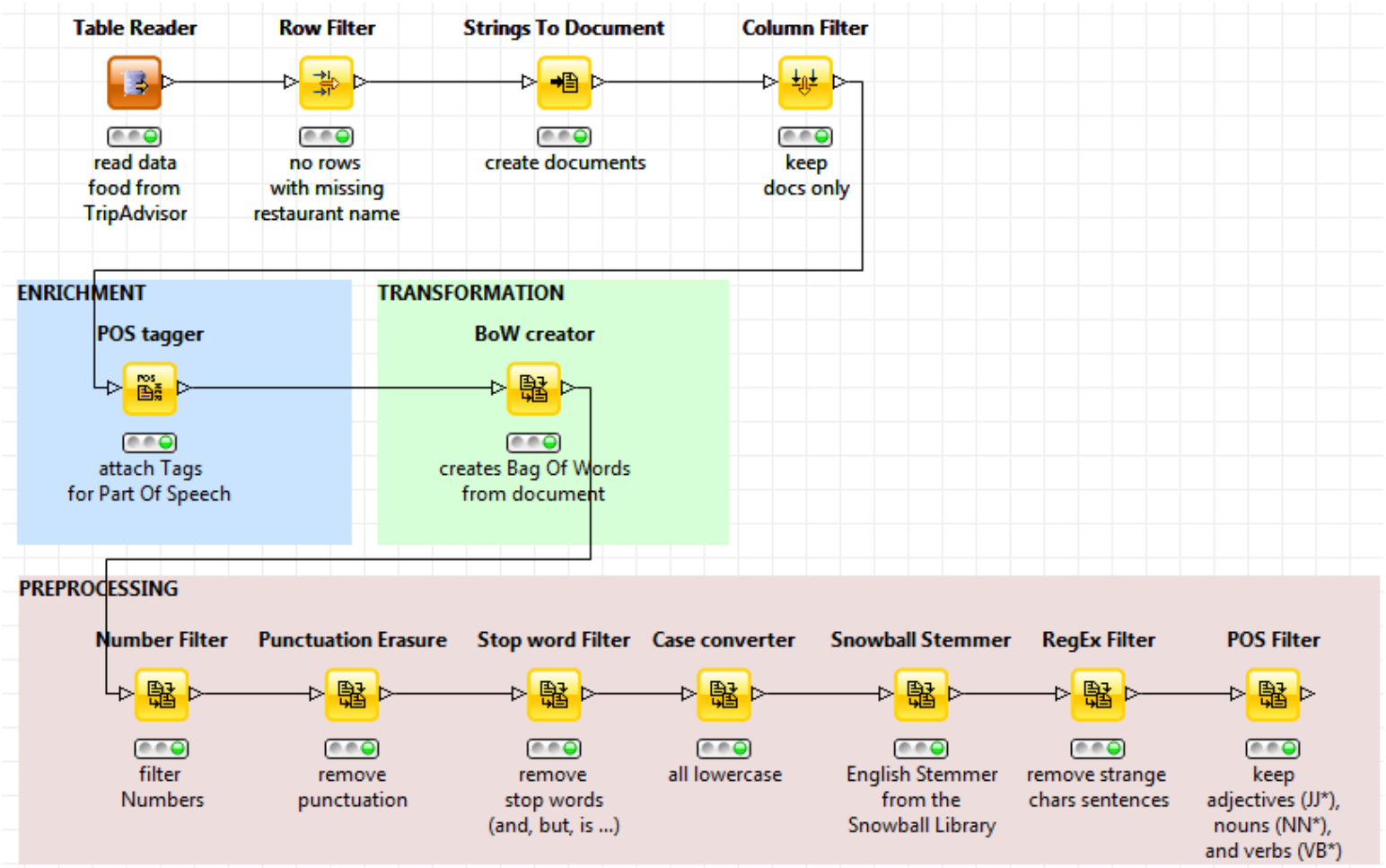


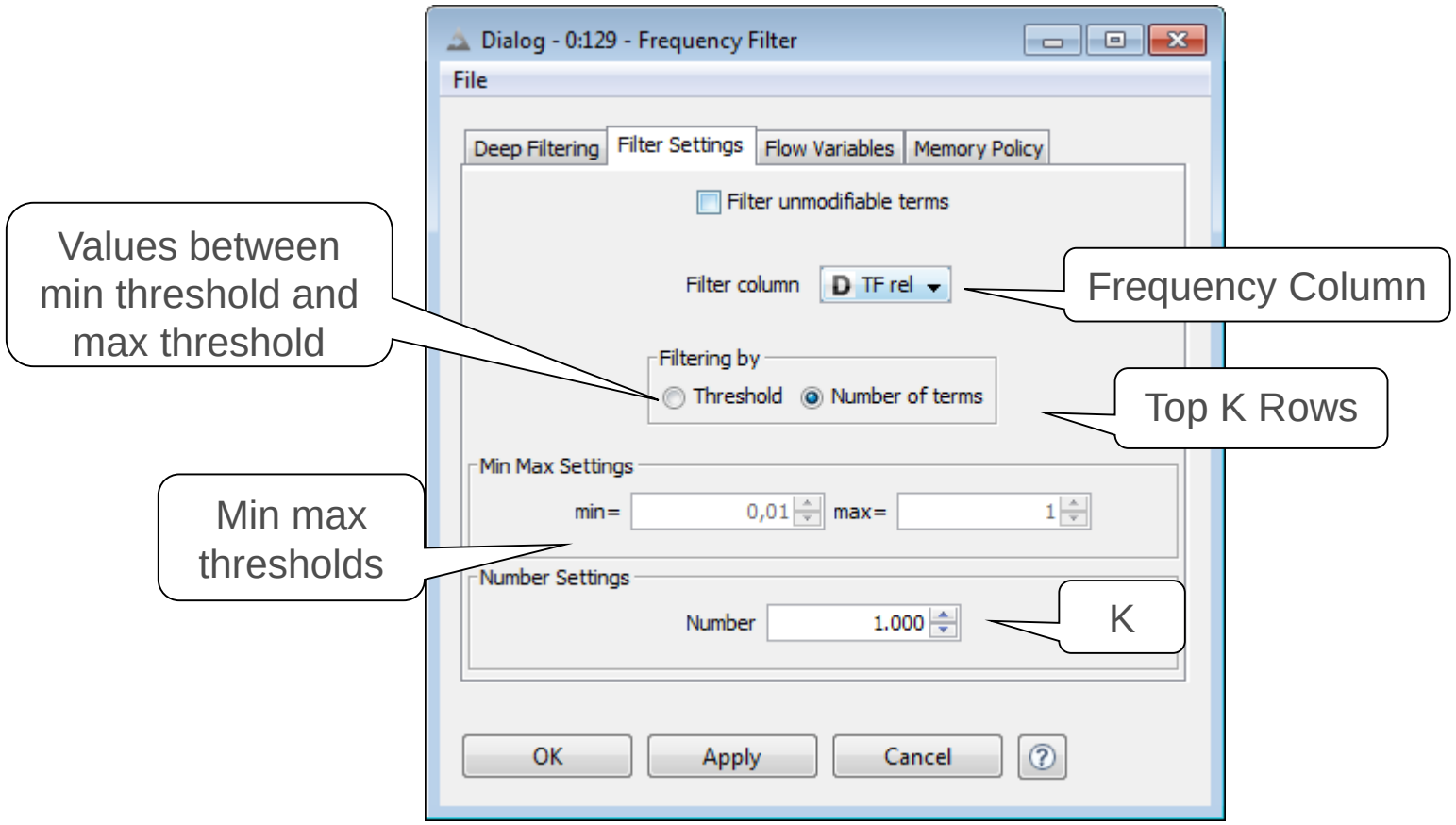
Table Reader
Row Filter
Strings To Document
Column Filter
read data food from TripAdvisor
no rows with missing restaurant name
create documents
keep docs only
ENRICHMENT
POS tagger
attach Tags for Part Of Speech
TRANSFORMATION
BoW creator
creates Bag Of Words from document
PREPROCESSING
Number Filter
Punctuation Erasure
Stop word Filter
Case converter
Snowball Stemmer
RegEx Filter
POS Filter
filter Numbers
remove punctuation
remove stop words (and, but, is ...)
all lowercase
English Stemmer from the Snowball Library
remove strange chars sentences
keep adjectives (JJ*), nouns (NN*), and verbs (VB*)

5 - Frequencies

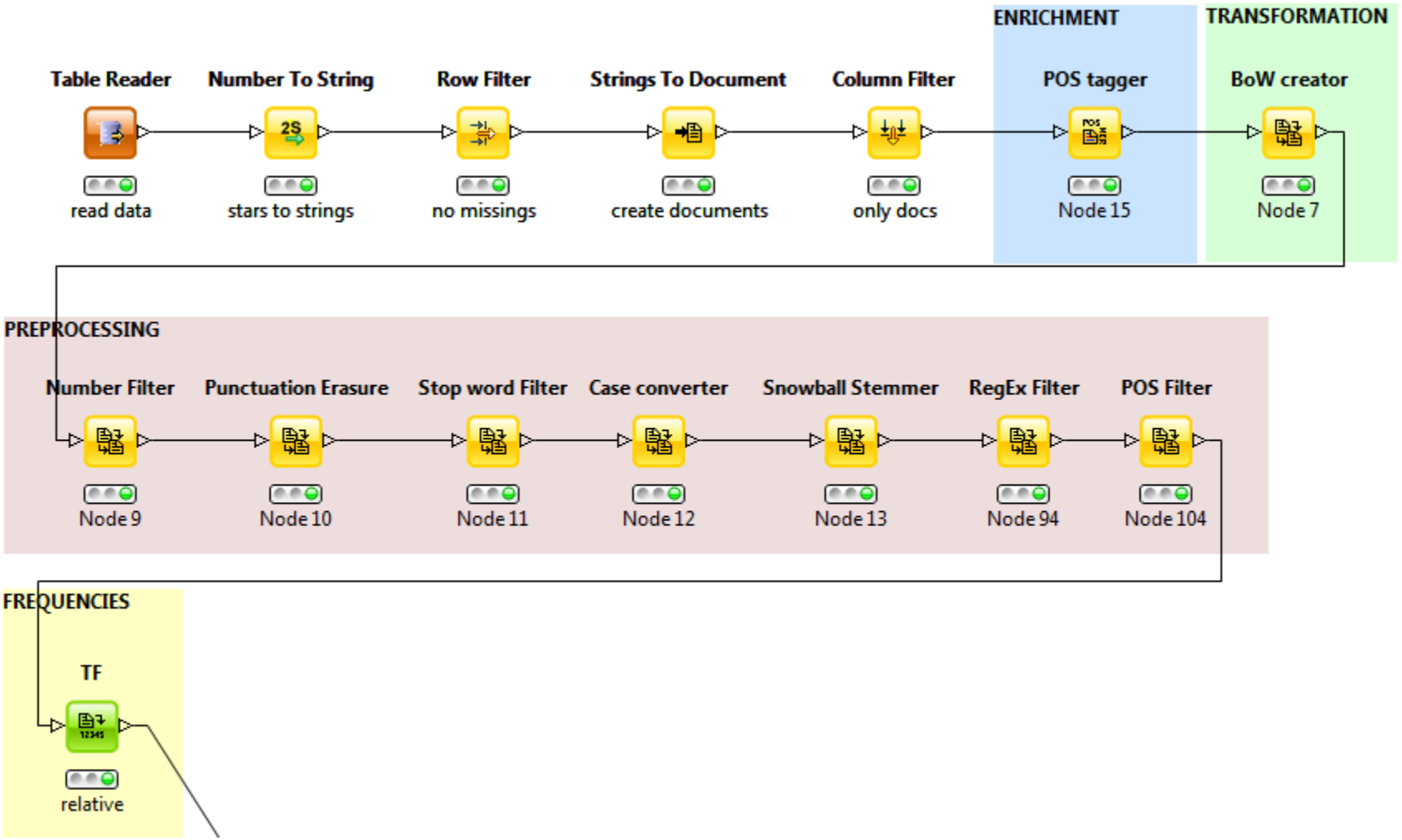
Frequency Measures

- TF (Term Frequency)
  - TF absolute = # occurr. of term t
  - TF relative = # occurr. of term t/ # terms
- IDF = log(1 + # docs / # docs with term t) (Inverse Document Frequency)
- ICF = log(1 + # cat./ # cat. with term t) (Inverse Category Frequency)
- IDF * TF

Frequency Filter

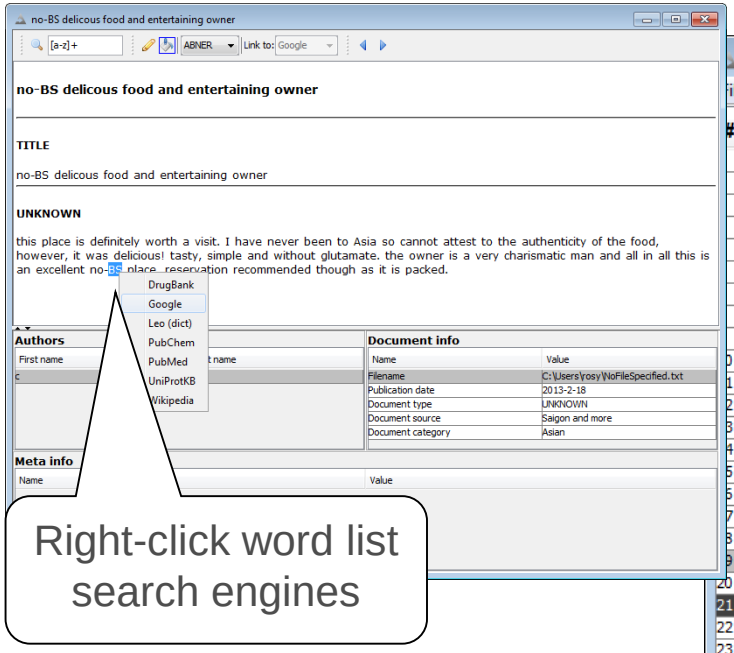


Workflow

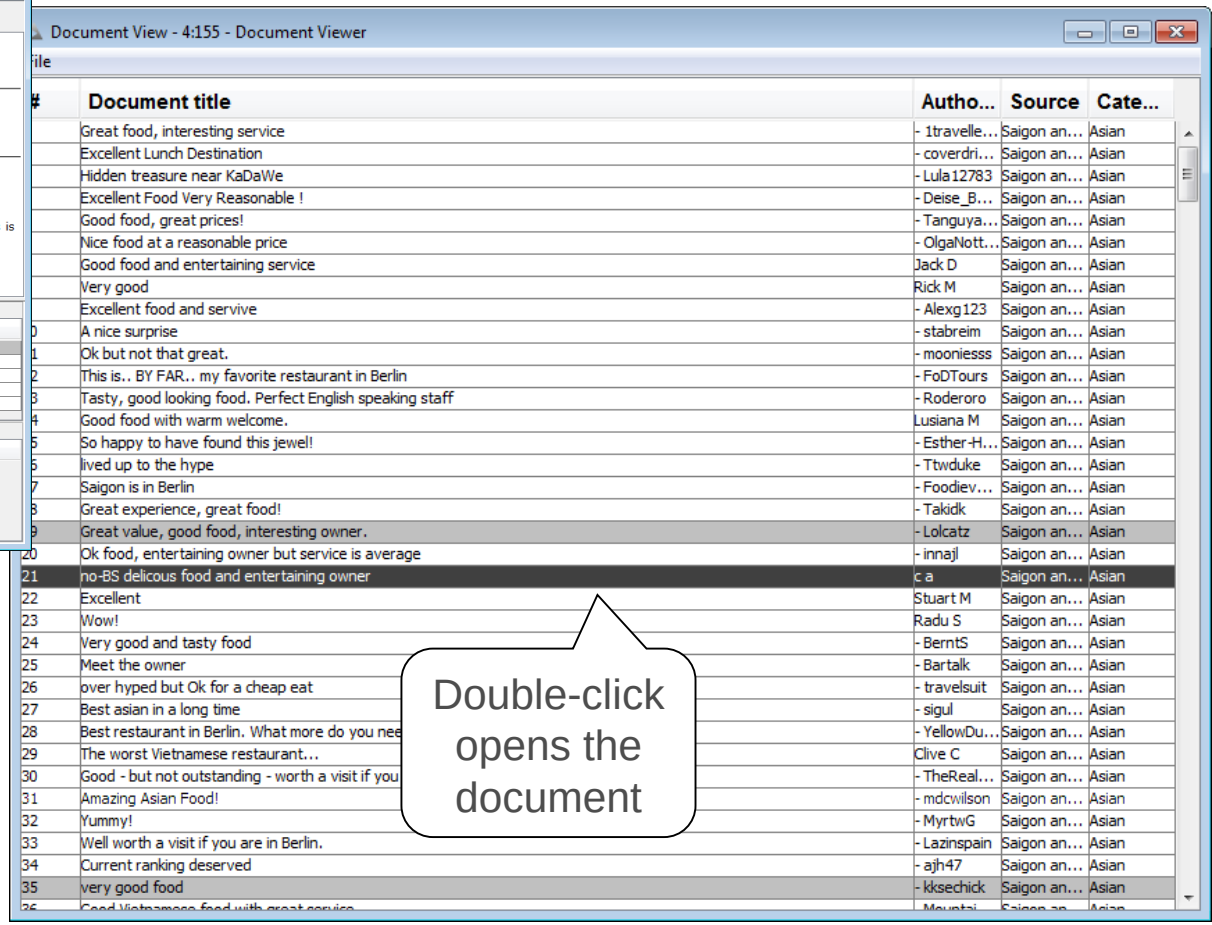

ENRICHMENT
TRANSFORMATION
Table Reader
Number To String
Row Filter
Strings To Document
Column Filter
POS tagger
BoW creator
read data
stars to strings
no missings
create documents
only docs
Node 15
Node 7
PREPROCESSING
Number Filter
Punctuation Erasure
Stop word Filter
Case converter
Snowball Stemmer
RegEx Filter
POS Filter
Node 9
Node 10
Node 11
Node 12
Node 13
Node 94
Node 104
FREQUENCIES
TF
relative

6 - Visualization

Document Viewer



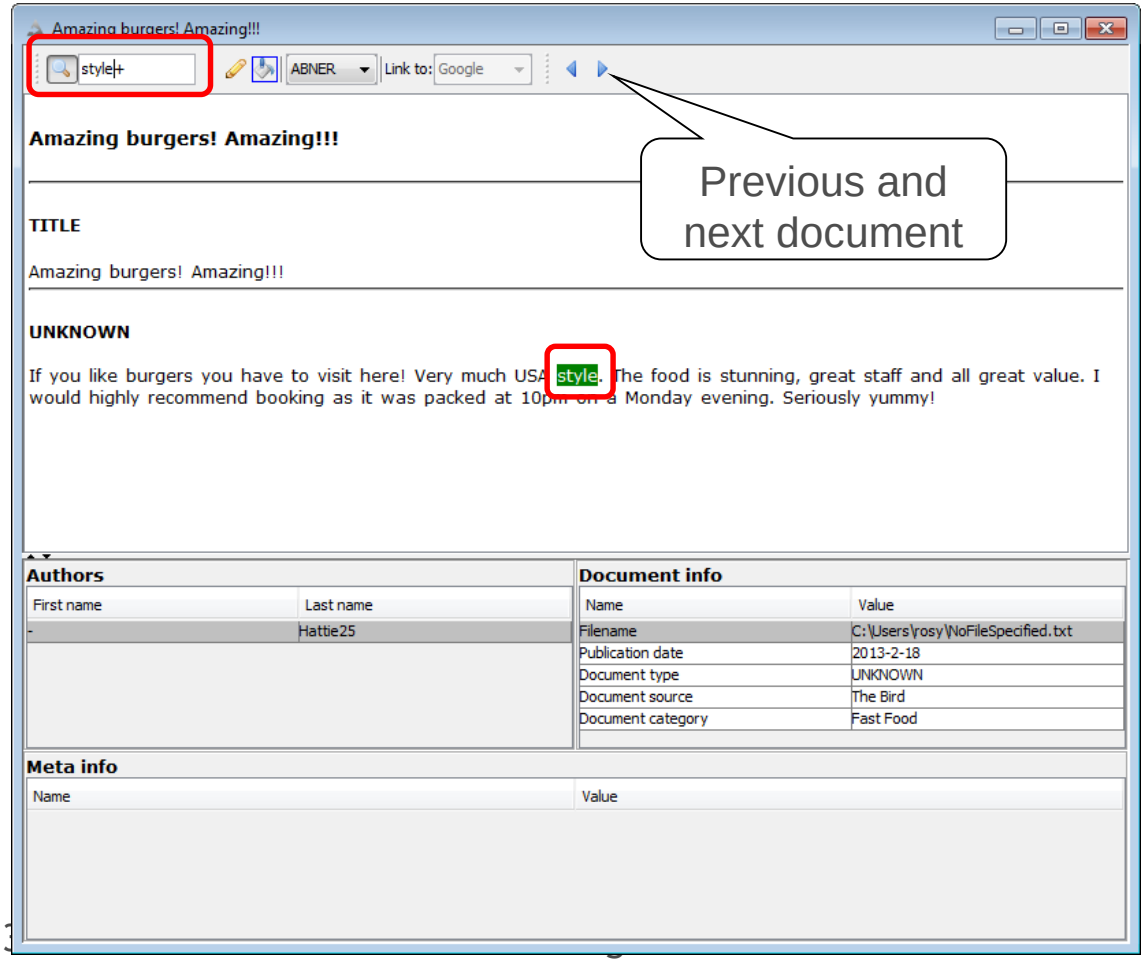
Right-click word list search engines



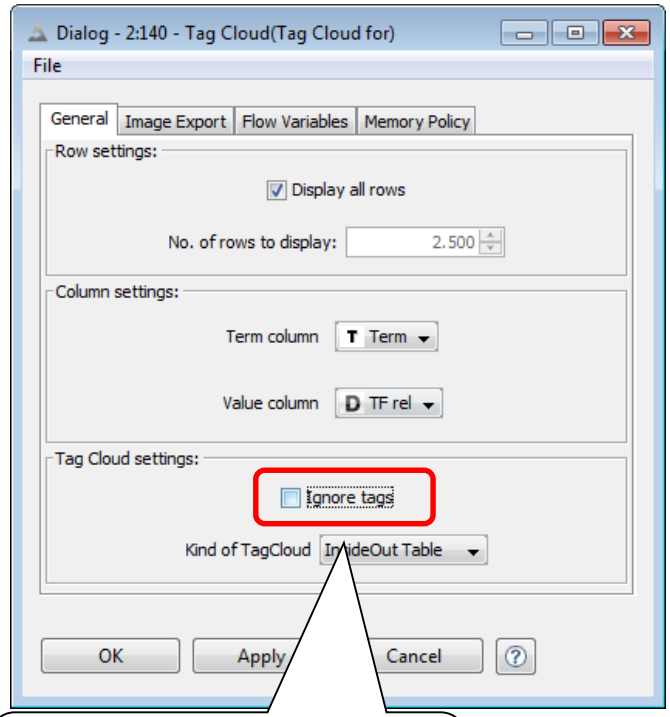
Double-click opens the document

Search engines listed in KNIME->Textprecessing->Search Engine Preferences

Document Viewer



Tag Cloud



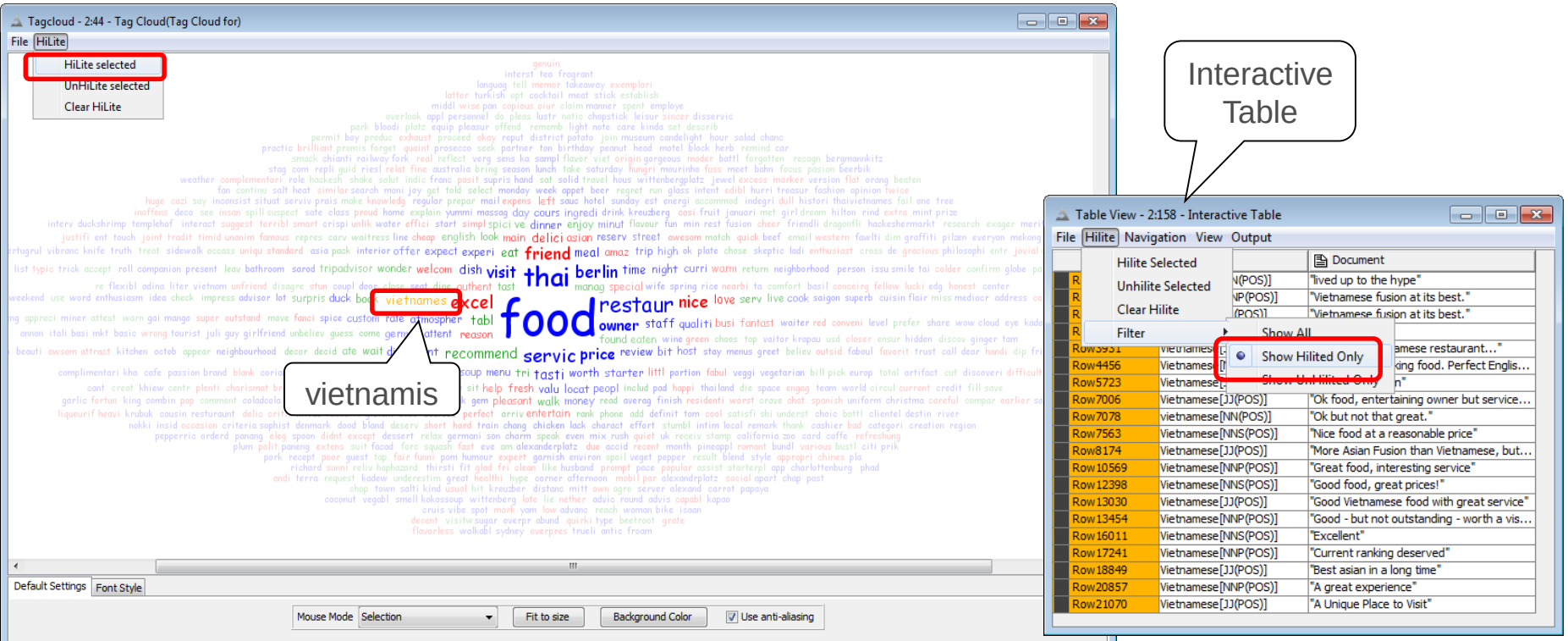
Adj, verbs and nouns as same word

Asian Restaurants

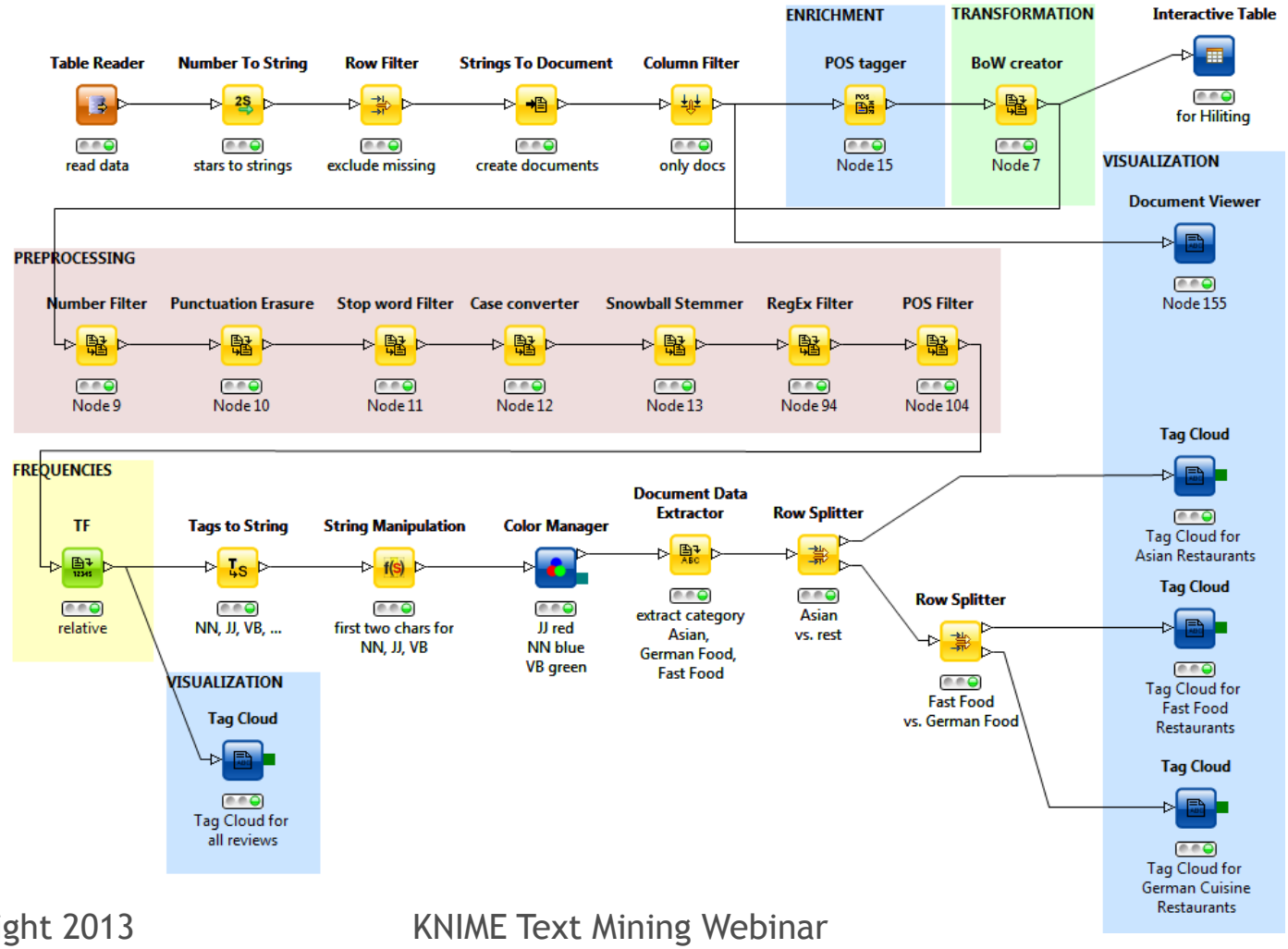
German Food Restaurants

Fast Food Restaurants

Hiliting in Tag Clouds

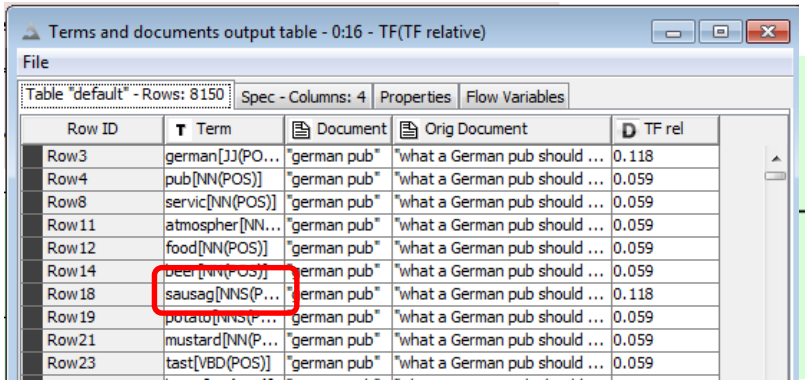


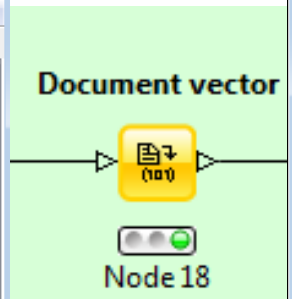
Workflow

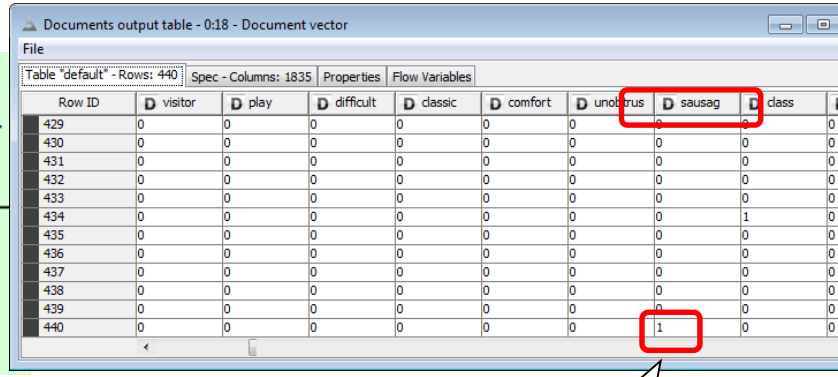

ENRICHMENT
TRANSFORMATION
Interactive Table
Table Reader
Number To String
Row Filter
Strings To Document
Column Filter
POS tagger
BoW creator
for Hiliting
read data
stars to strings
exclude missing
create documents
only docs
Node 15
Node 7
VISUALIZATION
Document Viewer
Node 155
PREPROCESSING
Number Filter
Punctuation Erasure
Stop word Filter
Case converter
Snowball Stemmer
RegEx Filter
POS Filter
Node 9
Node 10
Node 11
Node 12
Node 13
Node 94
Node 104
FREQUENCIES
TF
Tags to String
String Manipulation
Color Manager
Document Data Extractor
Row Splitter
Tag Cloud
relative
NN, JJ, VB, ...
first two chars for NN, JJ, VB
JJ red
NN blue
VB green
extract category Asian, German Food, Fast Food
Asian vs. rest
Tag Cloud for Asian Restaurants
Row Splitter
Tag Cloud
Fast Food vs. German Food
Tag Cloud for Fast Food Restaurants
VISUALIZATION
Tag Cloud
Tag Cloud
Tag Cloud for all reviews
Tag Cloud for German Cuisine Restaurants

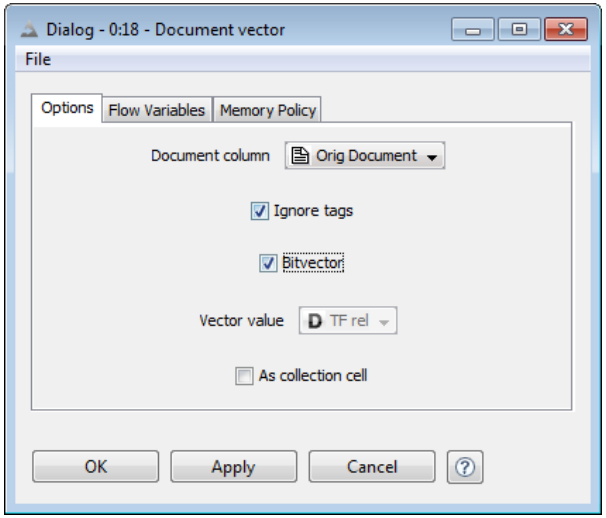
7 - Topic Classification

Document Vector





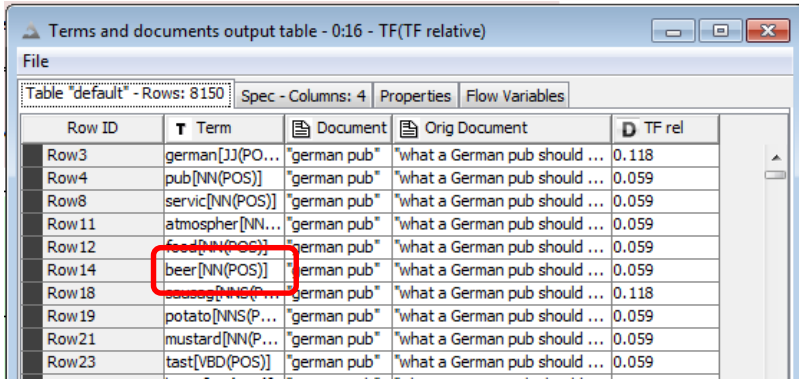


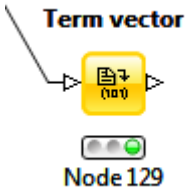


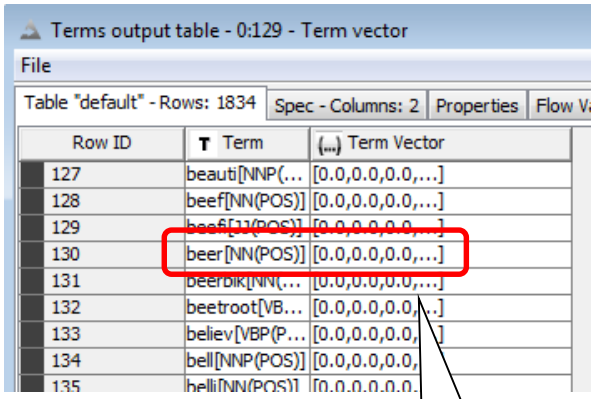
Bitvector or frequency measure

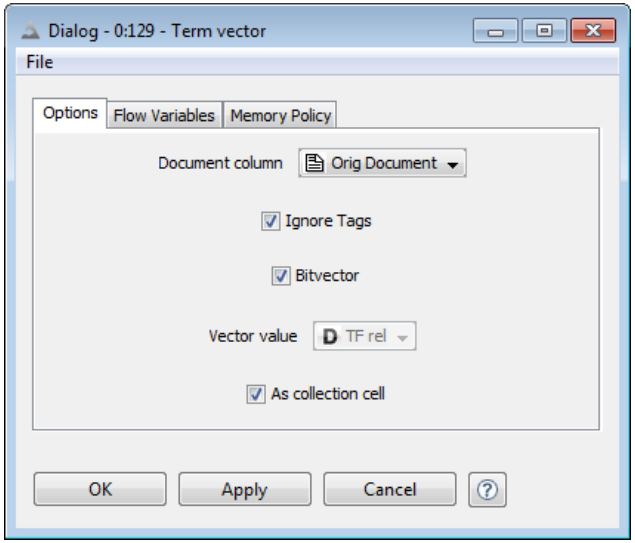
Document Vector: Documents represented in the terms space

Term Vector









Term Vector: Terms represented in the documents space

Bitvector or frequency measure

Topic Detection Goal

Possible Topics:

- Asian Restaurants
- German Food
- Fast Food

Pre-labeled data set available!

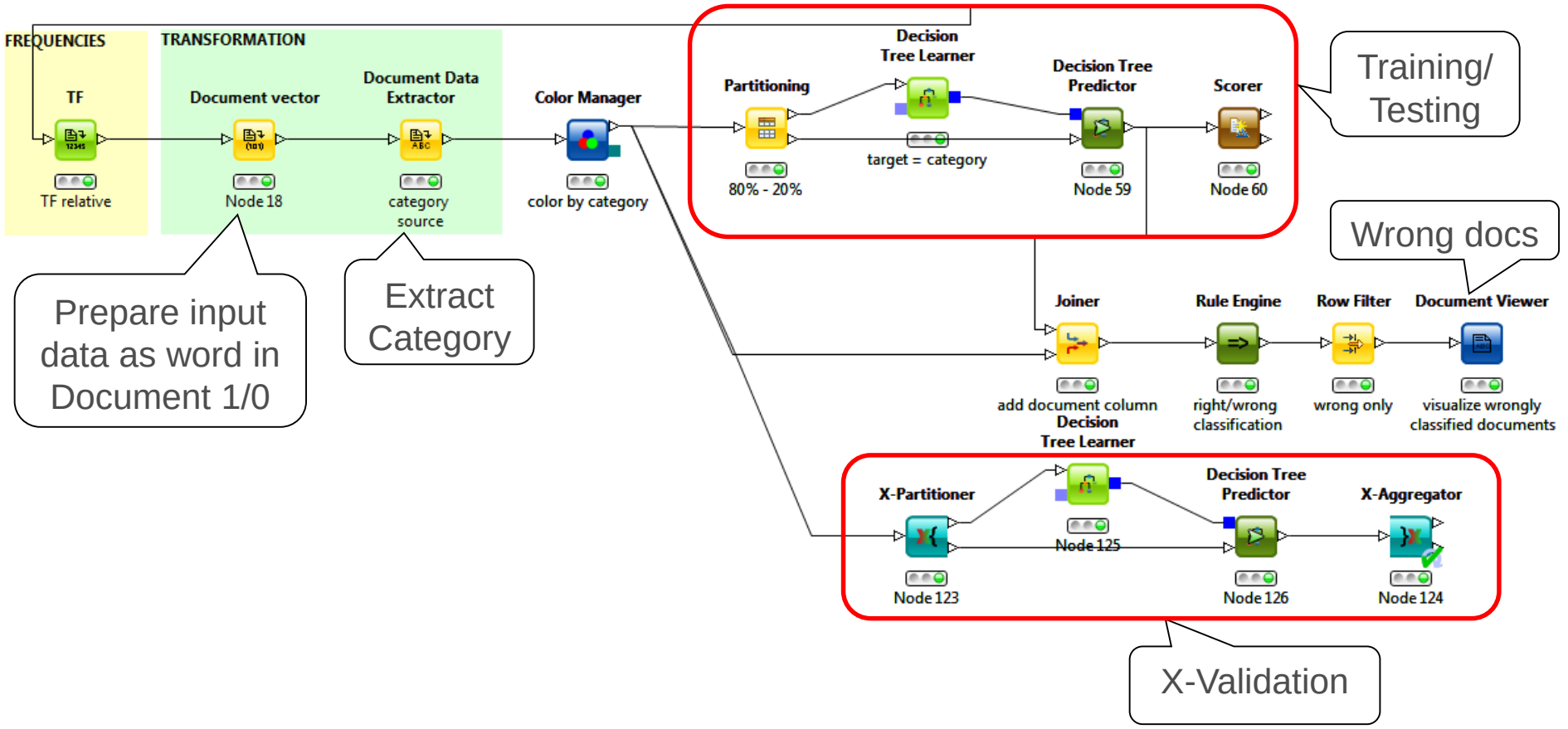
Target Topic is in Document Category.

Topic Detection

After the Document Vector Transformation, topic detection becomes just another data analytics problem.

80% for training set, 20% for test set.
Target = Category

Classification Sub-Workflow



Problem 1

Sometimes a pre-labeled data set is not available.

1. Use a Clustering technique
2. Find a similar pre-labeled data sets that you can adapt to the current problem

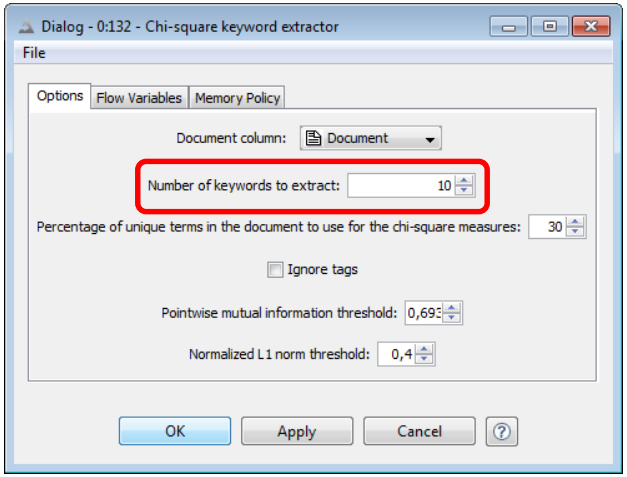
Problem 2

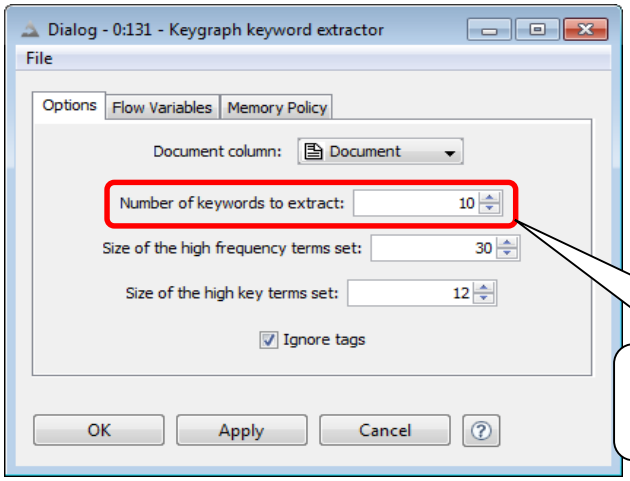
The vector generated by the Document Vector node can be high dimensional.

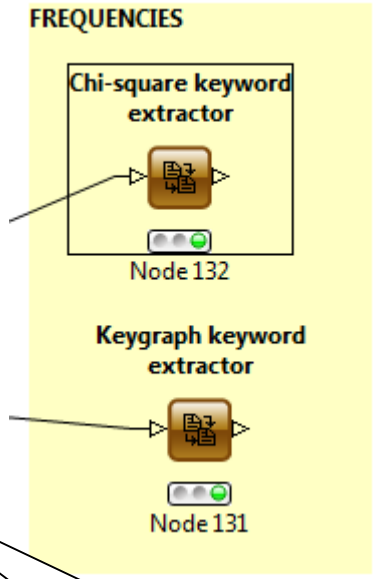
To reduce the input space dimensionality you can:

- Filter words by frequency
- Detect keywords and only use the most important ones.

Keywords Extractor Nodes







Max. # keywords per doc

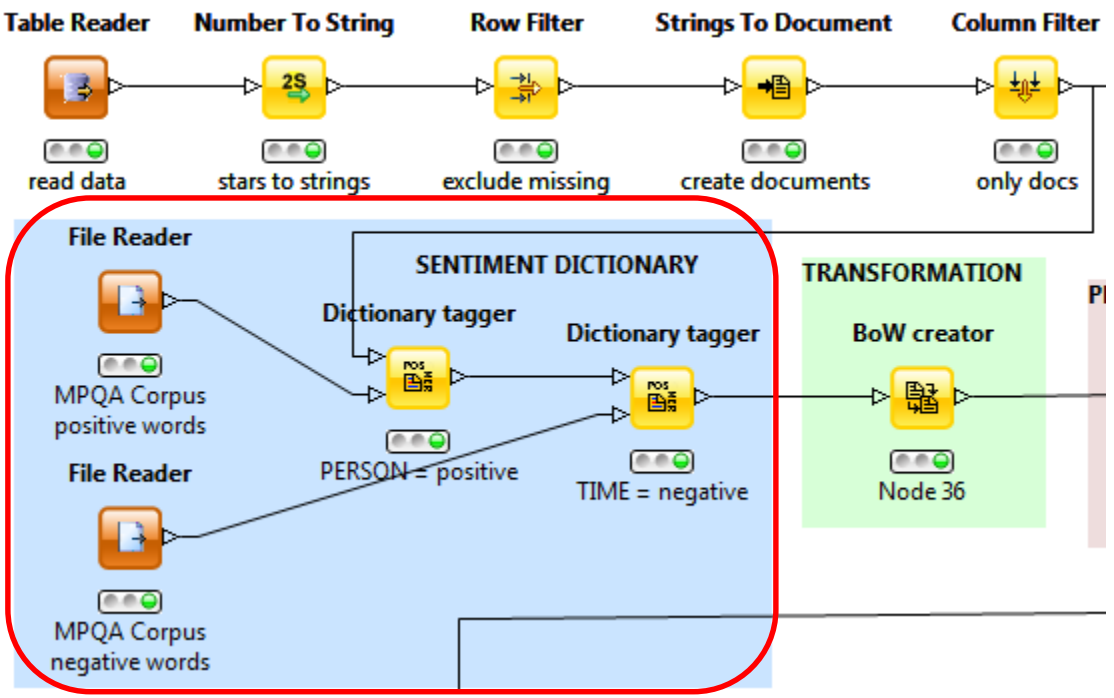
From:
"Keyword extraction from a single document using word co-occurrence statistical information" by Y.Matsuo and M. Ishizuka.

From:
"KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor" by Yukio Ohsawa.

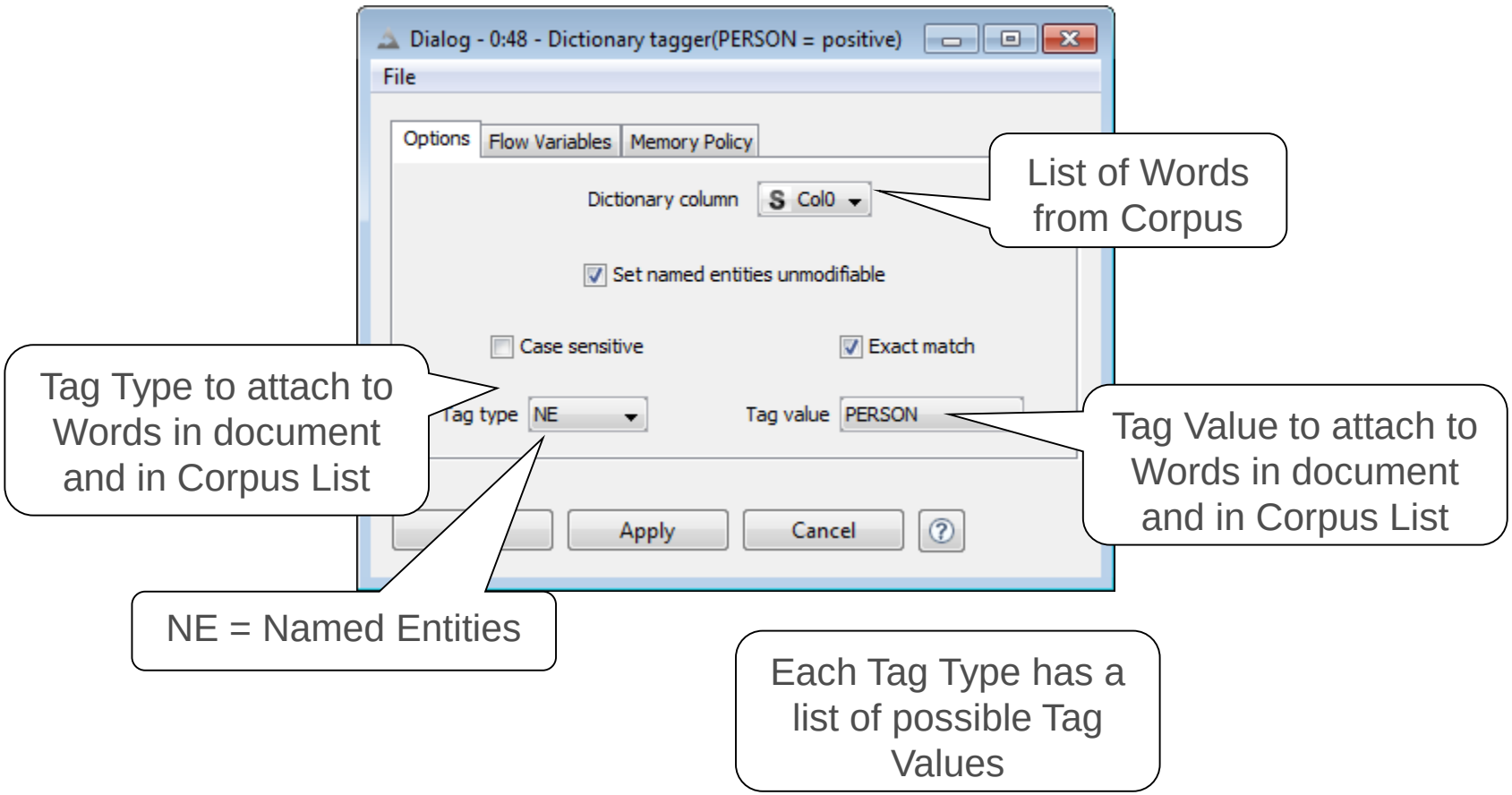
8 - Sentiment Analysis

Sentiment Corpus

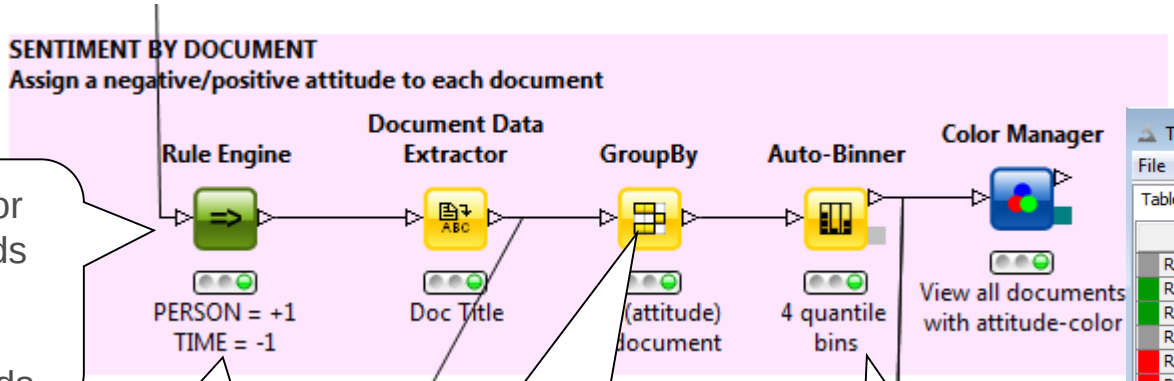
- MPQA Corpus with negative and positive words
- Tag Words according to Corpus with a Dictionary Tagger Node

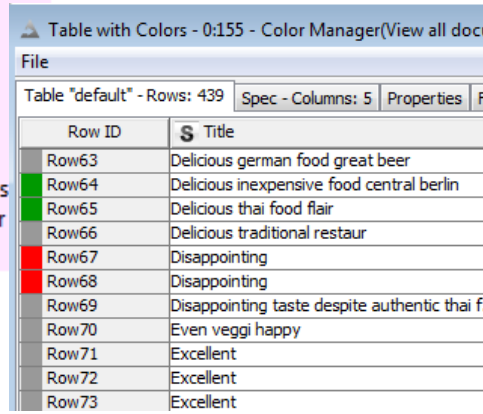


Dictionary Tagger Node



Sentiment By Document





PERSON Tag for POSITIVE Words

TIME Tag for NEGATIVE Words

ATTITUDE = +/-1 for POSITIVE or NEGATIVE Words

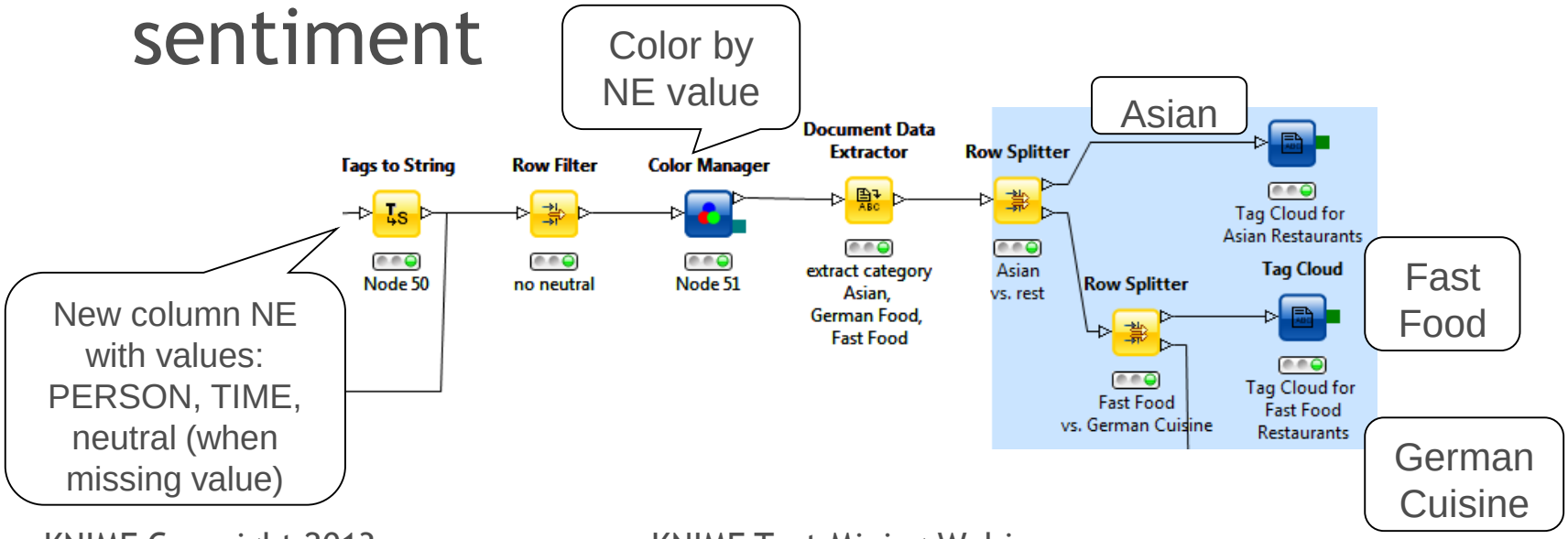
MEAN(ATTITUDE) on all words for each document

Bin1 = negative documents
Bin2, Bin3 = neutral documents
Bin4 = positive documents

Adjust Bin boundaries in Auto-Binner node for more accurate results

Sentiment by Category

- Tag Cloud on all words in all documents for a given category
- Words in tag clouds are colored by sentiment



Tag Cloud by Category

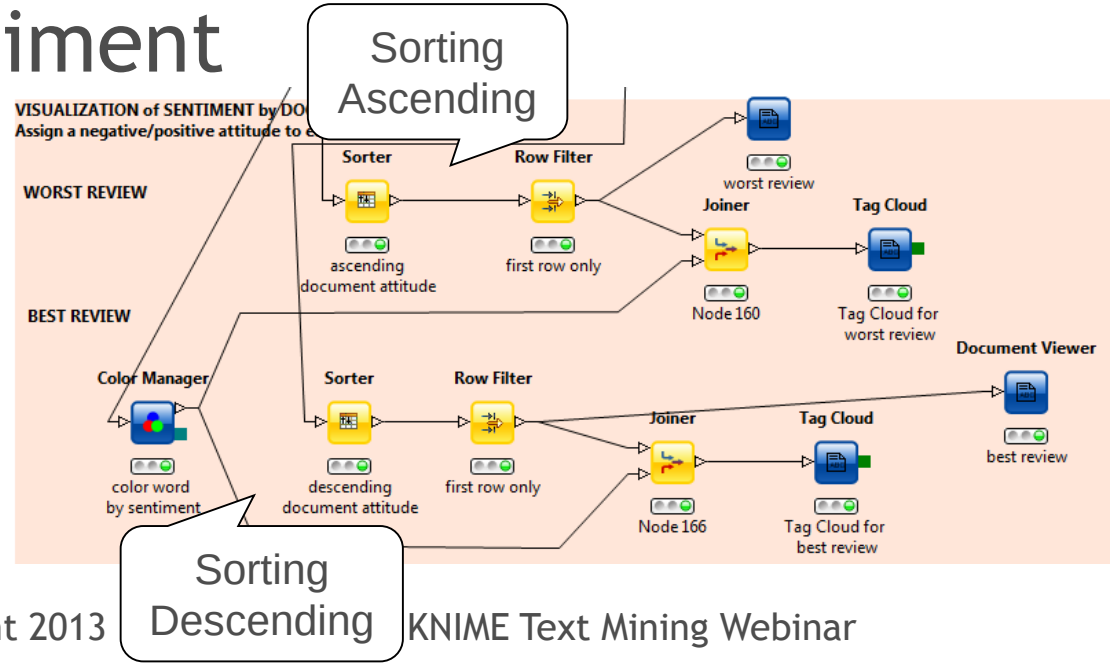
Asian

Fast Food

German

Sentiment by Word

- Find Most Positive/Most Negative Document
- Build Tag Cloud with words colored by sentiment



Tag Cloud of Worst/Best Doc

Most Positive

Most Negative

Improvements

- Add polarity change for negations
- Add polarity changes for enhancements
- Improve positive/negative dictionary
- Improve bin distribution (skeweness towards positive)
- Improve list of stop words
- Remove Names of Burgers?

Thank you

education@knime.com