



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

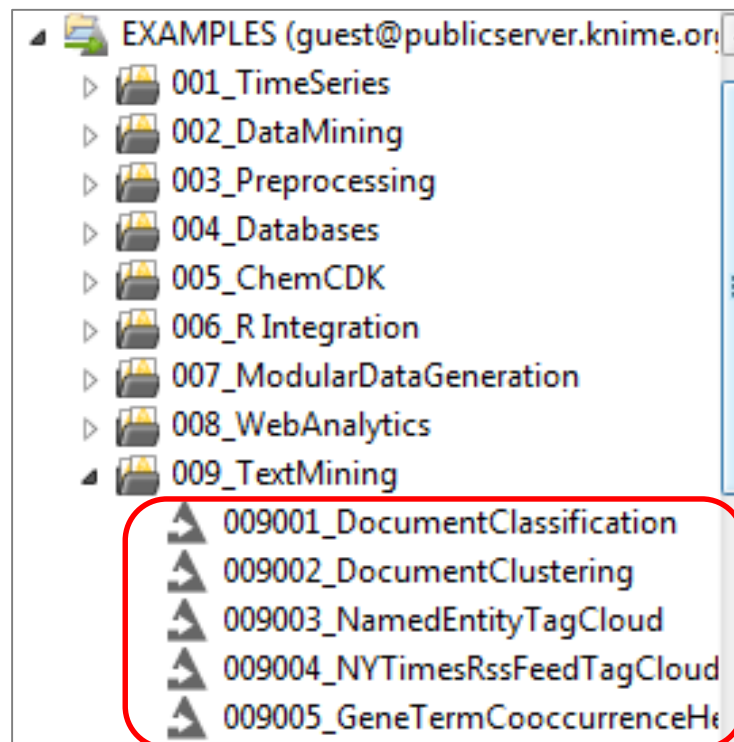
KNIME:

www.knime.org

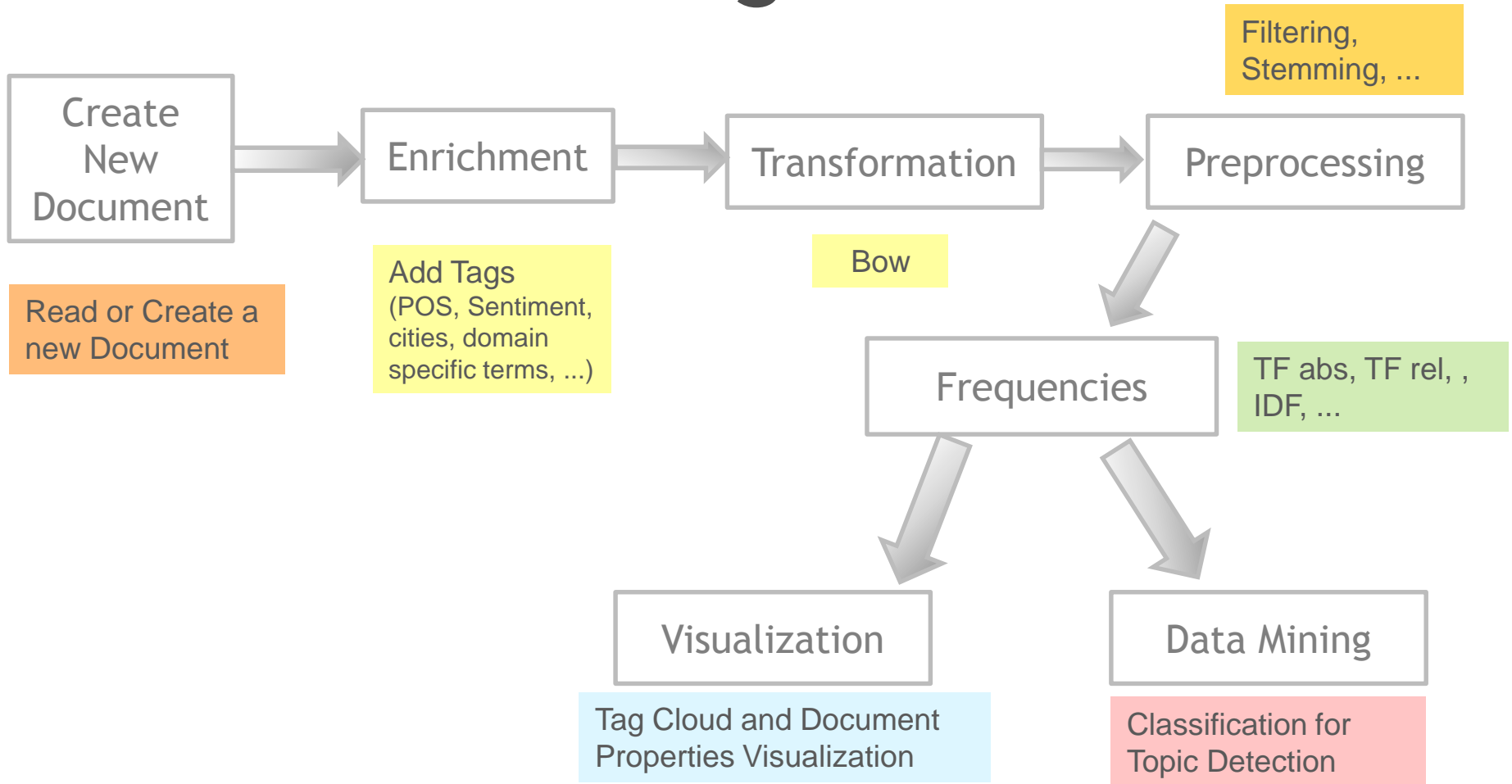
Install Textprocessing Extension
under KNIME Labs

Examples

Example Workflows available on the KNIME public server.



Text Mining Workflow

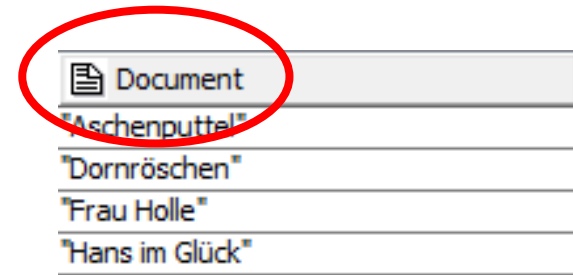


1 - Create a Document

New Data Types

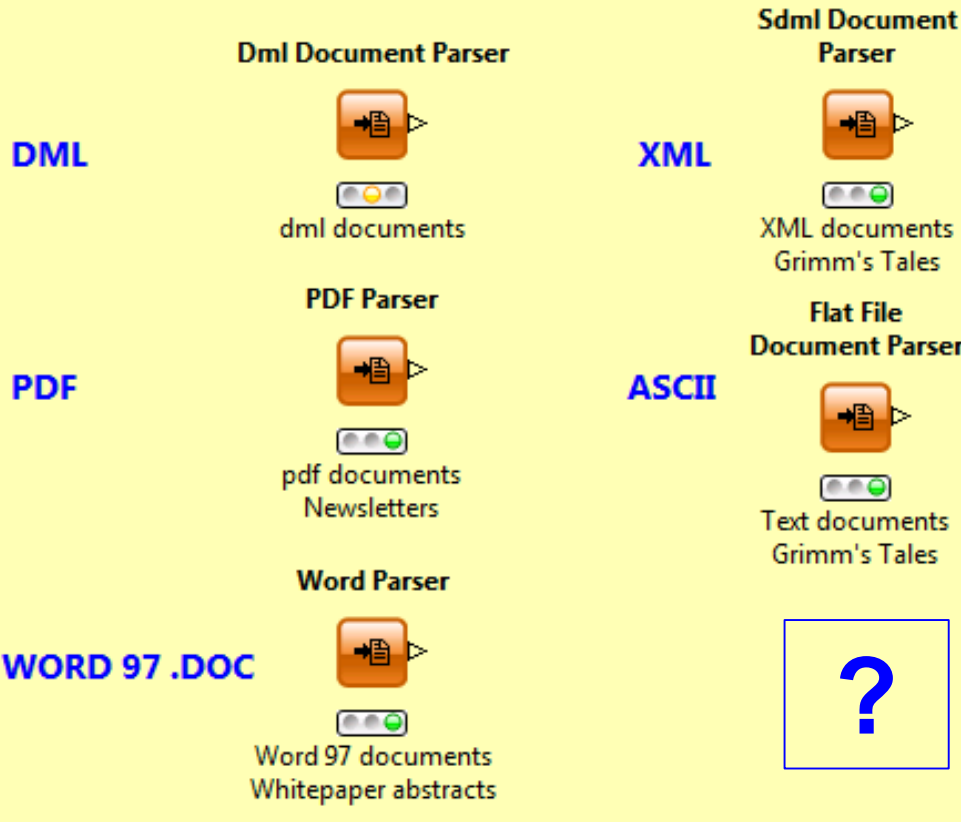
Document

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



From a Folder

Reading and parsing all files in a folder



The output is a list of Documents

Documents output table - 2:22 - Flat File Document Parser(Text do...

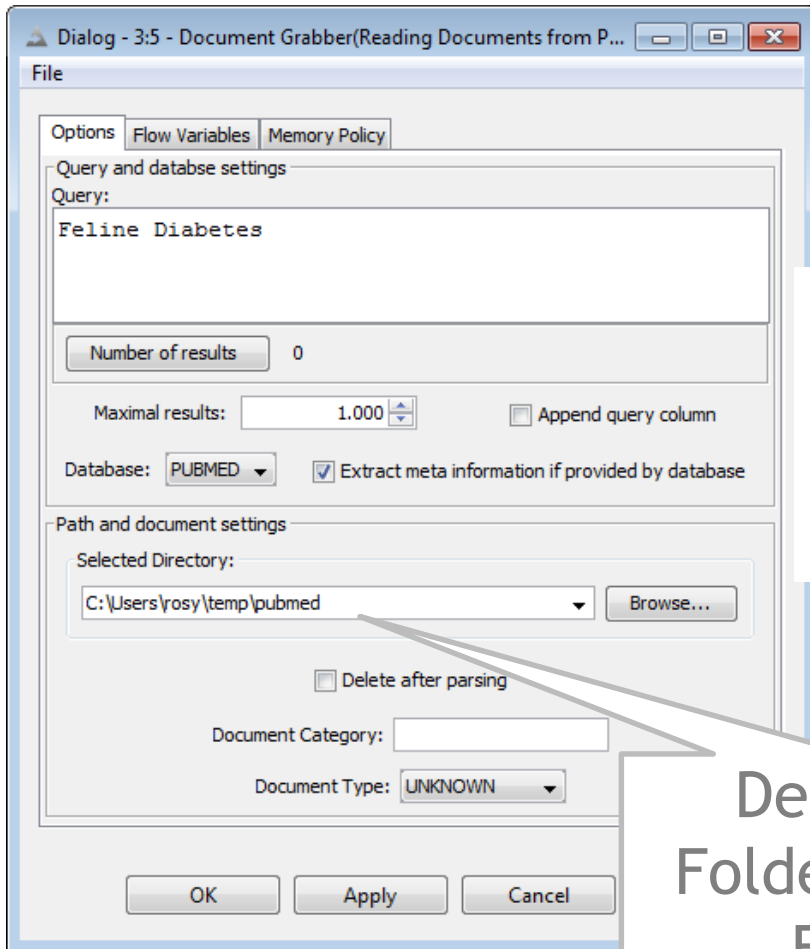
File

Table "default" - Rows: 211 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Document
Row 1	"A shoemaker, by no fault of his own, had become so poor that at last he f
Row 2	"Allerleirauh."
Row 3	"A Riddling Tale."
Row 4	"Bearskin."
Row 5	"Brides On Their Trial."
Row 6	"Brother and Sister."
Row 7	"Brother Lustig."
Row 8	"Cat and Mouse in Partnership."
Row 9	"Cinderella."
Row 10	"Clever Elsie."

From PUBMED

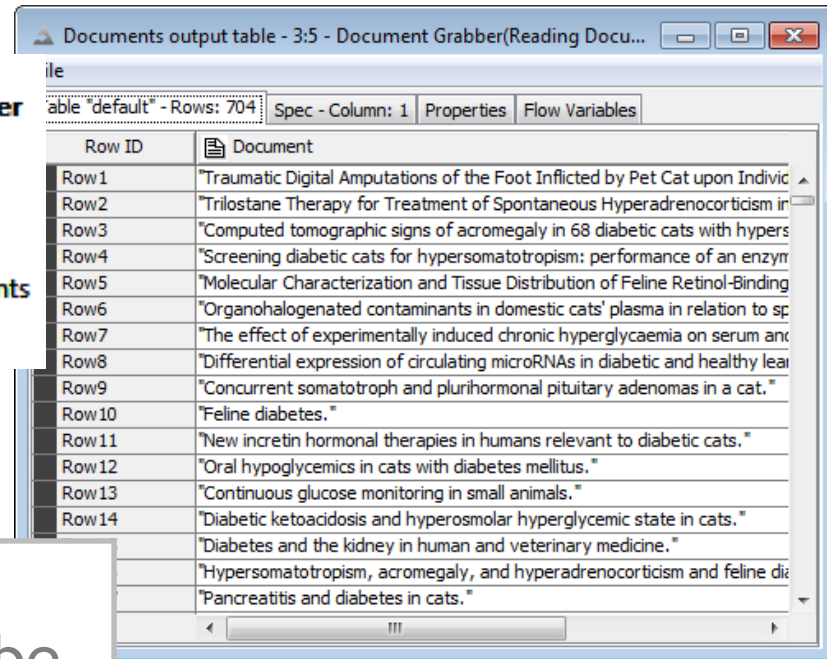
The output is a list of Documents



Document Grabber



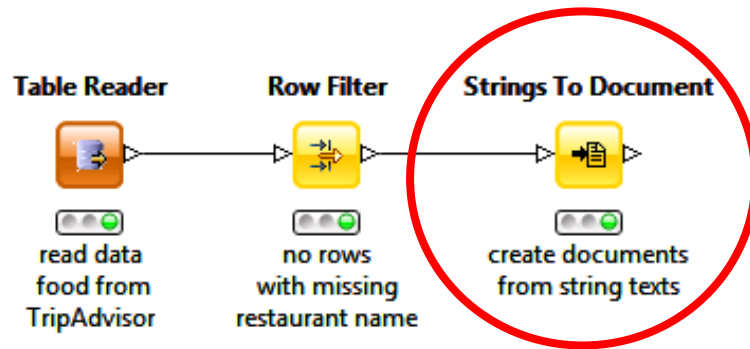
Retrieving Documents from PUBMED



Row ID	Document
Row 1	"Traumatic Digital Amputations of the Foot Inflicted by Pet Cat upon Individ...
Row 2	"Trilostane Therapy for Treatment of Spontaneous Hyperadrenocorticism in...
Row 3	"Computed tomographic signs of acromegaly in 68 diabetic cats with hypers...
Row 4	"Screening diabetic cats for hypersomatotropism: performance of an enzym...
Row 5	"Molecular Characterization and Tissue Distribution of Feline Retinol-Binding...
Row 6	"Organohalogenated contaminants in domestic cats' plasma in relation to sp...
Row 7	"The effect of experimentally induced chronic hyperglycaemia on serum and...
Row 8	"Differential expression of circulating microRNAs in diabetic and healthy lea...
Row 9	"Concurrent somatotroph and plurihormonal pituitary adenomas in a cat."
Row 10	"Feline diabetes."
Row 11	"New incretin hormonal therapies in humans relevant to diabetic cats."
Row 12	"Oral hypoglycemics in cats with diabetes mellitus."
Row 13	"Continuous glucose monitoring in small animals."
Row 14	"Diabetic ketoacidosis and hyperosmolar hyperglycemic state in cats." "Diabetes and the kidney in human and veterinary medicine." "Hypersomatotropism, acromegaly, and hyperadrenocorticism and feline di...
	"Pancreatitis and diabetes in cats."

Destination Folder MUST be EMPTY!

Strings to Document



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

Row ID	Restaurant	Category	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!"

Dialog - 2:3 - Strings To Document(create documents)

File

Options: Flow Variables, Memory Policy

Text

Title: \$ review title, Full text: \$ review text

Authors

Authors: \$ user name, Author names separator: ,

Source and Category

Document source: [text box]

Use sources from column, Document source column: \$ Restaurant

Document category: [text box]

Use categories from column, Document category column: \$ Category

Type and Date

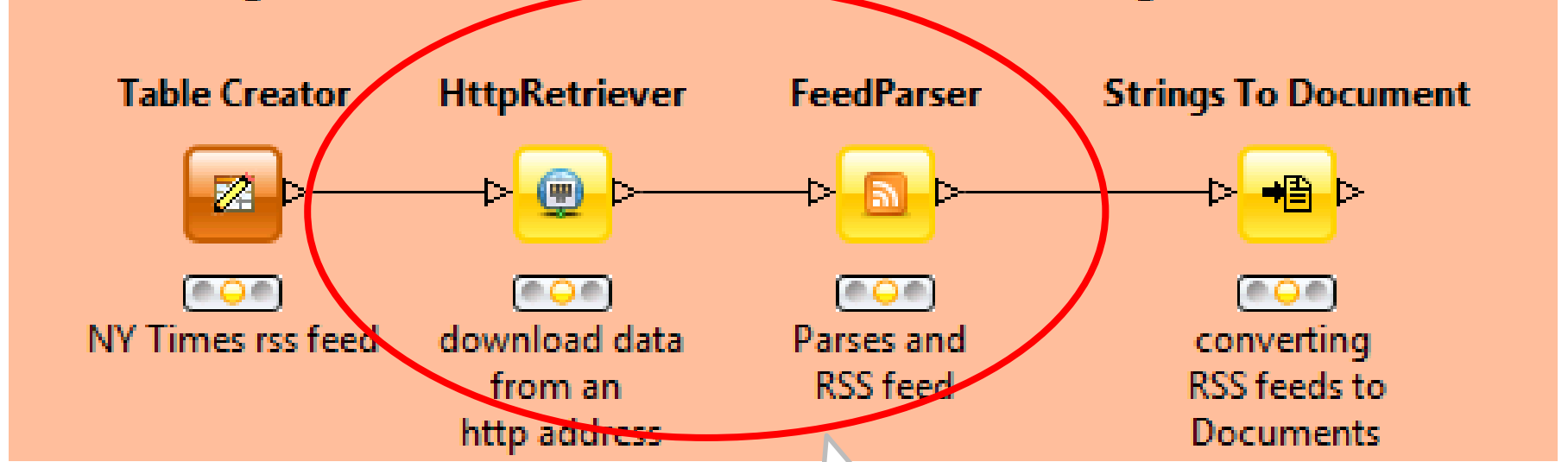
Document type: UNKNOWN

Publication date (dd-mm-yyyy): 18-02-2013

OK, Apply, Cancel, ?

From RSS Feeds

Downloading of latest NY Times RSS feed and transforming into document cells



<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	S user name	S review title	S review text	I stars	I Reviews	S Restaurant	S 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 da...	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.

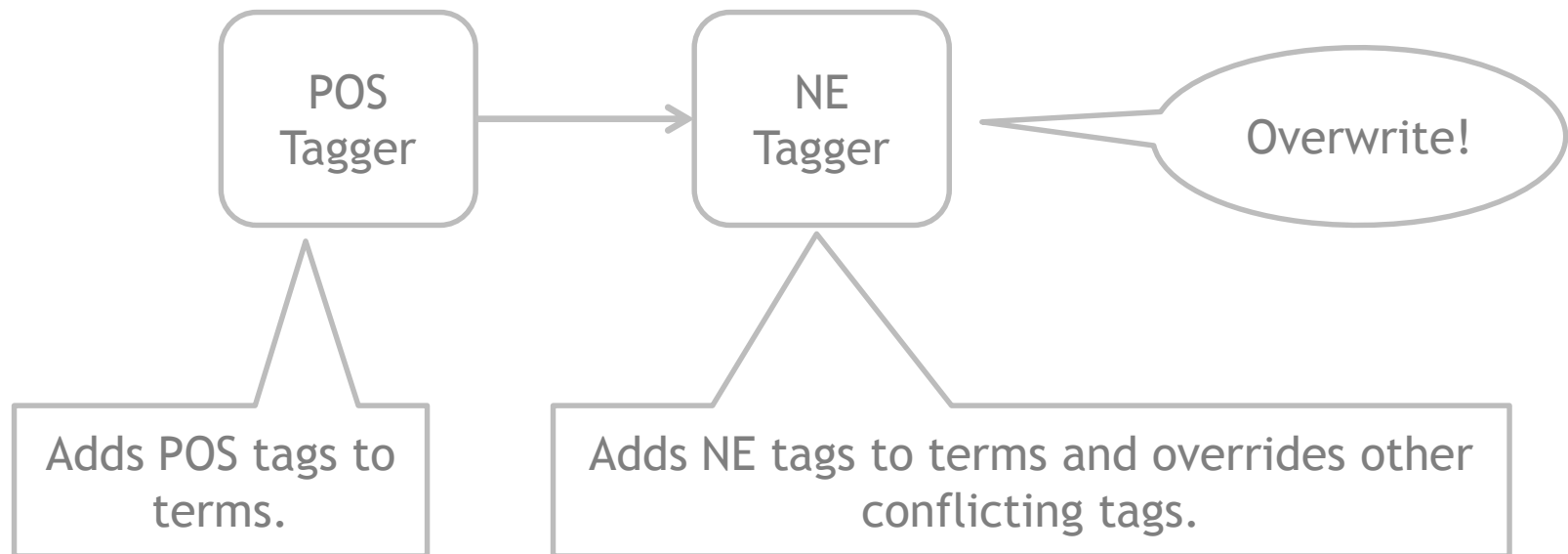
- Multiword detection, named entity recognition, part of speech definition, ...
- To each detected entity (term) a tag is added, specifying its type.
- To avoid intersection of granularity the **last node dominates**.

Tagger Conflict Resolution

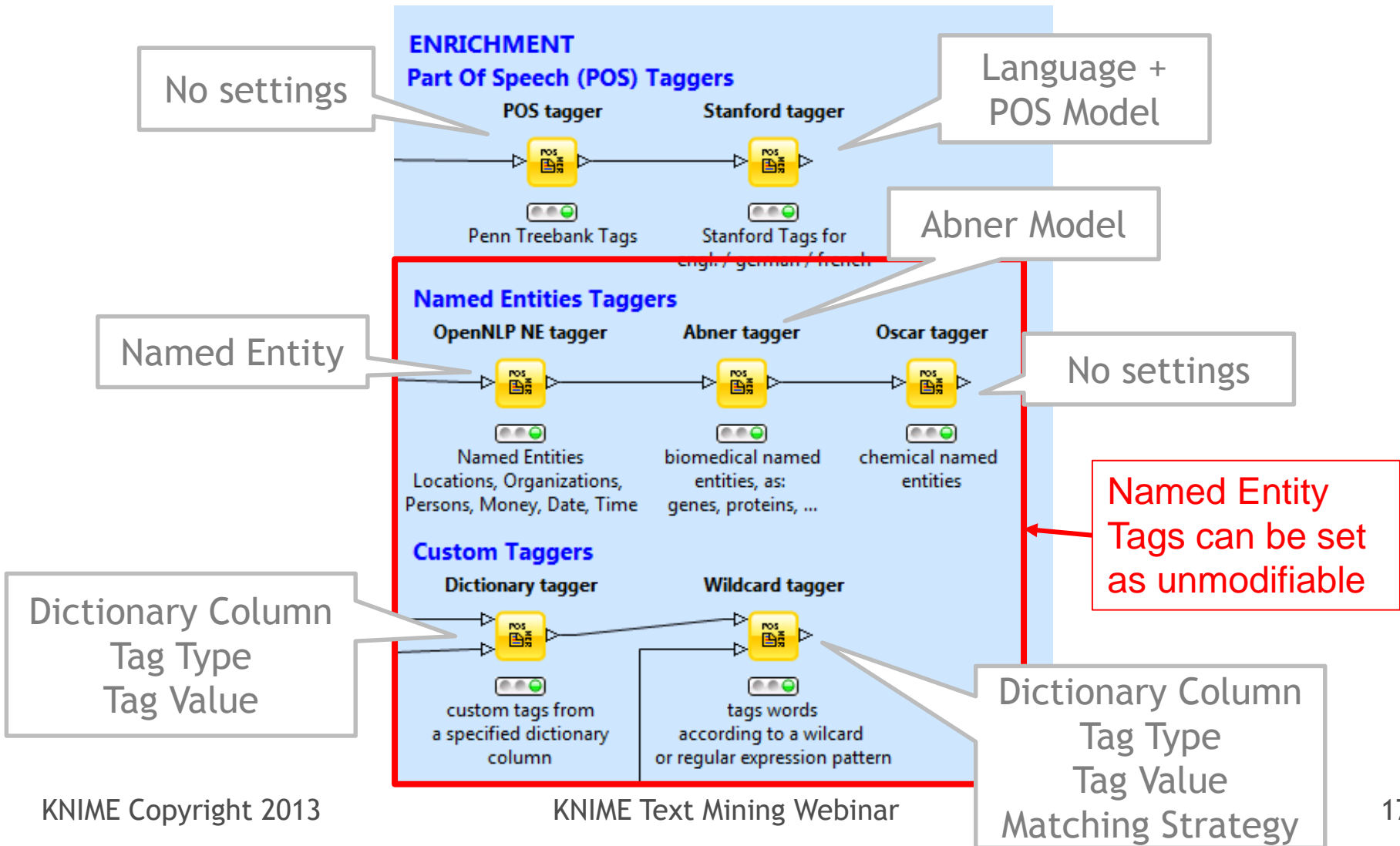
In case of intersections of granularity the last node overwrites.

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 ”



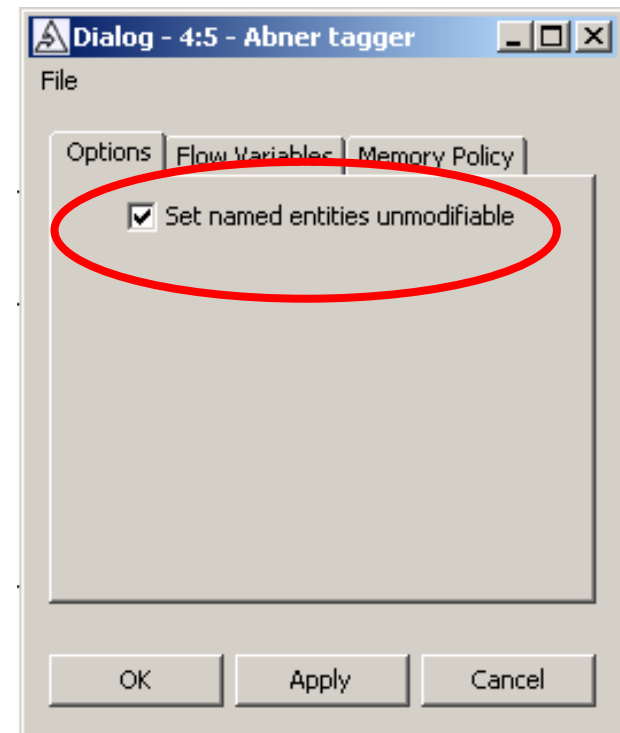
Tagging



Unmodifiable Named Entity Tags

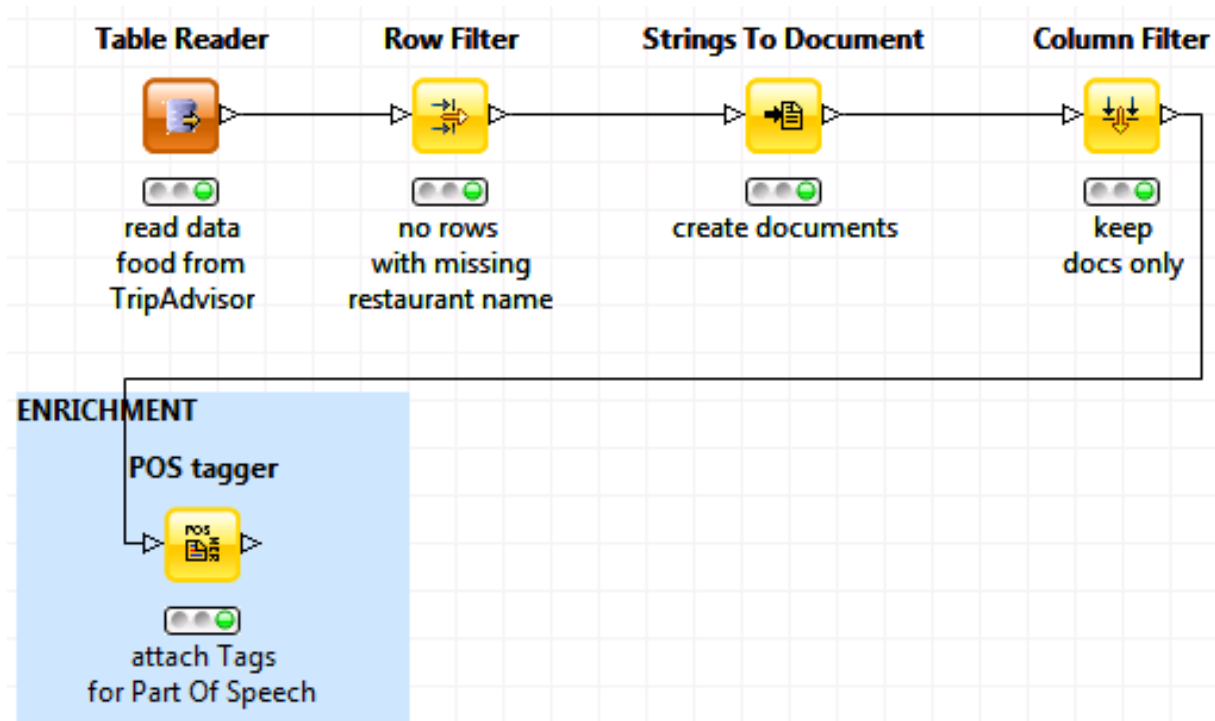
Named Entities Tags attached through enrichment nodes can be set as **unmodifiable**

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



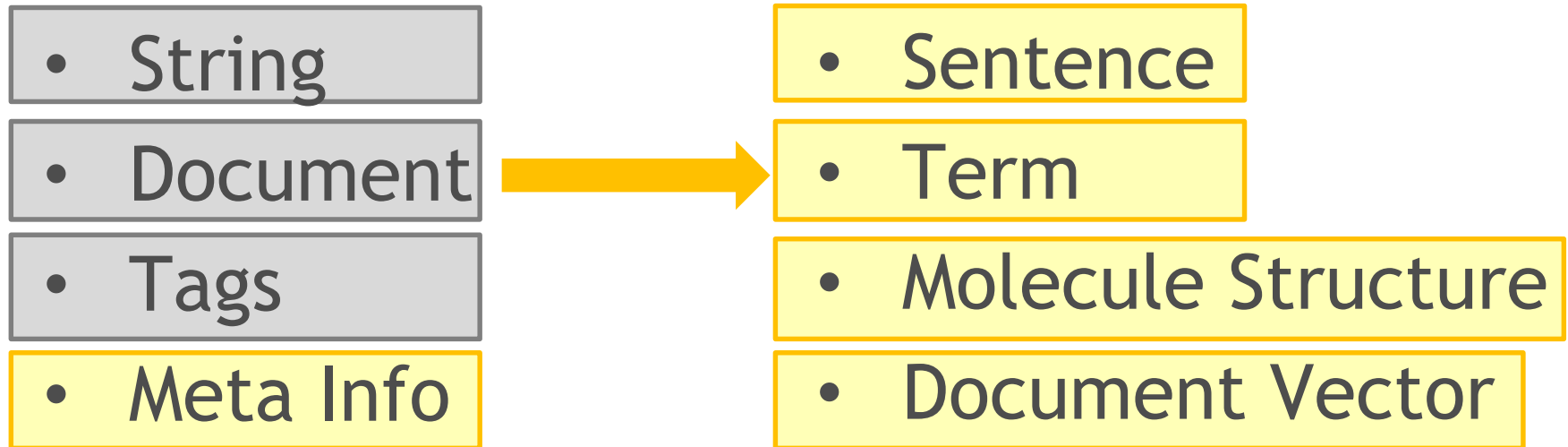
Workflow

Strings to Document Workflow + POS Tagger



3 - Transformation

Data Types and Features



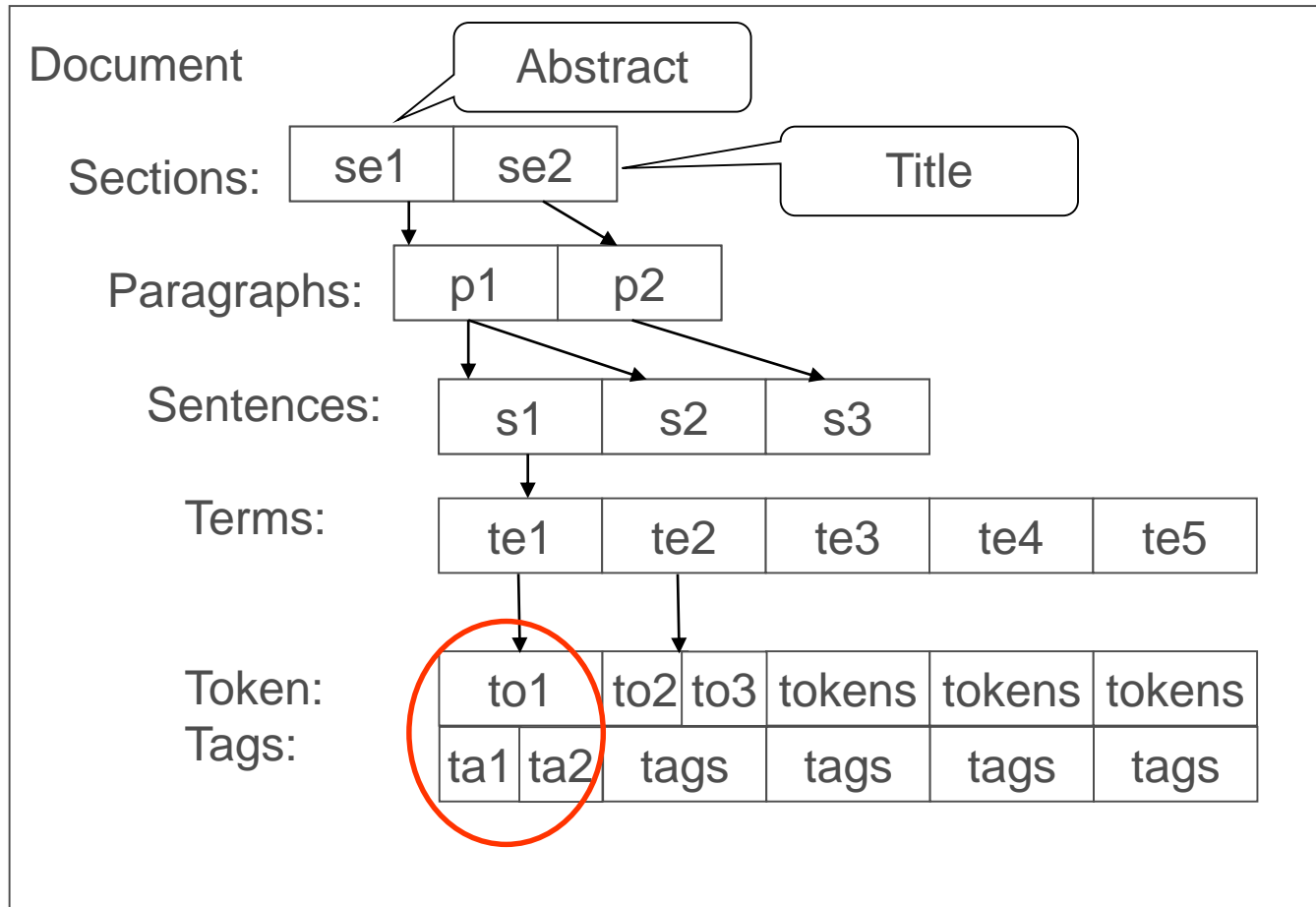
Parsing / Tokenization

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.

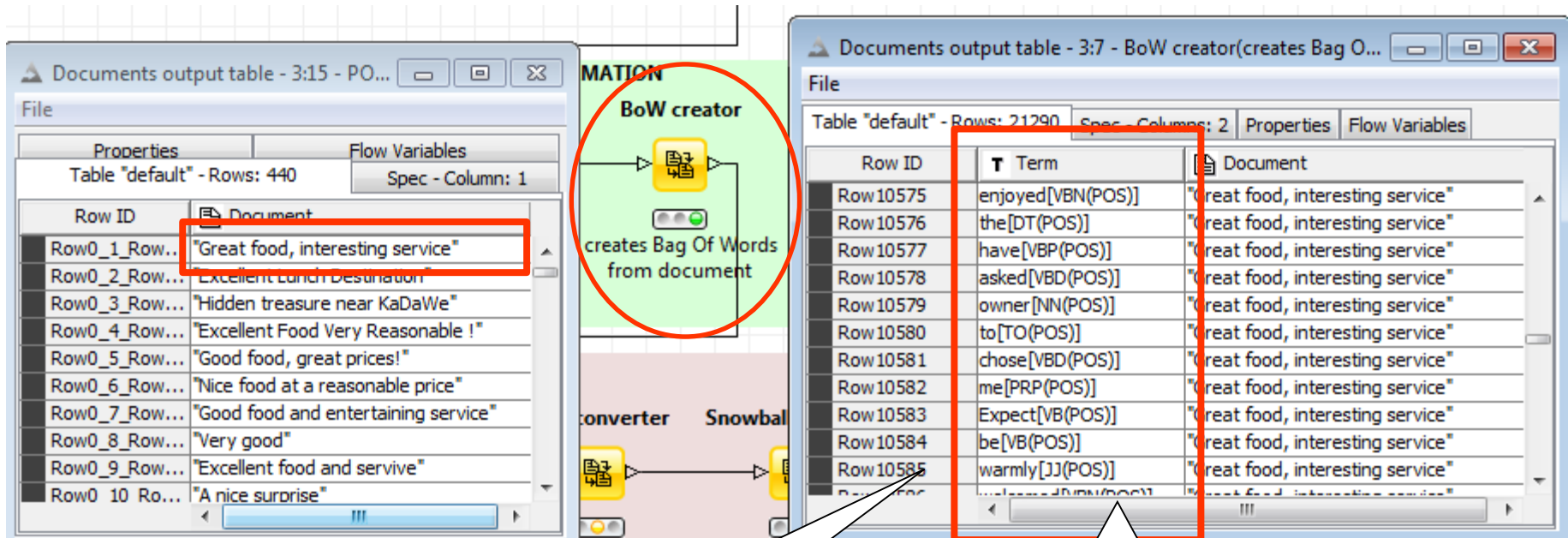
Parsing



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

Terms consist of tokens and tags

The Bag of Words



The screenshot shows the KNIME workflow interface. On the left, a 'Documents output table - 3:15 - PO...' window displays a list of documents. The first document, 'Great food, interesting service', is highlighted with a red box. In the center, a 'BoW creator' node is circled in red. A text box next to it states 'creates Bag Of Words from document'. On the right, a 'Documents output table - 3:7 - BoW creator(creates Bag O...' window displays a table with three columns: 'Row ID', 'Term', and 'Document'. The 'Term' column contains words with their part-of-speech tags, such as 'enjoyed[VBN(POS)]', 'the[DT(POS)]', and 'warmly[JJ(POS)]'. The 'Document' column contains the original text 'great food, interesting service'. A red box highlights the 'Term' and 'Document' columns, with a callout pointing to the text 'List (Bag) of Terms (Words) identified in Document'.

No Settings required

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

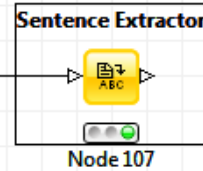
List (Bag) of Terms (Words) identified in Document

Data and Sentence Extractor

Filtered table - 3:5 - Column Filter(keep)

Table "default" - Rows: 440

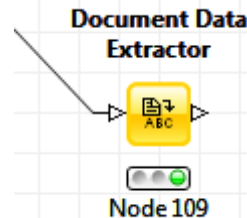
Row ID	Document
Row0_1_Row...	"Great food, interesting service"
Row0_2_Row...	"Excellent Lunch Destination"
Row0_3_Row...	"Hidden treasure near KaDaWe"
Row0_4_Row...	"Excellent Food Very Reasonable !"
Row0_5_Row...	"Good food, great prices!"



Documents and extracted sentences. - 3:107 - Sentence Extractor

Table "default" - Rows: 2361

Row ID	Document	Sentence
Row1	"Great food, interesting service"	Great food, interesting service
Row2	"Great food, interesting service"	this restaurant is seemingly known by Berliners fo...
Row3	"Great food, interesting service"	I enjoyed the food and I have asked the owner t...
Row4	"Great food, interesting service"	Expect to be warmly welcomed by the owner in a ...
Row5	"Excellent Lunch Destination"	Excellent Lunch Destination
Row6	"Excellent Lunch Destination"	Very much enjoyed our experience here.
Row7	"Excellent Lunch Destination"	Food was prompt, delicious and well presented.
Row8	"Excellent Lunch Destination"	I had earlier read about the owner's enthusiasm h...



Dialog - 3:109 - Document Data Extractor

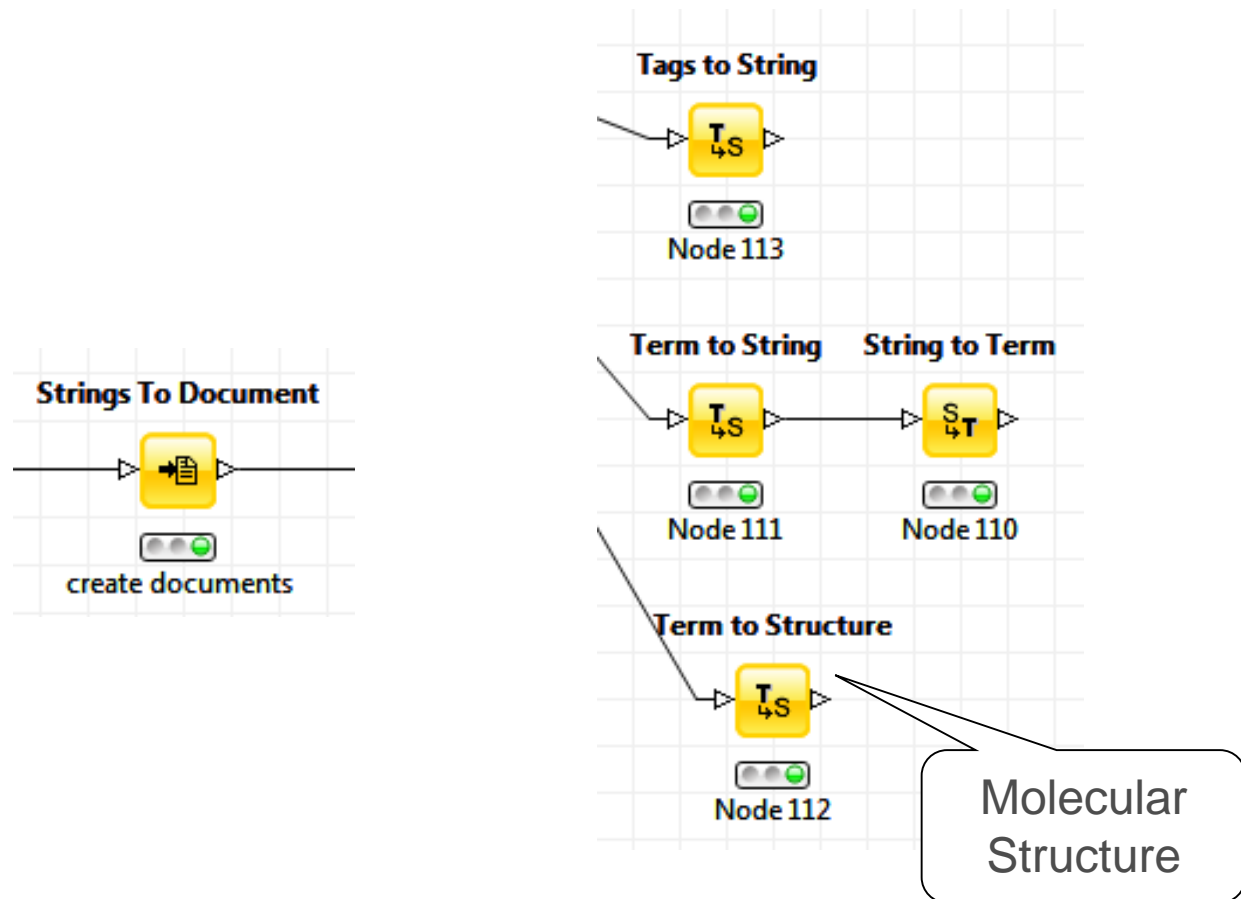
Document column: Document

Data extractors:

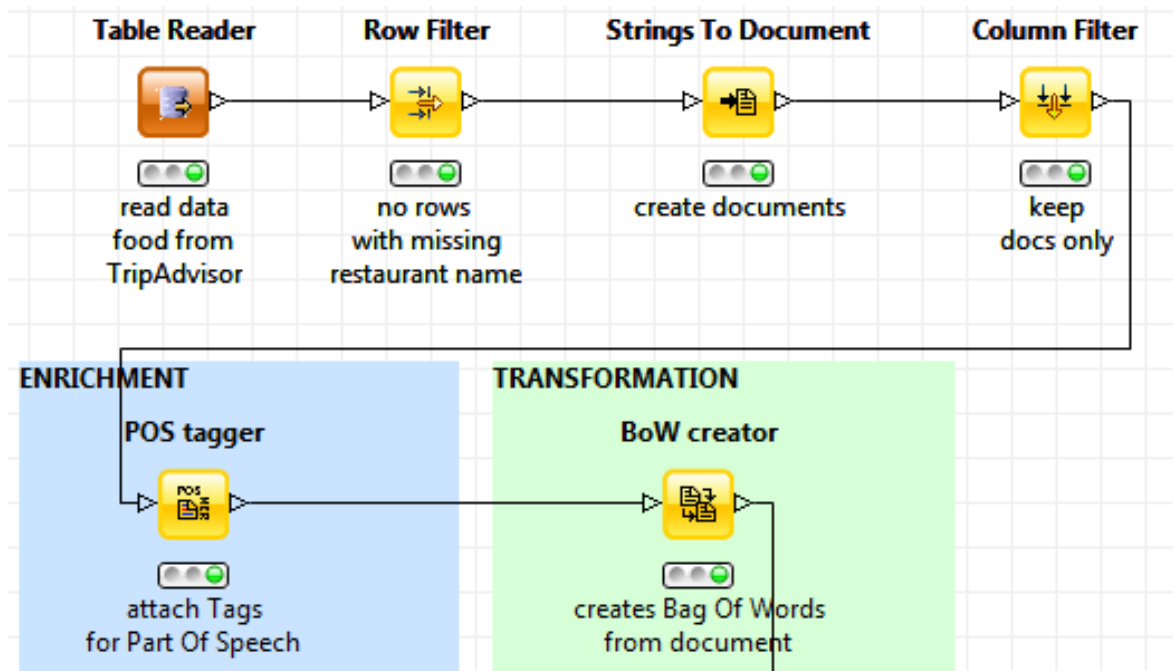
- Title
- Abstract
- Text
- Document body text
- Author
- Author set
- Category
- Category set

OK Apply Cancel ?

Conversions



Workflow



4 - Preprocessing

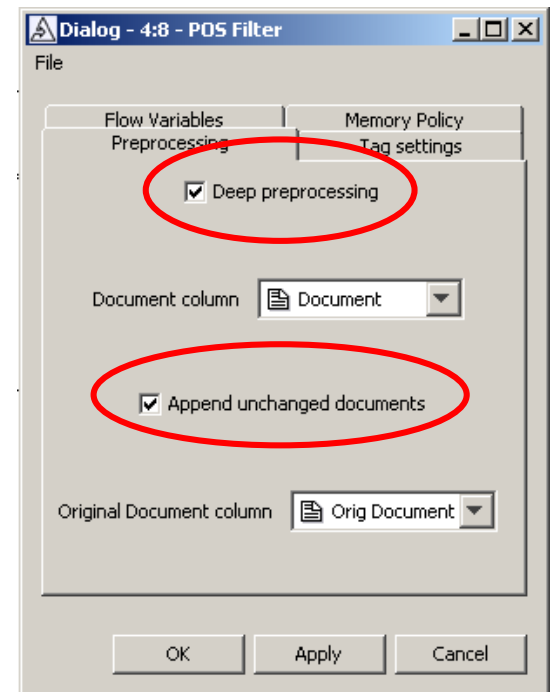
Preprocessing Mode

Normal

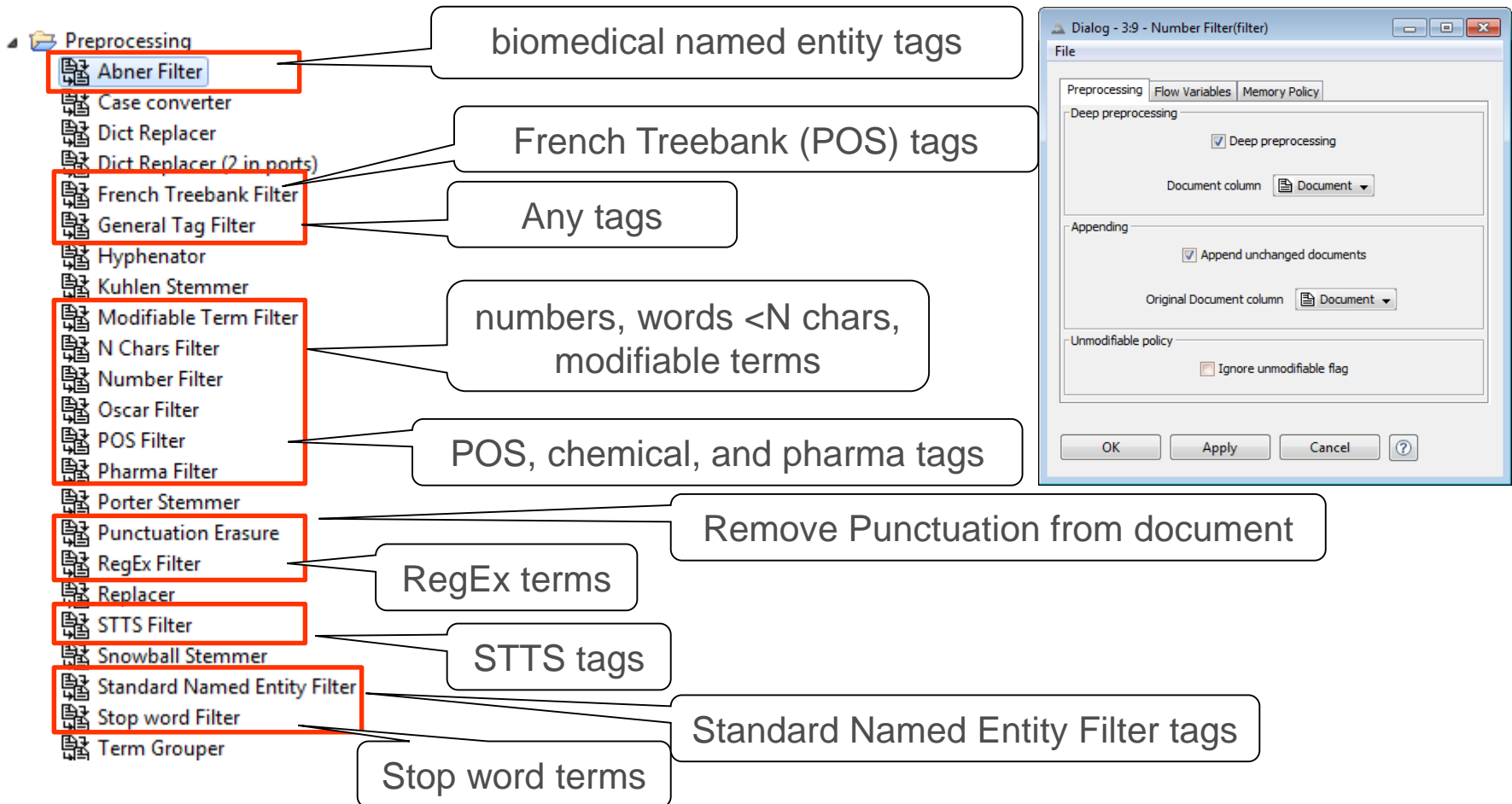
- Faster
- Preprocesses only terms of the term column
- Documents are not changed

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



The image shows a list of preprocessing filters in KNIME, with several highlighted in red. Callout boxes point to these filters, describing their functions. To the right, a screenshot of the 'Number Filter' dialog box is shown, which is used for filtering based on the number of characters in terms.

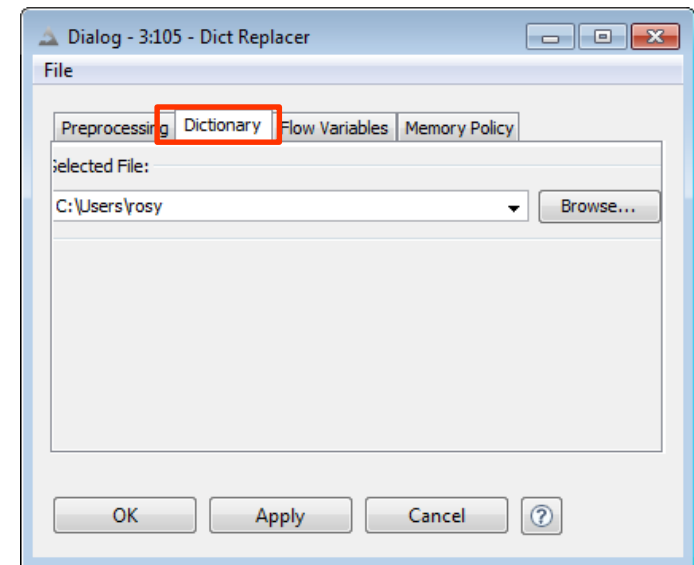
- Abner Filter**: biomedical named entity tags
- French Treebank Filter**: French Treebank (POS) tags
- General Tag Filter**: Any tags
- Modifiable Term Filter**: numbers, words <N chars, modifiable terms
- POS Filter**: POS, chemical, and pharma tags
- Punctuation Erasure**: Remove Punctuation from document
- RegEx Filter**: RegEx terms
- STTS Filter**: STTS tags
- Standard Named Entity Filter**: Standard Named Entity Filter tags
- Stop word Filter**: Stop word terms

The 'Number Filter' dialog box includes the following settings:

- Deep preprocessing: Deep preprocessing
- Document column: Document
- Appending: Append unchanged documents
- Original Document column: Document
- Unmodifiable policy: Ignore unmodifiable flag

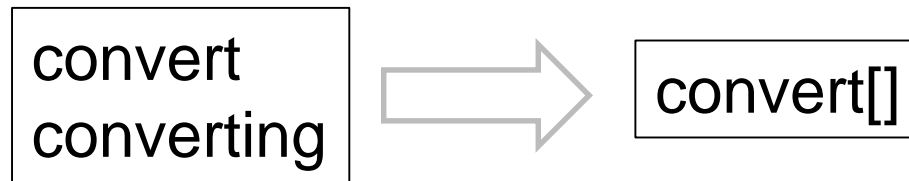
Converting and Replacing

- Preprocessing
 - Abner Filter
 - Case converter Term case conversion
 - Dict Replacer Term replacement with dictionary word
 - Dict Replacer (2 in ports)
 - French Treebank Filter
 - General Tag Filter
 - Hyphenator Introducing Hyphenation (Liang's algorithm)
 - Kuhlen Stemmer
 - Modifiable Term Filter
 - N Chars Filter
 - Number Filter
 - Oscar Filter
 - POS Filter
 - Pharma Filter
 - Porter Stemmer
 - Punctuation Erasure
 - RegEx Filter
 - Replacer RegEx based term replacer
 - STTS Filter
 - Snowball Stemmer
 - Standard Named Entity Filter
 - Stop word Filter
 - Term Grouper Groups rows by term value



Stemming

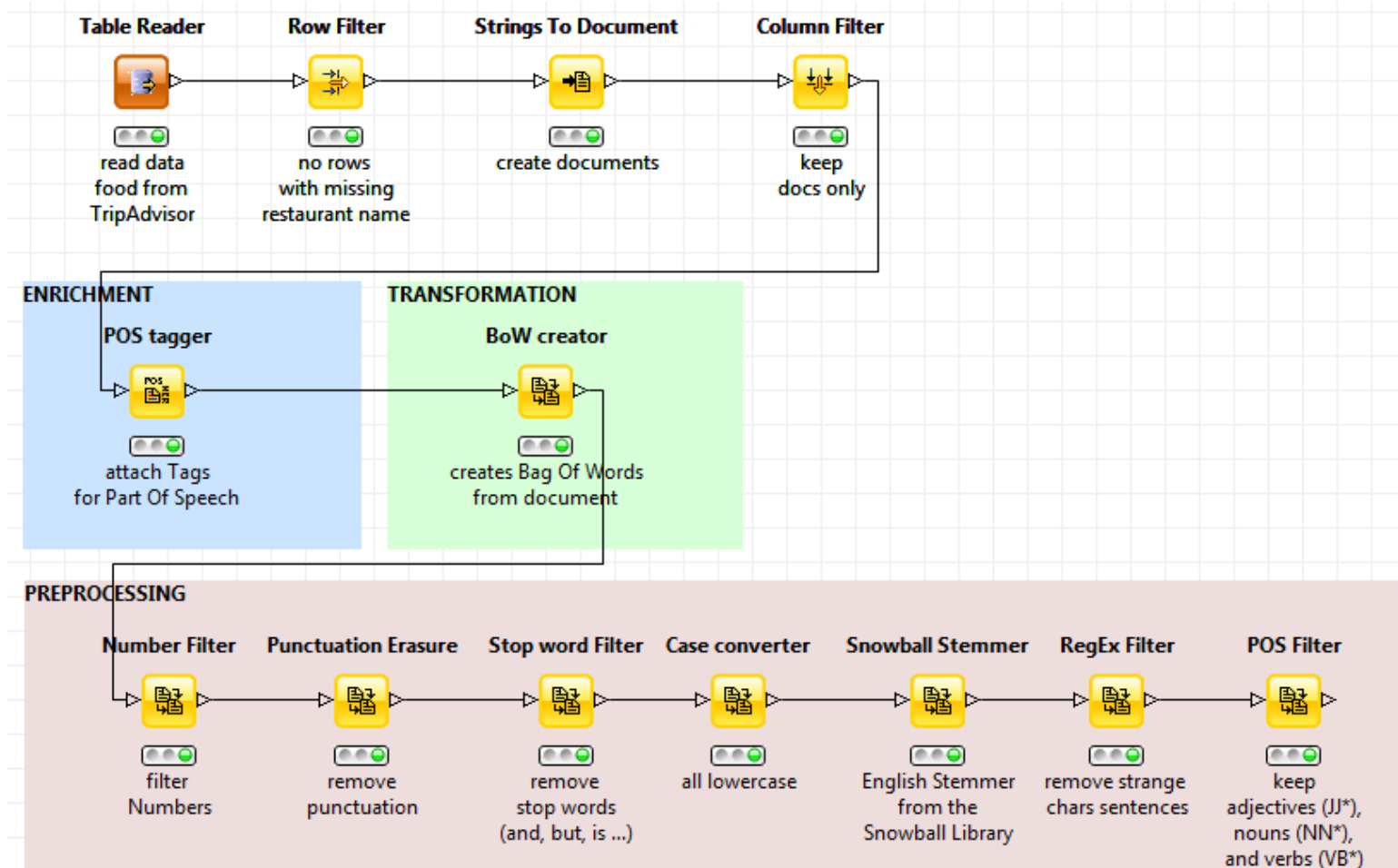
- Preprocessing
 - Abner Filter
 - Case converter
 - Dict Replacer
 - Dict Replacer (2 in ports)
 - French Treebank Filter
 - General Tag Filter
 - Hyphenator
 - Kuhlen Stemmer
 - Modifiable Term Filter
 - N Chars Filter
 - Number Filter
 - Oscar Filter
 - POS Filter
 - Pharma Filter
 - Porter Stemmer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



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

Stemmed term replaces original term!

Workflow

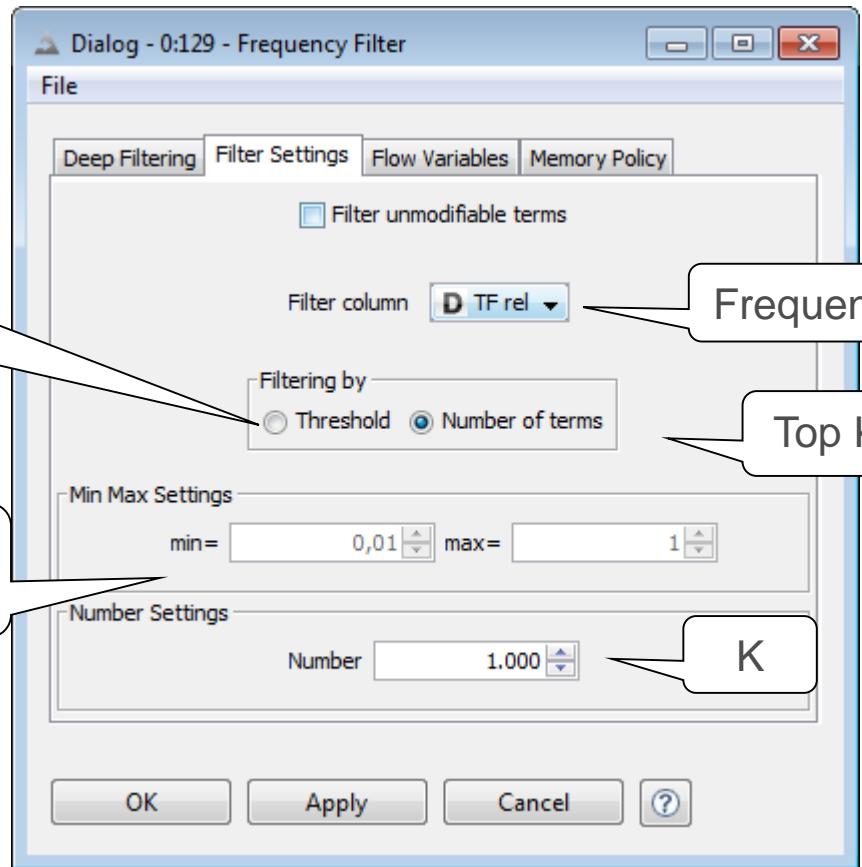


5 - Frequencies

Frequency Measures

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

Frequency Filter



Values between min threshold and max threshold

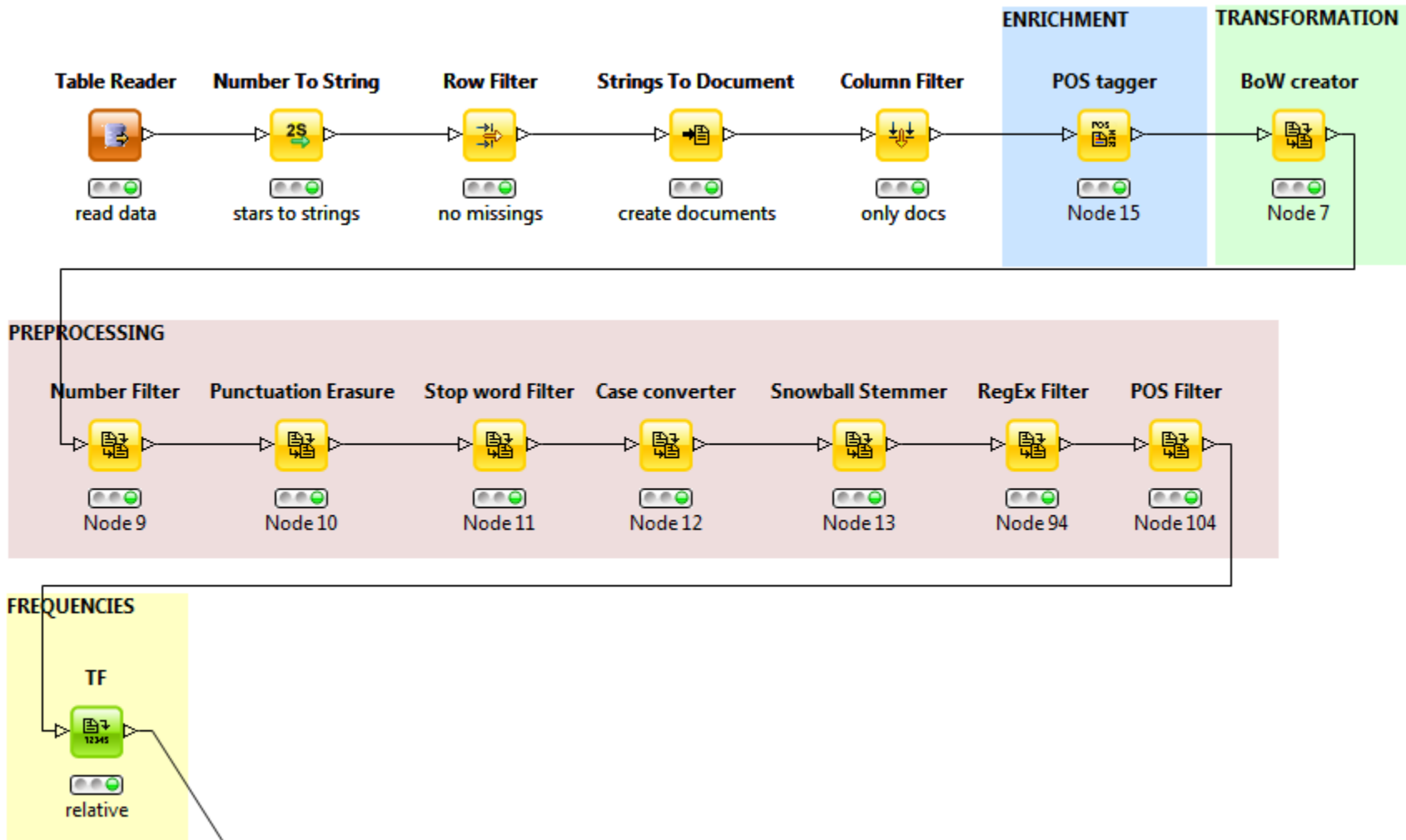
Frequency Column

Top K Rows

Min max thresholds

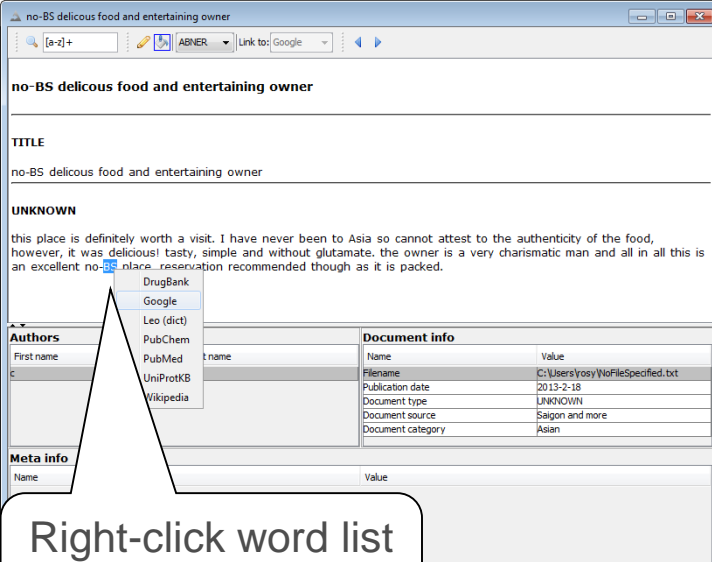
K

Workflow



6 - Visualization

Document Viewer



no-BS delicious food and entertaining owner

TITLE
no-BS delicious food and entertaining owner

UNKNOWN
this place is definitely worth a visit. I have never been to Asia so cannot attest to the authenticity of the food, however, it was delicious! tasty, simple and without glutamate, the owner is a very charismatic man and all in all this is an excellent no-BS place reservation recommended though as it is packed.

Authors

- DrugBank
- Google
- Leo (dict)
- PubChem
- PubMed
- UniProtKB
- Wikipedia

Document info

Name	Value
Filename	C:\Users\rosy\NoFilesSpecified.txt
Publication date	2013-2-18
Document type	UNKNOWN
Document source	Saigon and more
Document category	Asian

Meta info

Name	Value

Right-click word list search engines

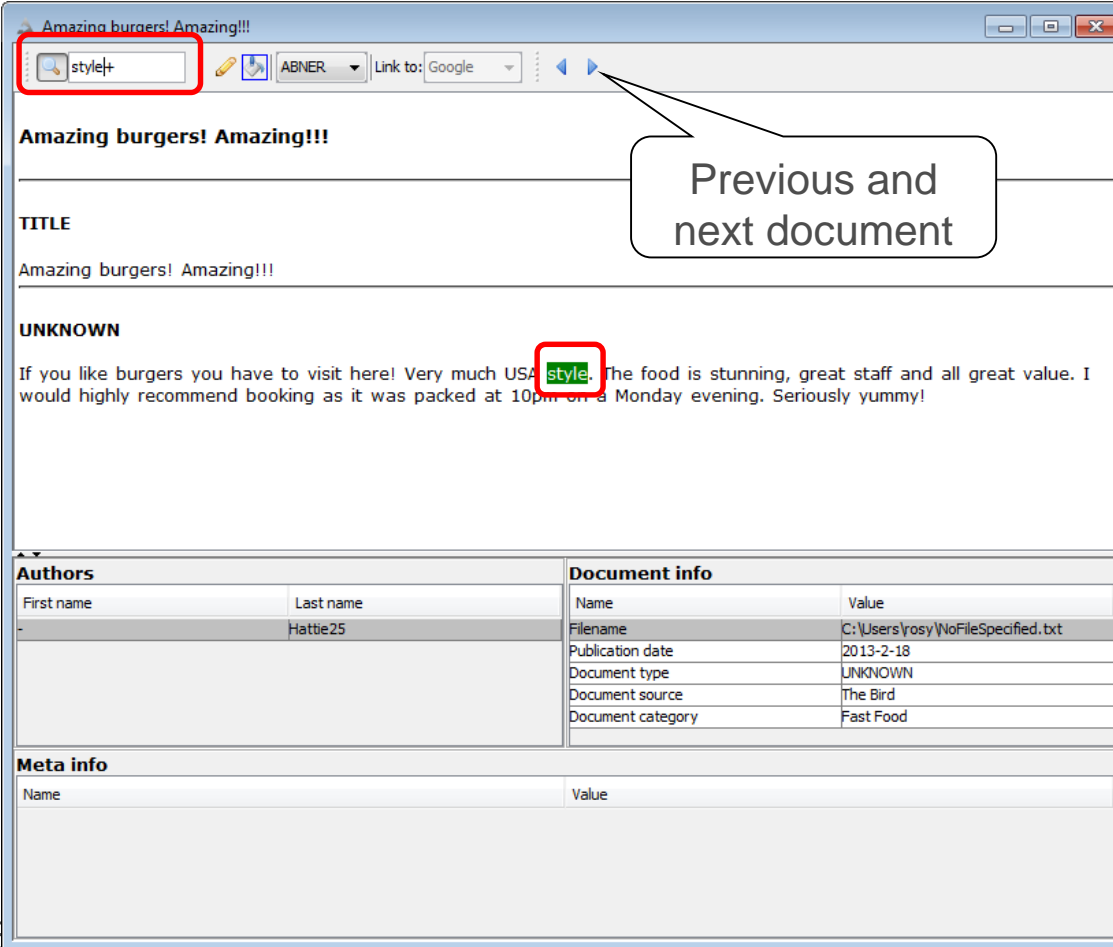
Search engines listed in KNIME->Textprecessing->Search Engine Preferences

Document View - 4:155 - Document Viewer

#	Document title	Autho...	Source	Cate...
	Great food, interesting service	- ltravelle...	Saigon an...	Asian
	Excellent Lunch Destination	- coverdri...	Saigon an...	Asian
	Hidden treasure near KaDaWe	- Lula12783	Saigon an...	Asian
	Excellent Food Very Reasonable !	- Deise_B...	Saigon an...	Asian
	Good food, great prices!	- Tanguya...	Saigon an...	Asian
	Nice food at a reasonable price	- OlgaNott...	Saigon an...	Asian
	Good food and entertaining service	Jack D	Saigon an...	Asian
	Very good	Rick M	Saigon an...	Asian
	Excellent food and serve	- Alexg123	Saigon an...	Asian
0	A nice surprise	- stabreim	Saigon an...	Asian
1	Ok but not that great.	- mooniesss	Saigon an...	Asian
2	This is.. BY FAR.. my favorite restaurant in Berlin	- FoDTours	Saigon an...	Asian
3	Tasty, good looking food. Perfect English speaking staff	- Roderoro	Saigon an...	Asian
4	Good food with warm welcome.	Lusiana M	Saigon an...	Asian
5	So happy to have found this jewel!	- Esther-H...	Saigon an...	Asian
5	lived up to the hype	- Tbwduke	Saigon an...	Asian
7	Saigon is in Berlin	- Foodiev...	Saigon an...	Asian
8	Great experience, great food!	- Takidk	Saigon an...	Asian
9	Great value, good food, interesting owner.	- Lolcatz	Saigon an...	Asian
20	Ok food, entertaining owner but service is average	- innajl	Saigon an...	Asian
21	no-BS delicious food and entertaining owner	c a	Saigon an...	Asian
22	Excellent	Stuart M	Saigon an...	Asian
23	Wow!	Radu S	Saigon an...	Asian
24	Very good and tasty food	- BernrS	Saigon an...	Asian
25	Meet the owner	- Bartalk	Saigon an...	Asian
26	over hyped but Ok for a cheap eat	- travelsuit	Saigon an...	Asian
27	Best asian in a long time	- sigul	Saigon an...	Asian
28	Best restaurant in Berlin. What more do you need	- YellowDu...	Saigon an...	Asian
29	The worst Vietnamese restaurant...	Clive C	Saigon an...	Asian
30	Good - but not outstanding - worth a visit if you	- TheReal...	Saigon an...	Asian
31	Amazing Asian Food!	- mdcwilson	Saigon an...	Asian
32	Yummy!	- MyrtrwG	Saigon an...	Asian
33	Well worth a visit if you are in Berlin.	- Lazinspain	Saigon an...	Asian
34	Current ranking deserved	- ajh47	Saigon an...	Asian
35	very good food	- kksechick	Saigon an...	Asian
36	Good Vietnamese food with great service	Mounti	Saigon an...	Asian

Double-click opens the document

Document Viewer



Amazing burgers! Amazing!!!

style+

ABNER Link to: Google

Previous and next document

Amazing burgers! Amazing!!!

TITLE

Amazing burgers! Amazing!!!

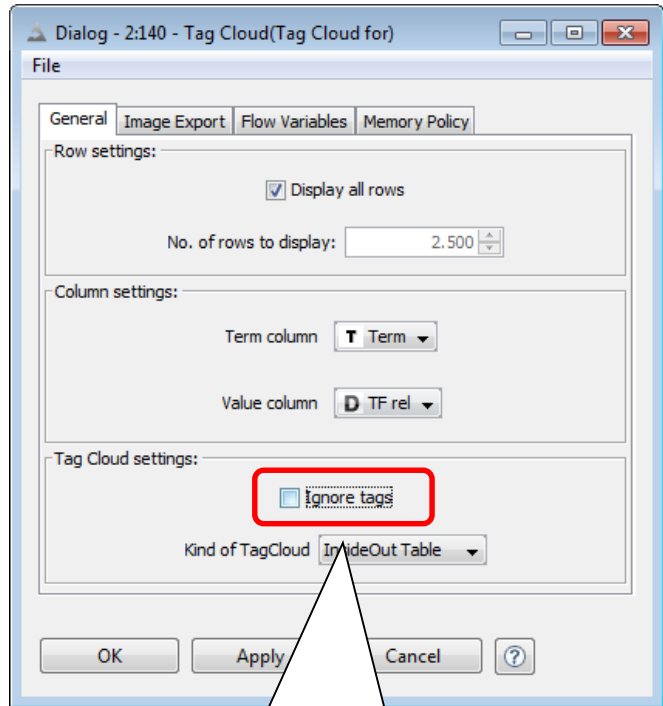
UNKNOWN

If you like burgers you have to visit here! Very much USA **style**. The food is stunning, great staff and all great value. I would highly recommend booking as it was packed at 10pm on a Monday evening. Seriously yummy!

Authors		Document info	
First name	Last name	Name	Value
-	Hattie25	Filename	C:\Users\yosy\NoFileSpecified.txt
		Publication date	2013-2-18
		Document type	UNKNOWN
		Document source	The Bird
		Document category	Fast Food

Meta info	
Name	Value

Tag Cloud



Dialog - 2:140 - Tag Cloud(Tag Cloud for)

File

General Image Export Flow Variables Memory Policy

Row settings:

Display all rows

No. of rows to display: 2.500

Column settings:

Term column T Term

Value column D TF rel

Tag Cloud settings:

Ignore tags

Kind of TagCloud InsideOut Table

OK Apply Cancel ?

Adj, verbs and nouns as same word



Tagcloud - 2:140 - Tag Cloud(Tag Cloud for)

File HiLite

Default Settings: Color Legend Font Style

VBN	Change...	VB	Change...	VBP	Change...
VBZ	Change...	VBG	Change...	VBD	Change...
NNS	Change...	NNP	Change...	NN	Change...
JJ	Change...	JJS	Change...	JJR	Change...
NNPS	Change...				

Asian Restaurants



German Food Restaurants



Hiliting in Tag Clouds



Tagcloud - 2:44 - Tag Cloud(Tag Cloud for)

File | HiLite

- HiLite selected
- UnHiLite selected
- Clear HiLite

vietnamis

Other tags: thai, berlin, time, night, curri, warm, return, neighborhood, person, issu, smile, tai, colder, confirm, globe, pa...

Interactive Table

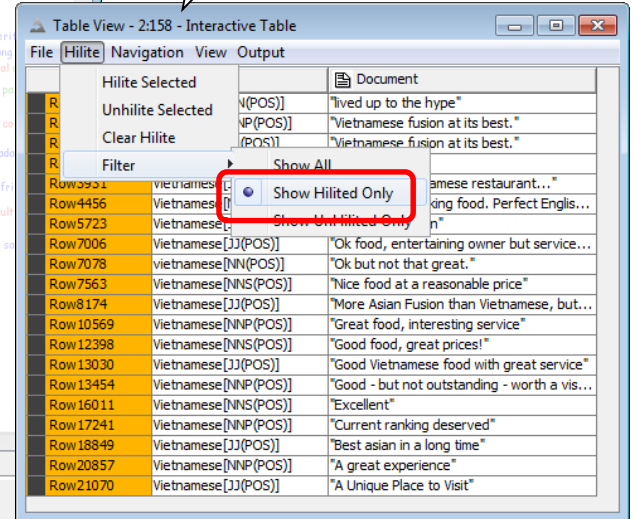
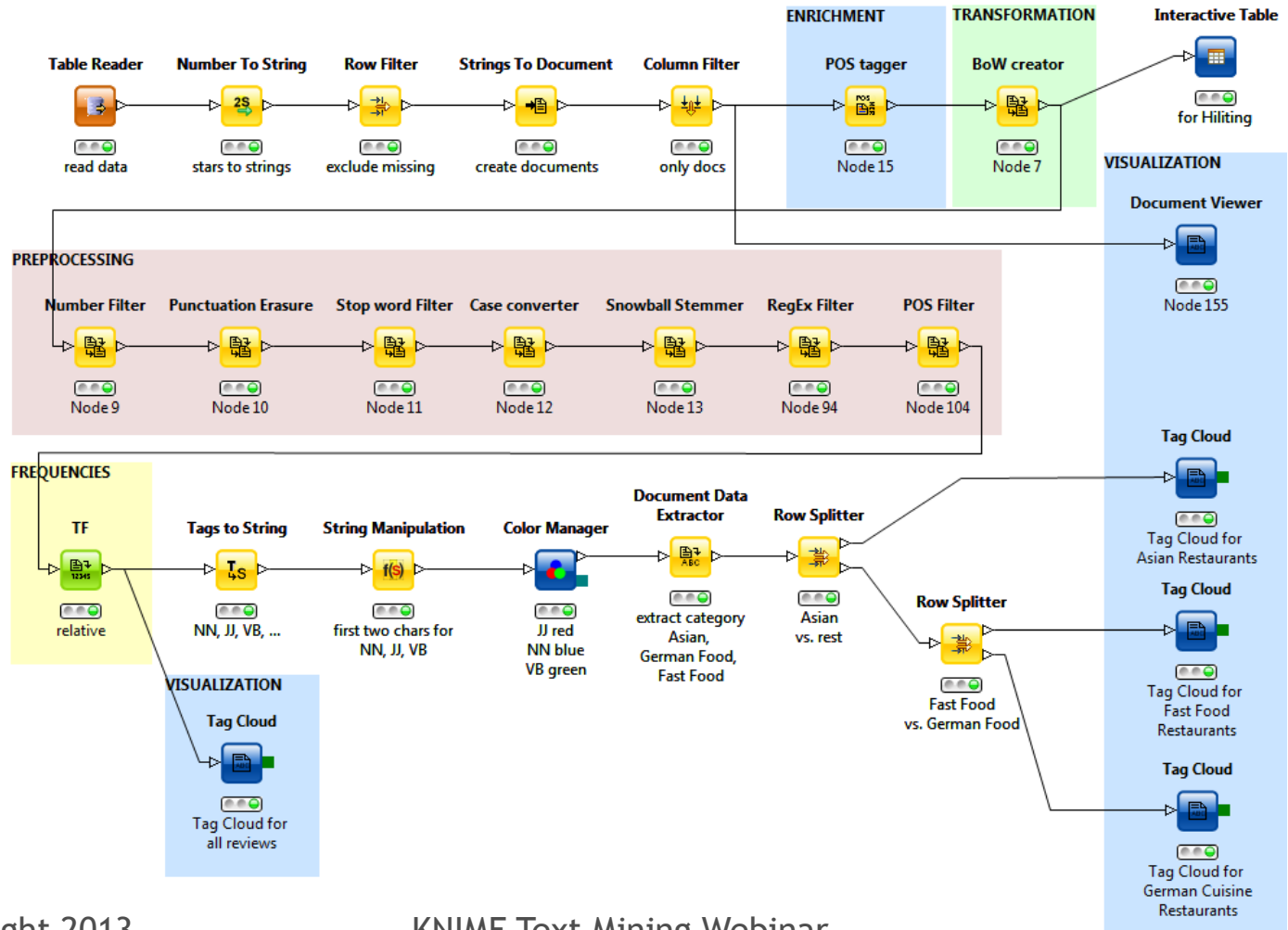


Table View - 2:158 - Interactive Table

File | Hilite | Navigation | View | Output

	Hilite Selected	Document
R	Unhilite Selected	N(POS) "lived up to the hype"
R	Clear Hilite	NP(POS) "Vietnamese fusion at its best."
R	Filter	(POS) "Vietnamese fusion at its best."
	Filter	Show All
		Show Hilited Only
		Show UnHilited Only
Row2931	vietnames	amese restauran..."
Row4456	Vietnames	king food. Perfect Englis...
Row5723	Vietnamese	n"
Row7006	Vietnamese [JJ(POS)]	"Ok food, entertaining owner but service..."
Row7078	vietnamese [NN(POS)]	"Ok but not that great."
Row7563	Vietnamese [NNS(POS)]	"Nice food at a reasonable price"
Row8174	Vietnamese [JJ(POS)]	"More Asian Fusion than Vietnamese, but..."
Row10569	Vietnamese [NNP(POS)]	"Great food, interesting service"
Row12398	Vietnamese [NNS(POS)]	"Good food, great prices!"
Row13030	Vietnamese [JJ(POS)]	"Good Vietnamese food with great service"
Row13454	Vietnamese [NNP(POS)]	"Good - but not outstanding - worth a vis..."
Row16011	Vietnamese [NNS(POS)]	"Excellent"
Row17241	Vietnamese [NNP(POS)]	"Current ranking deserved"
Row18849	Vietnamese [JJ(POS)]	"Best asian in a long time"
Row20857	Vietnamese [NNP(POS)]	"A great experience"
Row21070	Vietnamese [JJ(POS)]	"A Unique Place to Visit"

Workflow

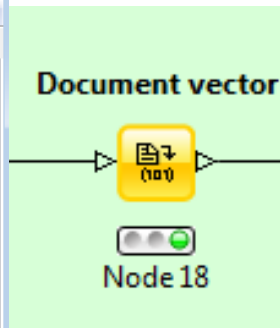


7 - Topic Classification

Document Vector

Terms and documents output table - 0:16 - TF(TF relative)

Row ID	T Term	Document	Orig Document	D TF rel
Row3	german[JJ(PO...]	"german pub"	"what a German pub should ...	0.118
Row4	pub[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row8	servic[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row11	atmosph[er][NN...]	"german pub"	"what a German pub should ...	0.059
Row12	food[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row14	beer[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row18	sausage[NN(S(P...]	"german pub"	"what a German pub should ...	0.118
Row19	potato[NN(S(P...]	"german pub"	"what a German pub should ...	0.059
Row21	mustard[NN(P...]	"german pub"	"what a German pub should ...	0.059
Row23	tast[VBD(POS)]	"german pub"	"what a German pub should ...	0.059



Documents output table - 0:18 - Document vector

Row ID	D visitor	D play	D difficult	D classic	D comfort	D unobtrus	D sausage	D class
429	0	0	0	0	0	0	0	0
430	0	0	0	0	0	0	0	0
431	0	0	0	0	0	0	0	0
432	0	0	0	0	0	0	0	0
433	0	0	0	0	0	0	0	0
434	0	0	0	0	0	0	0	1
435	0	0	0	0	0	0	0	0
436	0	0	0	0	0	0	0	0
437	0	0	0	0	0	0	0	0
438	0	0	0	0	0	0	0	0
439	0	0	0	0	0	0	0	0
440	0	0	0	0	0	0	1	0

Document Vector: Documents represented in the terms space

Dialog - 0:18 - Document vector

Options | Flow Variables | Memory Policy

Document column: **Orig Document**

Ignore tags

Bitvector

Vector value: **D TF rel**

As collection cell

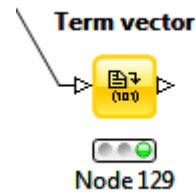
OK Apply Cancel ?

Bitvector or frequency measure

Term Vector

Terms and documents output table - 0:16 - TF(TF relative)

Row ID	T Term	Document	Orig Document	D TF rel
Row3	german[JJ(PO...]	"german pub"	"what a German pub should ...	0.118
Row4	pub[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row8	servic[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row11	atmosph[er][NN...]	"german pub"	"what a German pub should ...	0.059
Row12	food[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row14	beer[NN(POS)]	"german pub"	"what a German pub should ...	0.059
Row18	sausag[er][NNS(P...]	"german pub"	"what a German pub should ...	0.118
Row19	potato[NN(S(P...]	"german pub"	"what a German pub should ...	0.059
Row21	mustard[NN(P...]	"german pub"	"what a German pub should ...	0.059
Row23	tast[VBD(POS)]	"german pub"	"what a German pub should ...	0.059



Terms output table - 0:129 - Term vector

Row ID	T Term	(...) Term Vector
127	beauti[NNP(...]	[0.0,0.0,0.0,0,...]
128	beef[NN(POS)]	[0.0,0.0,0.0,0,...]
129	beer[JJ(POS)]	[0.0,0.0,0.0,0,...]
130	beer[NN(POS)]	[0.0,0.0,0.0,0,...]
131	beerdik[NN(...]	[0.0,0.0,0.0,0,...]
132	beetroot[VB...]	[0.0,0.0,0.0,0,...]
133	believ[VBP(P...]	[0.0,0.0,0.0,0,...]
134	bell[NNP(POS)]	[0.0,0.0,0.0,0,...]
135	helli[NN(POS)]	[0.0,0.0,0.0,0,...]

Term Vector: Terms represented in the documents space

Bitvector or frequency measure

Dialog - 0:129 - Term vector

Options | Flow Variables | Memory Policy

Document column:

Ignore Tags

Bitvector

Vector value:

As collection cell

OK Apply Cancel ?

Topic Detection Goal

Possible Topics:

- Asian Restaurants
- German Food
- Fast Food

Pre-labeled data set available!

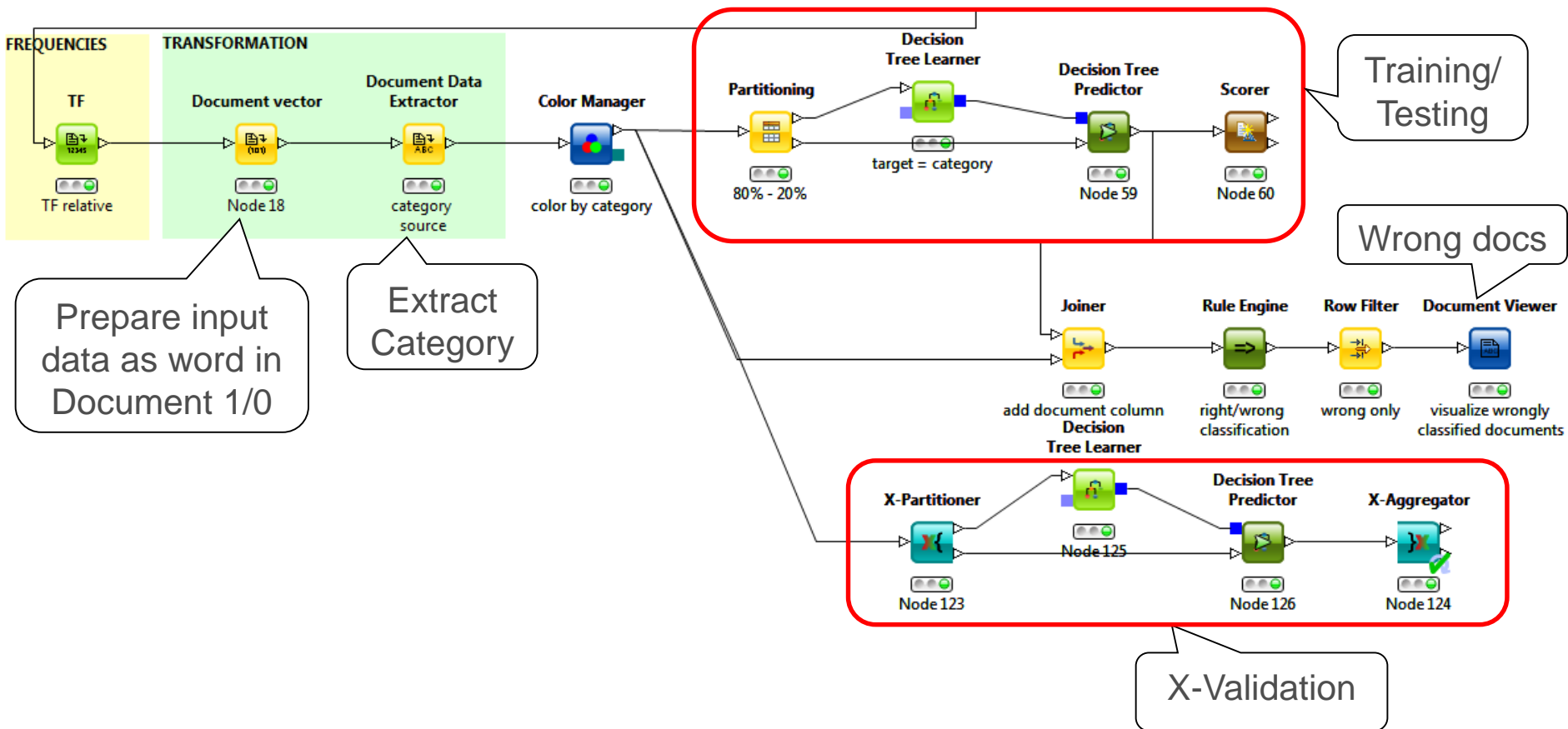
Target Topic is in Document Category.

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

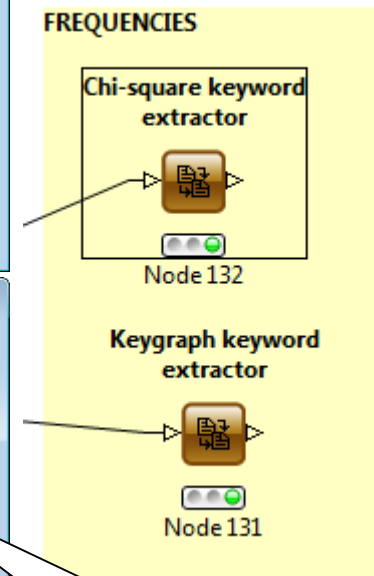
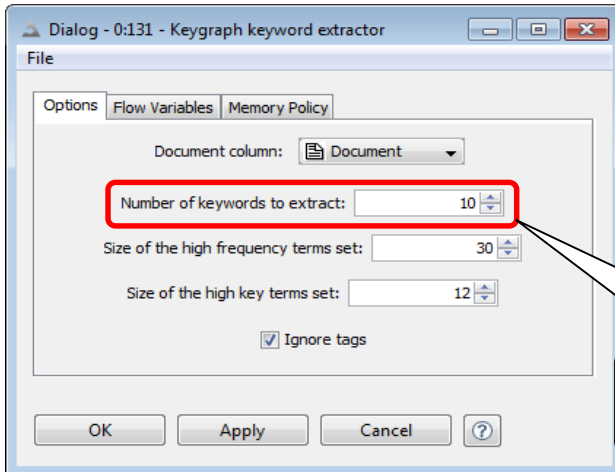
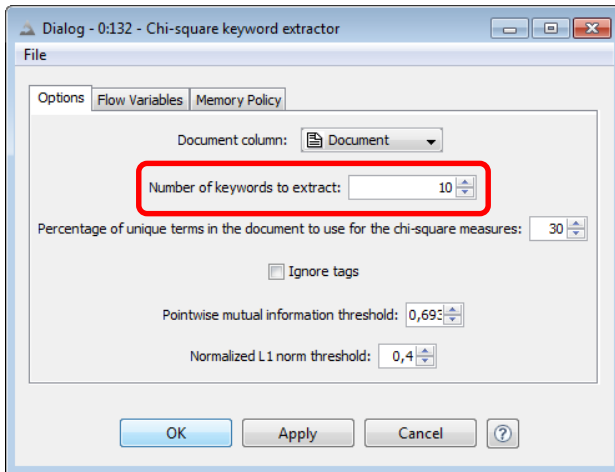
Problem 2

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



From:

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

From:

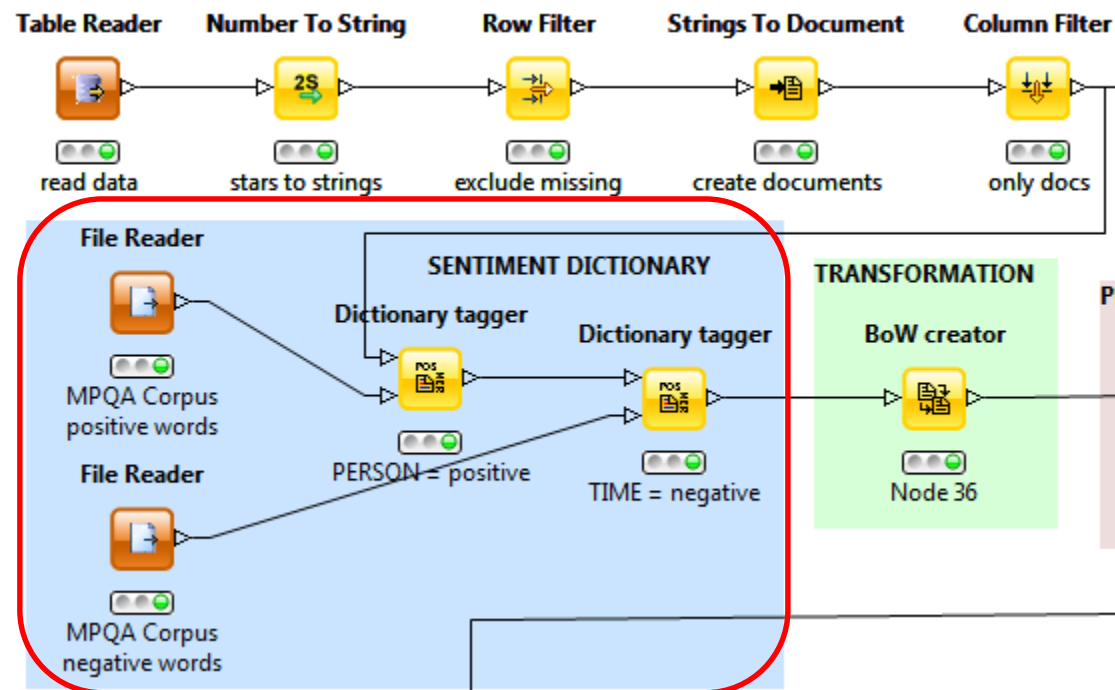
"KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor" by Yukio Ohsawa.

Max. # keywords per doc

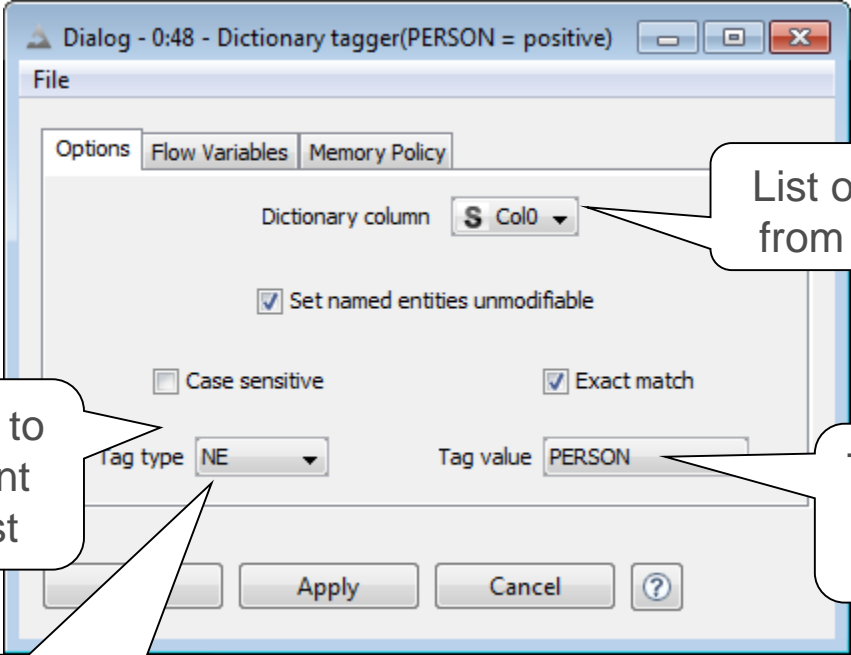
8 - Sentiment Analysis

Sentiment Corpus

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



Dictionary Tagger Node



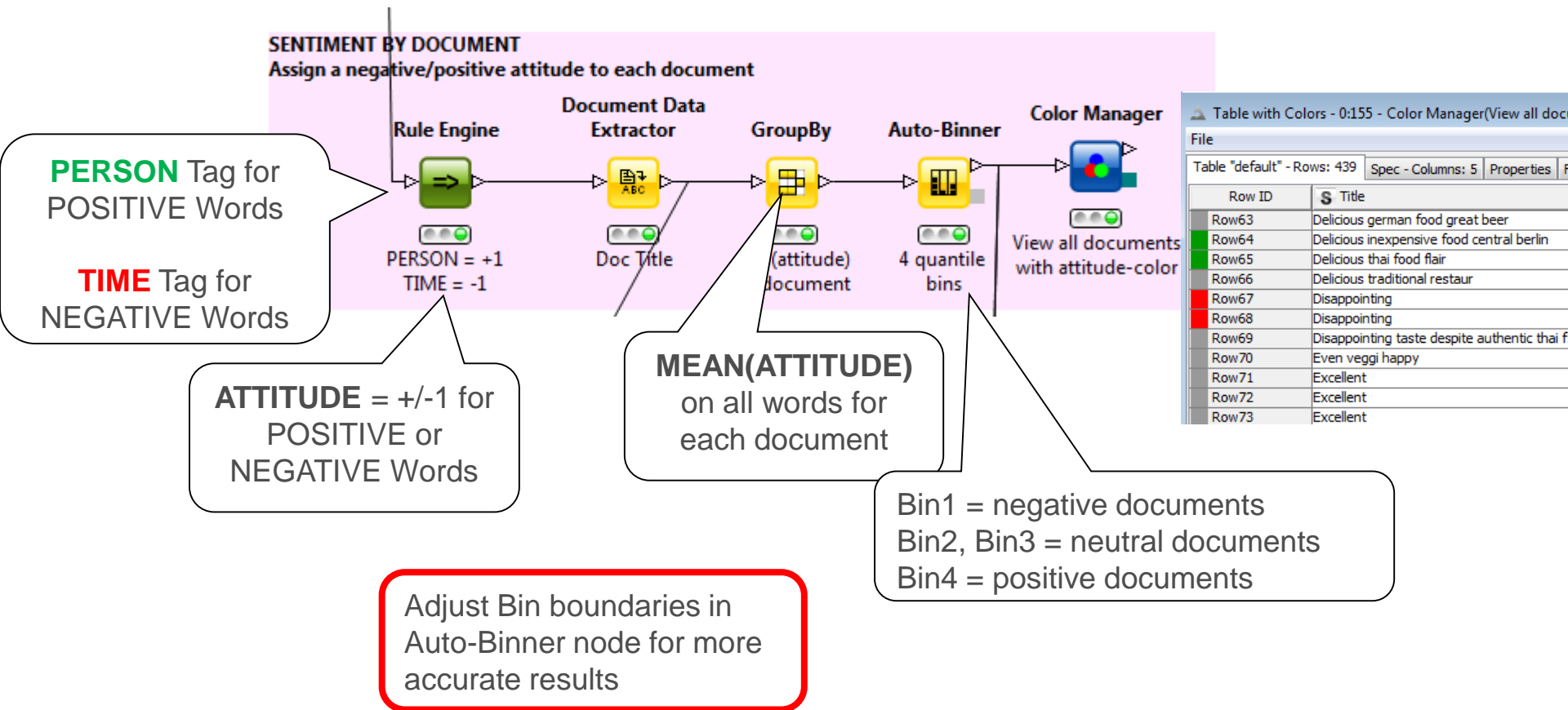
The screenshot shows the 'Dictionary Tagger' dialog box with the following settings and callouts:

- Dictionary column:** Set to 'Col0'. Callout: 'List of Words from Corpus'.
- Set named entities unmodifiable:** Checked.
- Case sensitive:** Unchecked.
- Exact match:** Checked.
- Tag type:** Set to 'NE'. Callout: 'Tag Type to attach to Words in document and in Corpus List'.
- Tag value:** Set to 'PERSON'. Callout: 'Tag Value to attach to Words in document and in Corpus List'.

Additional callouts:

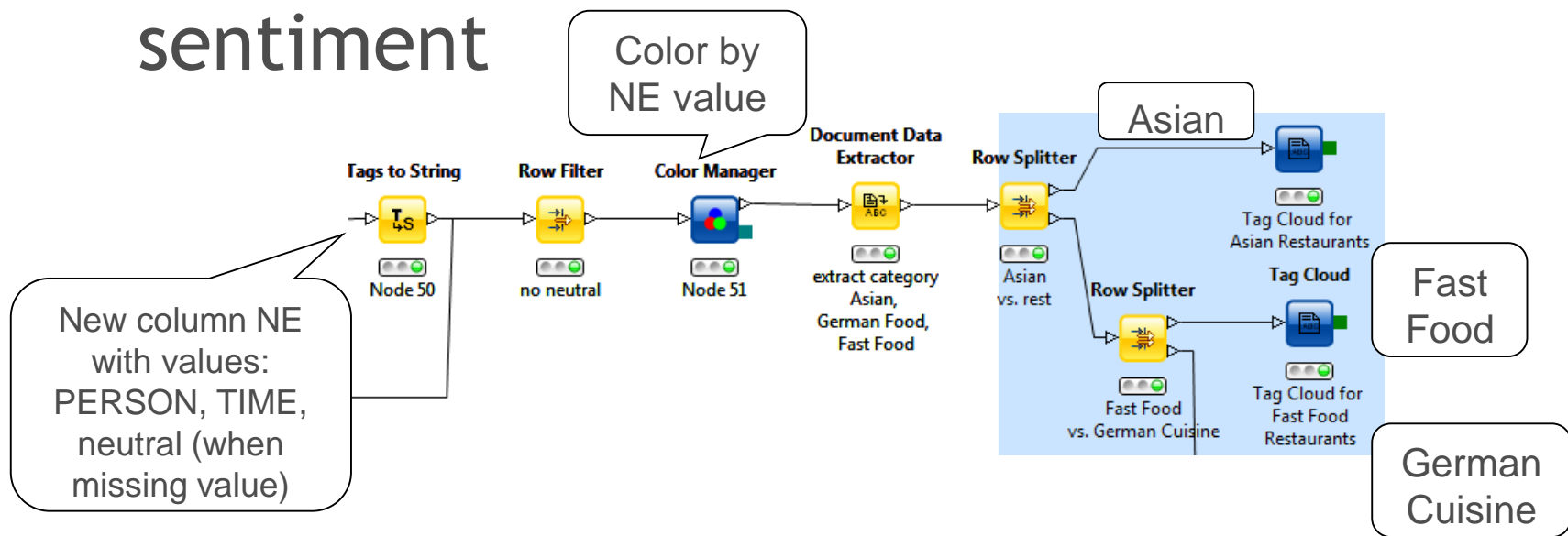
- 'NE = Named Entities' (pointing to the 'Tag type' dropdown).
- 'Each Tag Type has a list of possible Tag Values' (pointing to the 'Tag value' field).

Sentiment By Document



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

Close Pretty handy capable wise abundance fair Beautifully
 Enthusiastic Sophisticated healthy cold wrong shy want Worth exemplary Awesome
 Think insanely complimentary Better skeptical poor romantic understated Will excessive truth
 Unbelievably lucky correctly Fabulous slow prefer unfortunately pleasure regret try Usually exceptional sincere
 Fine memorable cozy Fantastic spot prompt worst cross stunning pleasing quaint timid Solid enthusiastic passion
 deserved Inconsistent traditional real Amazing friends smack Well salute perfectly superb Simple ensure fit
 disagree trick recommendation Bland happy helpful cheerfulness Delicious problem miss problems solid engaging
 underestimate Disappointing help humour convenient too extremely perfect Best positive lack prize smile
 uniformly interested beautifully reasonably warm worth cheap special need worryingly popular intrusive
 unlikely knife despite large entertaining like Good well awesome fancy lively warmly perfection
 TRY might simple sure Excellent just nice little even taste expert fun unique open
 live recommend excellent good friendly value efficient
 Word prepared close will welcome back good delicious reasonable Nice super
 smart Flexible pretty right gem busy fresh great best authentic disappointed lies down
 okay drive advice worry sunny Gorgeous rightly knowledgeable seasoned Reasonable disservice enjoy
 Attentive Just thirst although try definitely Great pleasant amazing Definitely established Joy
 basic Dear Extremely enthusiasm above better love lovely recommended wonderfulespecially friendliness
 plain clean Friendly Fresh friend lacking expensive attentive least easily recommendations famous impressions
 polite favorite courteous Mediocre Superb funny less Fabulous unfriendly praise fair free jovial relax Perf
 gracious critical Unfortunately Try interesting outstanding mediocre cut fortunate hungry proud AWESOME
 Unique poorly hilarious disappointment attest Wonderful long comforting eager hesitant preferably wow difficult
 Wow WILL satisfying honest exhaustive craving bland disappointing haphazard plenty worries clearly disappoint
 accept Waste suspect impress forget delicious genuine pain wonder charming comfortable
 agree attractive sweet inoffensive generous original unbelievably charismatic care
 amazing battle terrible moderate treasure brilliant bad

Fast Food

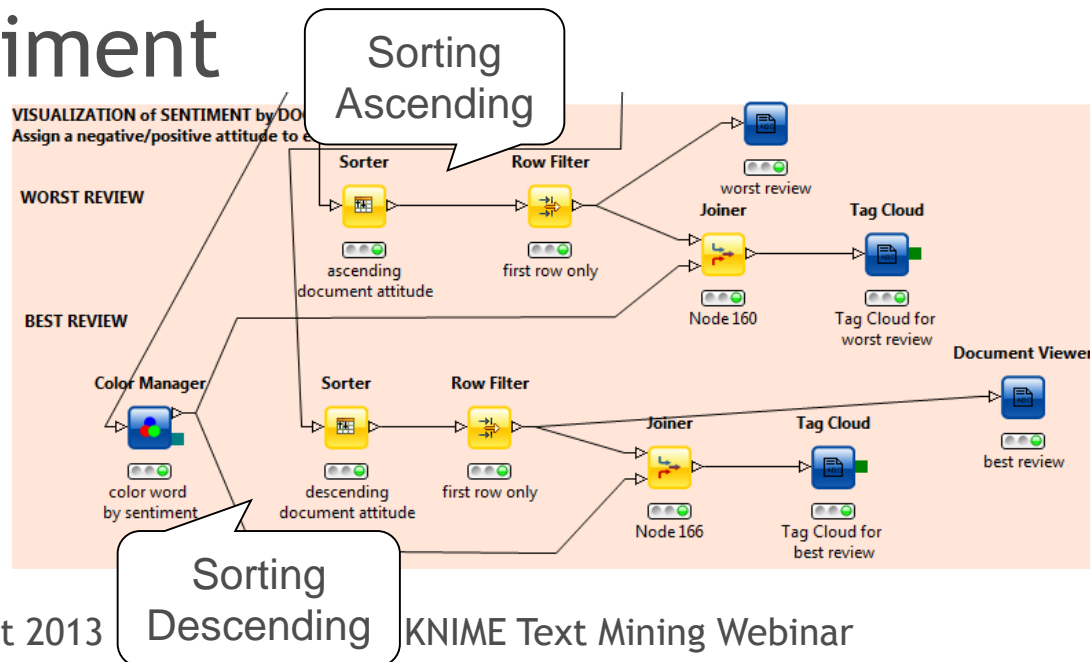
important Large STARS
 Back Level AMAZING Hope
 Disappointing Score BEST gem whatever Fine
 Fatty Seriously Brilliant happily Too fair welcome Expensive
 LOVE Ultimately Classic helpful comfort pain Quar extremelydearly truly Easily
 Lead beautiful Clearly hilarious criticism poor easily lively HOT excitement suggest Cheap
 Sadly cold DELICIOUS hot difficult real fall impressive complaints highlight Filthy elevated strong
 disappoint Fantastic lack enjoy reason loyal less Nice heart clever despite word downside stars Although
 lacking frozen supposed mind need beloved recommended Just decent clear close utterly disappointing
 knife understand miss polite grill reservations value happy terrible clearly although chaos cute star
 okay overwhelmingly slow least interesting will love Excellent crowded shake Yes Awesome care
 Decent special perfect clean down Best worth amazing cheap attractive ready taste yes
 fantastic fresh too friendly great like well Try superb limited smart
 lover perfectly Perfect positive stunning might super help popular wonderfully sickly
 messy friends delicious Great good best just Heaven definitely
 treat Delicious lost awesome busy little sure bad fuss pleasant vibrant
 pricey ultimate expensive simple even bar nice back better right wonderful fine open
 Special problem attentive spot tired long excellent try pretty top brilliant negative prepared
 Unfortunately relaxed efficient cut Good recommend Worth large FUN lovely noisy Noisy craving
 Fresh herbaceous refreshing Vile sluggish hungry easy fun kind Amazing extremely manic Easy unpretentious
 Suger Fun normal mid advanced unable impression hard wide Please low want unique charming serious slack
 disappointed GOOD perfection rules shame wrong sound Fat incredible trying terrific break raving ridiculous
 Well entertaining Incredible plain spectacular classic Even doubt true tantalizing basic paradise overbearing
 above especially Lovely problems stand condescending three sweet Wonderful odd original
 disappointment excited Waving reasonable tender speedy Unique importantly memorable
 Everest's fault Open refuse simple hope high
 Free Friend Phenomenal about heaven

German

strange
 tender fuzzy popular
 trendy glad beautiful fun incredibly
 true speedy Best smile Admittedly famous honest
 ultimate hard Busy steady spirit slow worse fair gem
 worthwhile humorous Pricey strong truly low sady profane weird exuberant divine
 Affordable hungry Unfortunately struggle welcome reason Friend hospitable playful pricey unobtainable erratic discreet
 Attentive ignore Unique super word skeptical genuine exceptional flawless desperate last polite tire enjoyable complain
 Awesome impressive Want surely Genuine treat heart stew black crowded dream comfortable lively perfectly terrible easily
 Wide sweet Truly unforgettable less Well disappointed refined awesome avoid decent close historic patient superbly darn
 terrific appalling unusual prepared favorite especially top perfect need Superb Cozy charming agree heavily painful stumble
 hunt profit interesting extremely Nice recommended Good warm difficult sure truth Stars Traditional fat loopy
 celebratory wish rude spot free bar Excellent stars too just special friends superb stiff wide Regal fairly
 stunning limited pleasant beautiful star nice great like better helpful elegant steal spacious
 marginally excellent reasonable back well good delicious best little unique mind
 Fancy simple wonderful value Fantastic although night definitely live real relaxed Lovely exquisite
 Perfect tradition able want kind will traditional Great friendly long disappointment cheap large
 worsly trying commitment affordable miss authentic lovely worth bad Delicious fantastic least reasonably GREAT
 wrong cold Star Amazing disturbed attractive recommend even amazing taste efficient gorgeous rage useful deserved
 Friendly competent beautifully Definitely enjoy expensive try busy dark eclectic problem respect delighted irritation stuffy
 Bad Worth Lonely disappointing bland Super recommendation fresh Try down poor promising deft intolerance splendid compromise
 Erratic lanky amusing Frilly excited clear attentive regret creative plenty pretty bulky hat slightly clearly afford
 Grand night celebration Reasonable fine cozy comfort intrusive noisy appropriate hopeful satisfactory classic Tern
 Help nasty correct Straight kid incredible lucky appealing happy remarkable cheerful Sport
 Fear abstruse dance Top lack omnia gladly relief bitter Perfection
 Cheap Splendid perfection easy above entrance proper appreciative Expensive
 disappoint Wonderful pleasing embarrassment positive amaze solid
 marvelous accept plush accurate pride

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

beautifulli
foodattent
friendli price eat
reason visit prepar
staff wondertwice
helpstai meal
love delici

Most Negative

drink
expenspoor burger
pricei tastinyc
servic unconsidstaff
fineslack london
head
bettercheap

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