



# 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

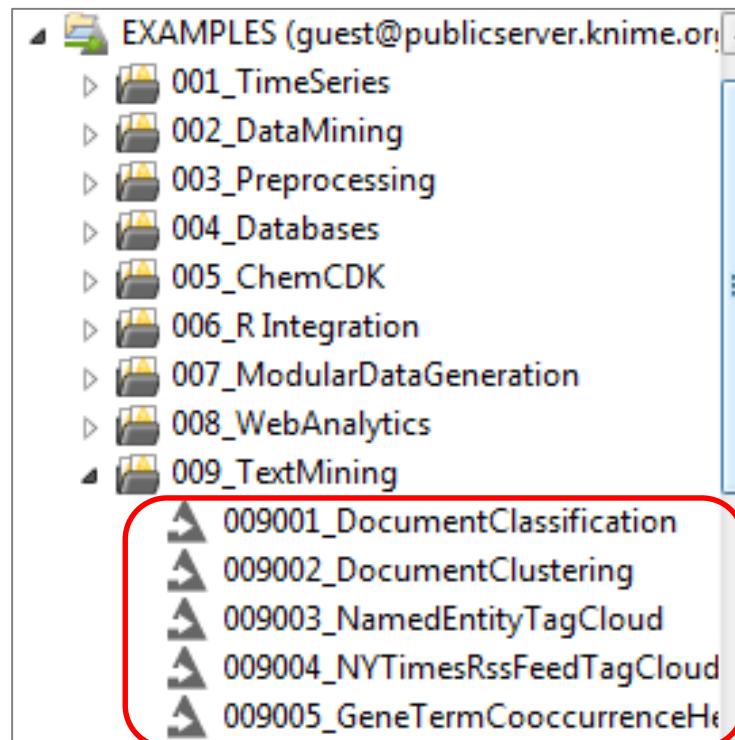
KNIME:

[www.knime.org](http://www.knime.org)

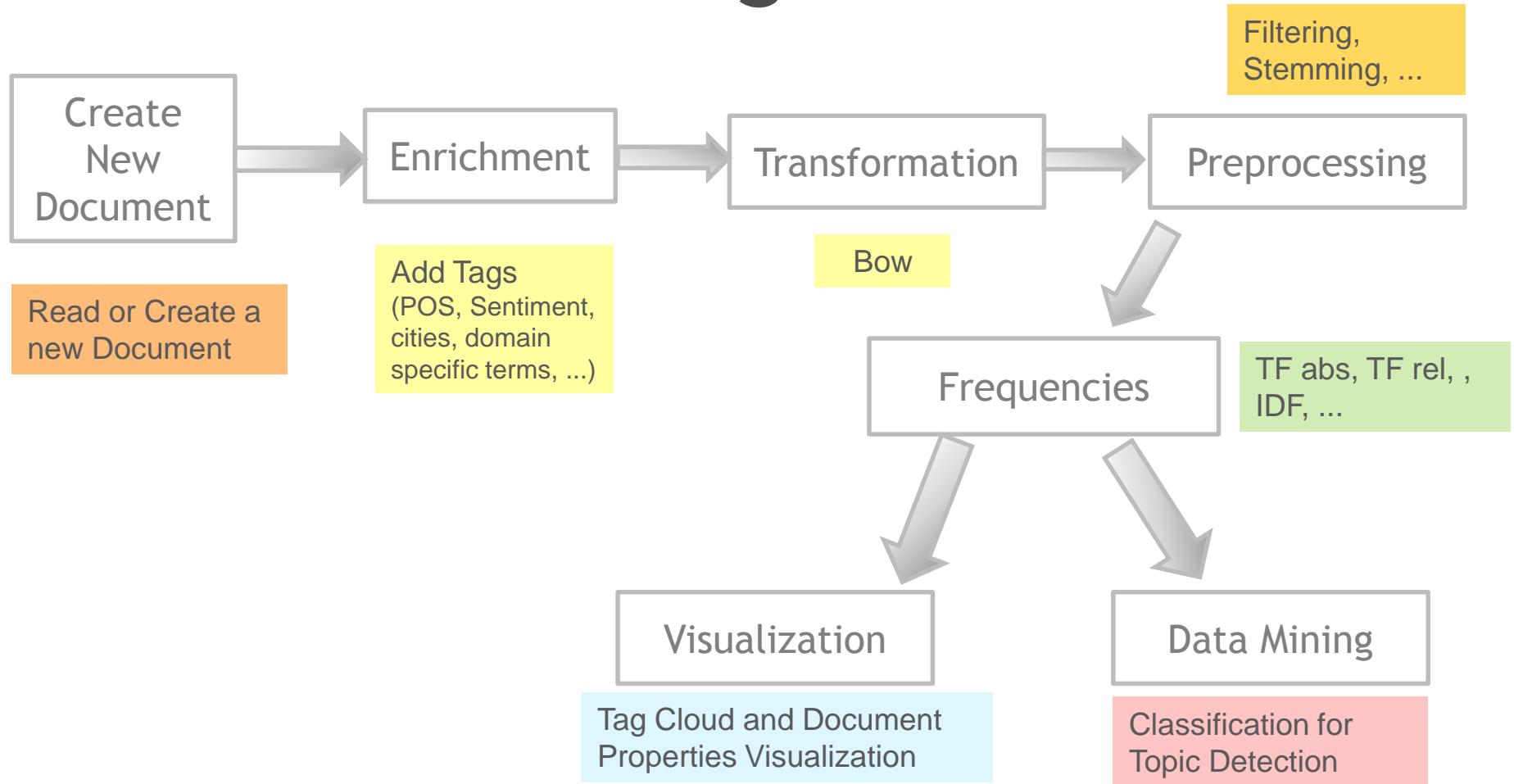
Install Textprocessing Extension  
under KNIME Labs

# Examples

Example Workflows available on the KNIME public server.



# Text Mining Workflow





# 1 - Create a Document

# New Data Types

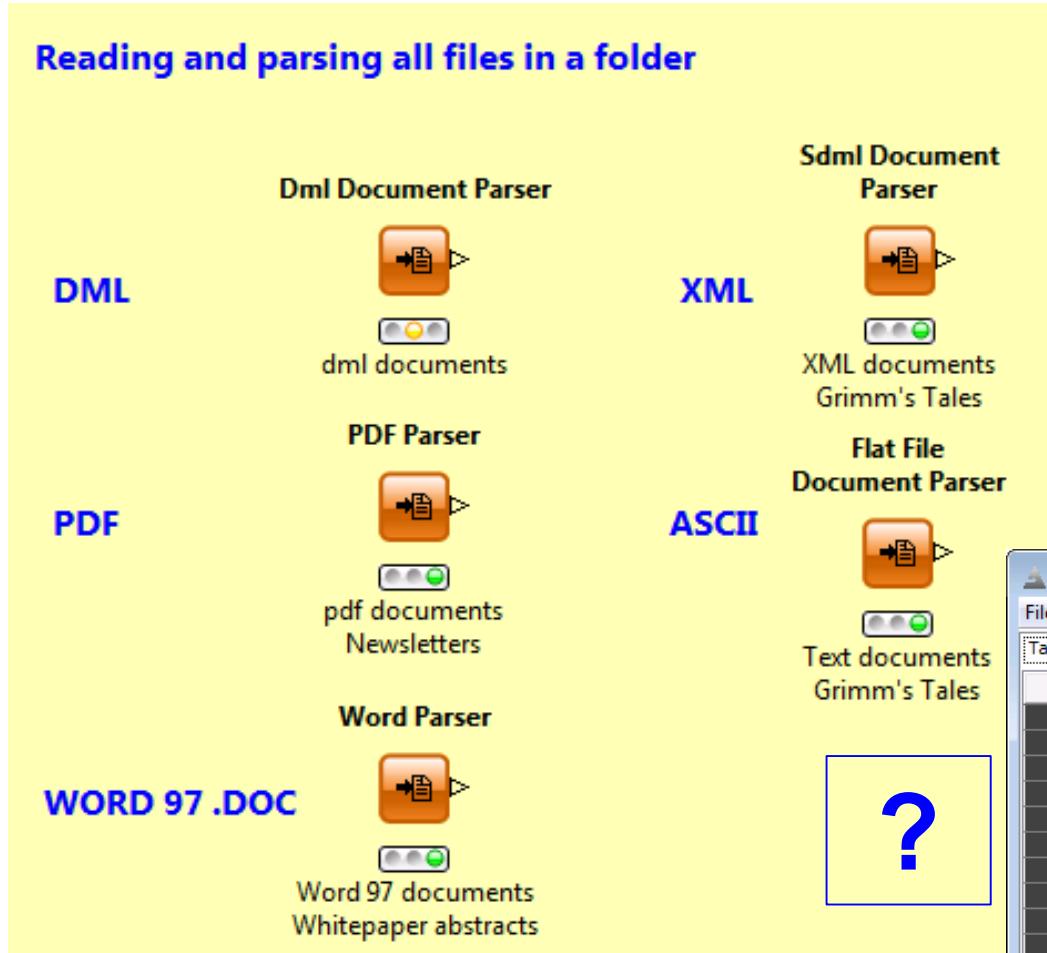
## Document

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

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

# From a Folder

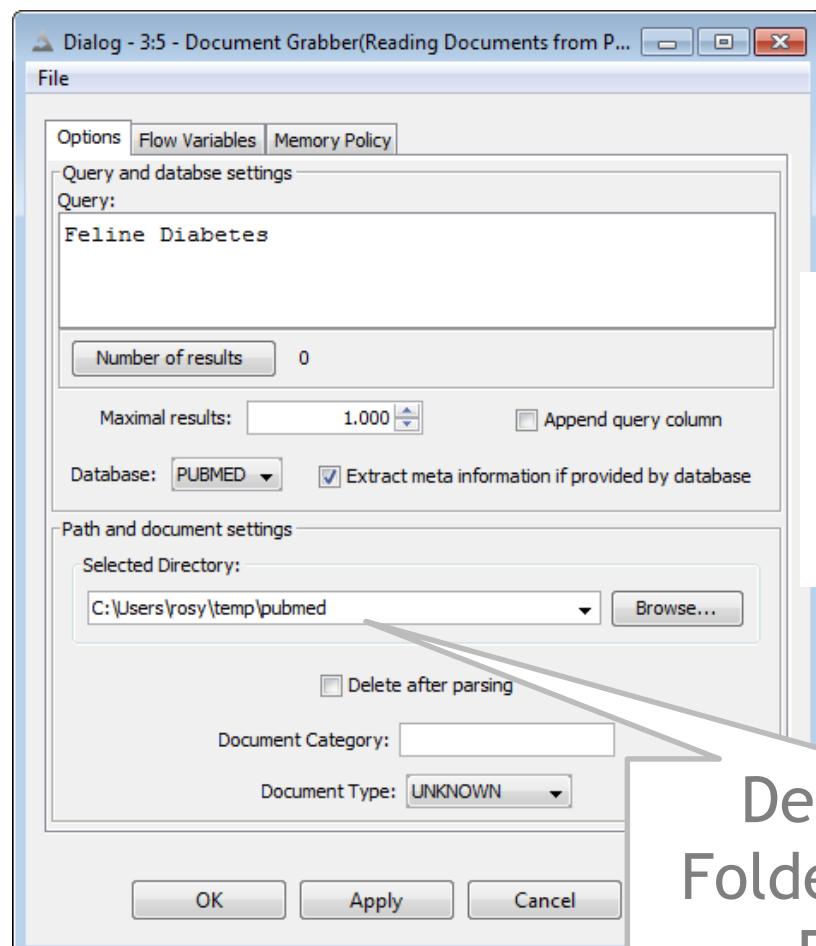
**Reading and parsing all files in a folder**



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
Row1	"A shoemaker, by no fault of his own, had become so poor that at last he h...
Row2	"Allerleirauh."
Row3	"A Riddling Tale."
Row4	"Bearskin."
Row5	"Brides On Their Trial."
Row6	"Brother and Sister."
Row7	"Brother Lustig."
Row8	"Cat and Mouse in Partnership."
Row9	"Cinderella."
Row10	"Clever Elsie."

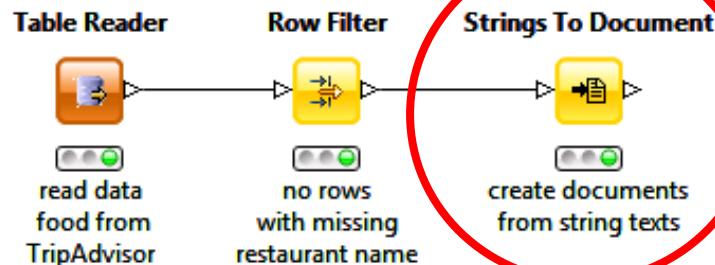
# From PUBMED



The output is a list of Documents

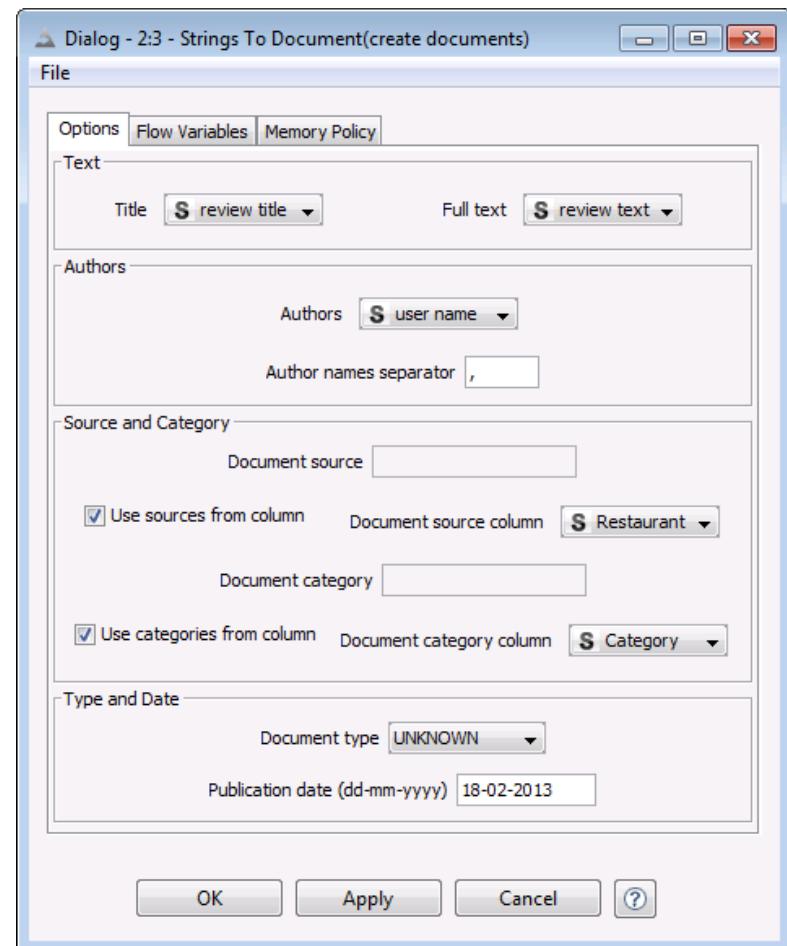
Row ID	Document
Row1	"Traumatic Digital Amputations of the Foot Inflicted by Pet Cat upon Individual..."
Row2	"Trilostane Therapy for Treatment of Spontaneous Hyperadrenocorticism in..."
Row3	"Computed tomographic signs of acromegaly in 68 diabetic cats with hypers..."
Row4	"Screening diabetic cats for hypersomatotropism: performance of an enzym..."
Row5	"Molecular Characterization and Tissue Distribution of Feline Retinol-Binding..."
Row6	"Organohalogenated contaminants in domestic cats' plasma in relation to sp..."
Row7	"The effect of experimentally induced chronic hyperglycaemia on serum and..."
Row8	"Differential expression of circulating microRNAs in diabetic and healthy lean..."
Row9	"Concurrent somatotroph and plurihormonal pituitary adenomas in a cat."
Row10	"Feline diabetes."
Row11	"New incretin hormonal therapies in humans relevant to diabetic cats."
Row12	"Oral hypoglycemics in cats with diabetes mellitus."
Row13	"Continuous glucose monitoring in small animals."
Row14	"Diabetic ketoacidosis and hyperosmolar hyperglycemic state in cats."

# Strings to Document



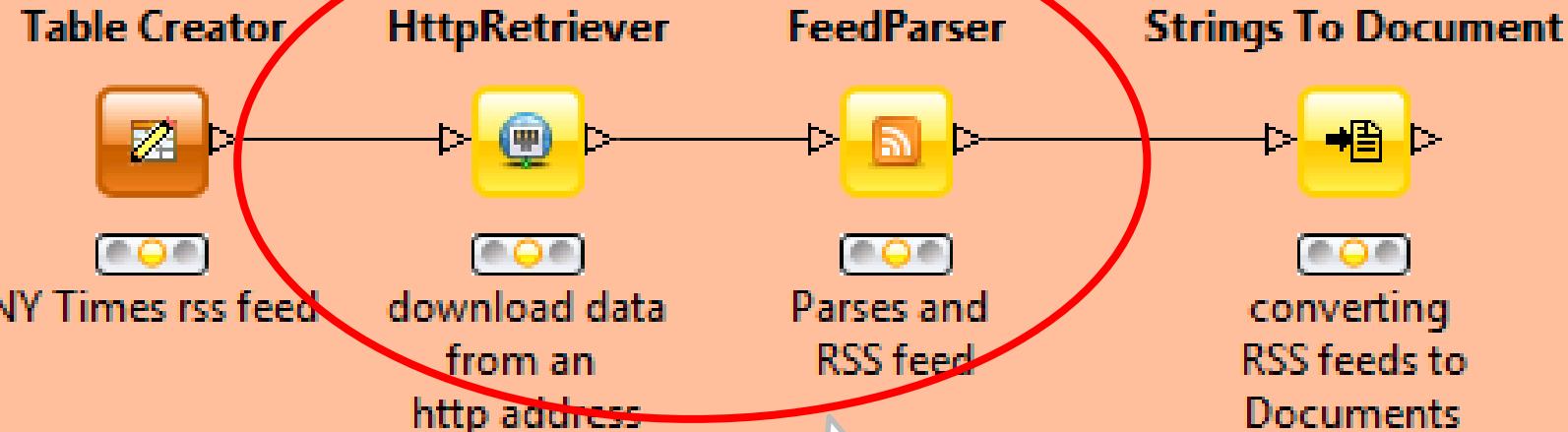
**String and document output table - 2:3 - Strings To Do...**

Row ID	Restaurant	Categor	Document
Row0_1_Row...	gon and m...	Asian	Create 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

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 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
zuckerin[NN(STTS)]
rapunzel[NE(STTS)]
frau[NN(STTS)]
königssohn[NN(STTS)]
haar[NN(STTS)]

# Enrichment (Tagging)

Enrichment nodes (mostly) **change the granularity** of terms.

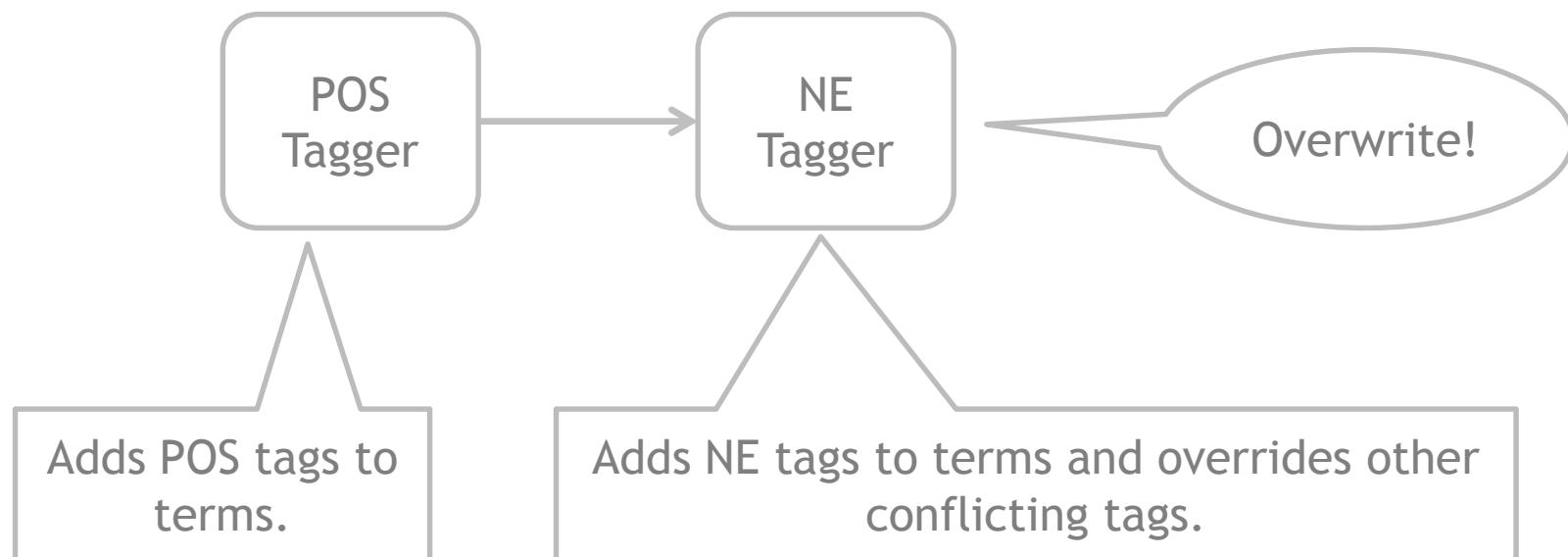
- Multiword detection, named entity recognition, part of speech definition, ...
- To each detected entity (term) a tag is added, specifying its type.
- To avoid intersection of granularity the **last node dominates**.

# Tagger Conflict Resolution

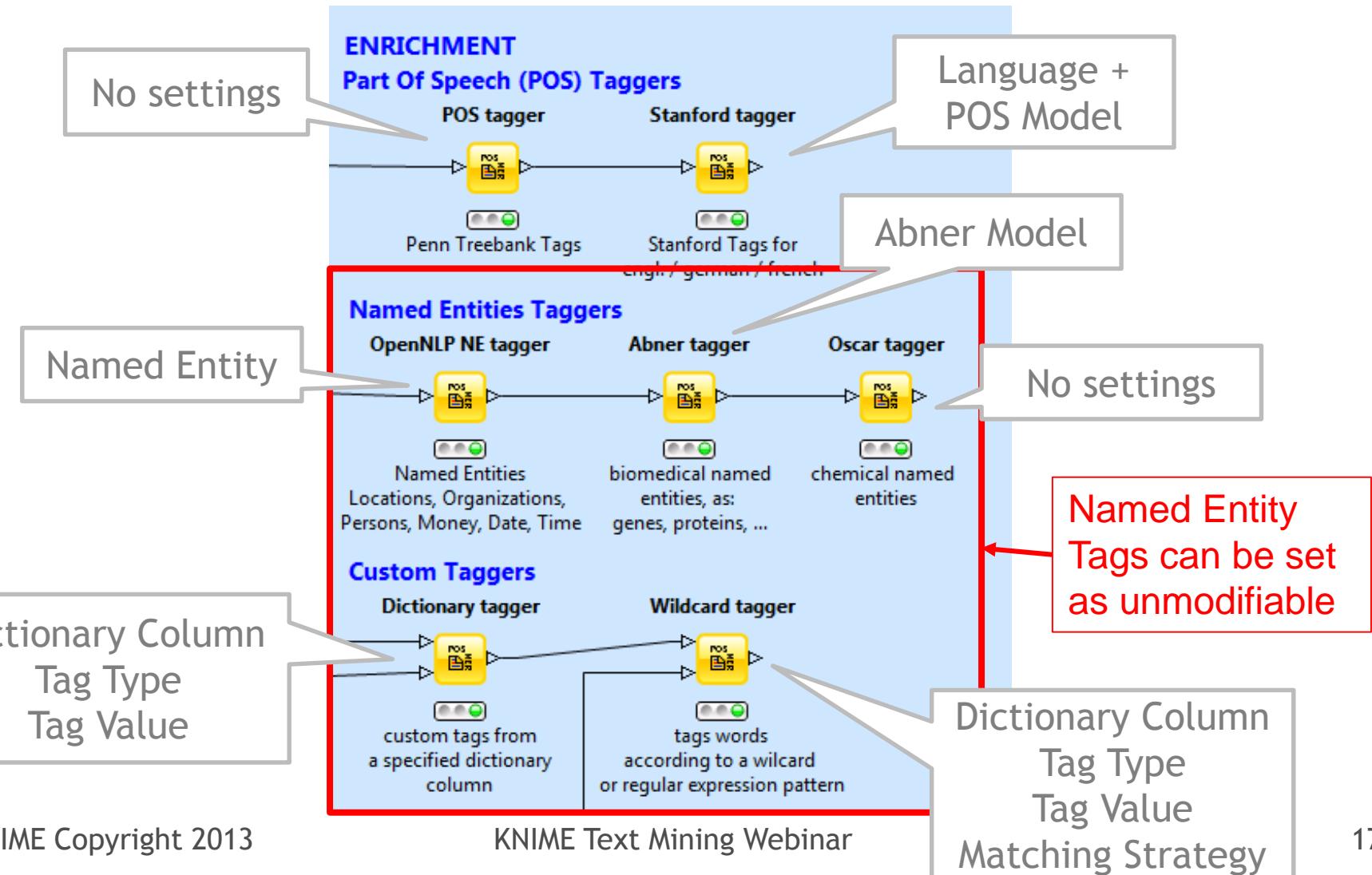
In case of intersections of granularity the last node overwrites.

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

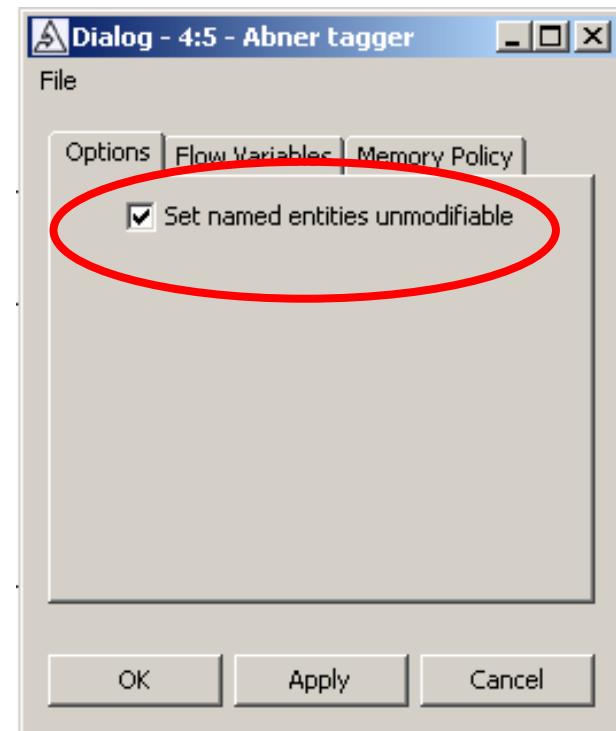


Dictionary Column  
Tag Type  
Tag Value

# Unmodifiable Named Entity Tags

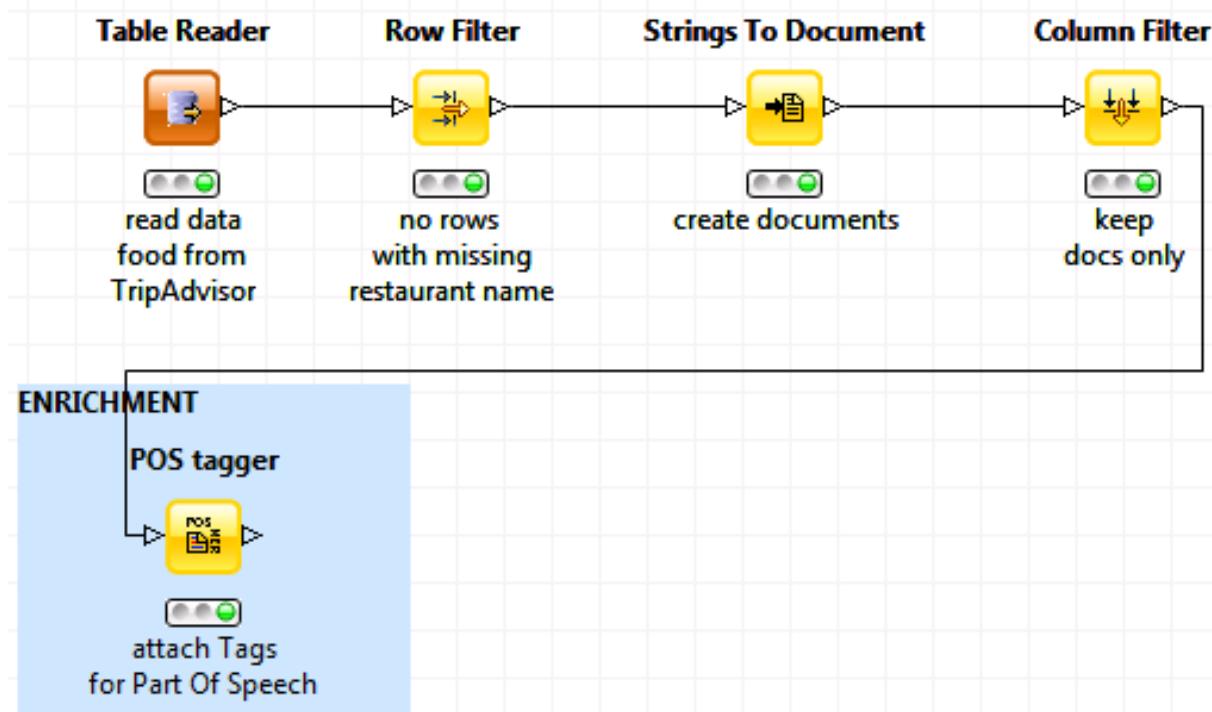
Named Entities Tags attached through enrichment nodes can be set as **unmodifiable**

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



# Workflow

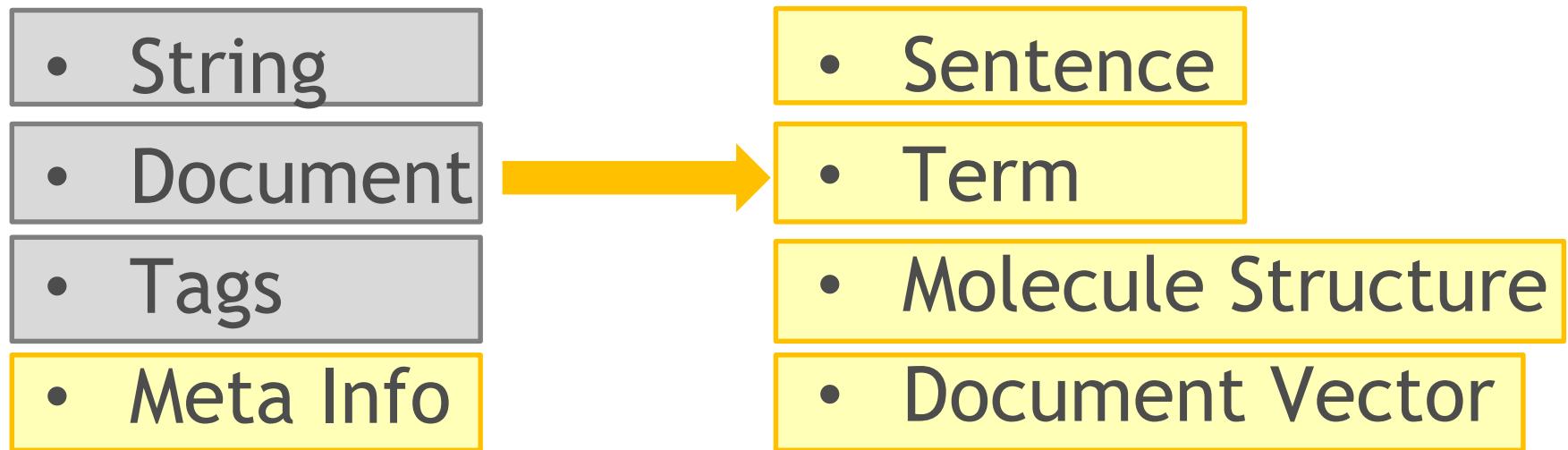
## Strings to Document Workflow + POS Tagger





# 3 - Transformation

# Data Types and Features



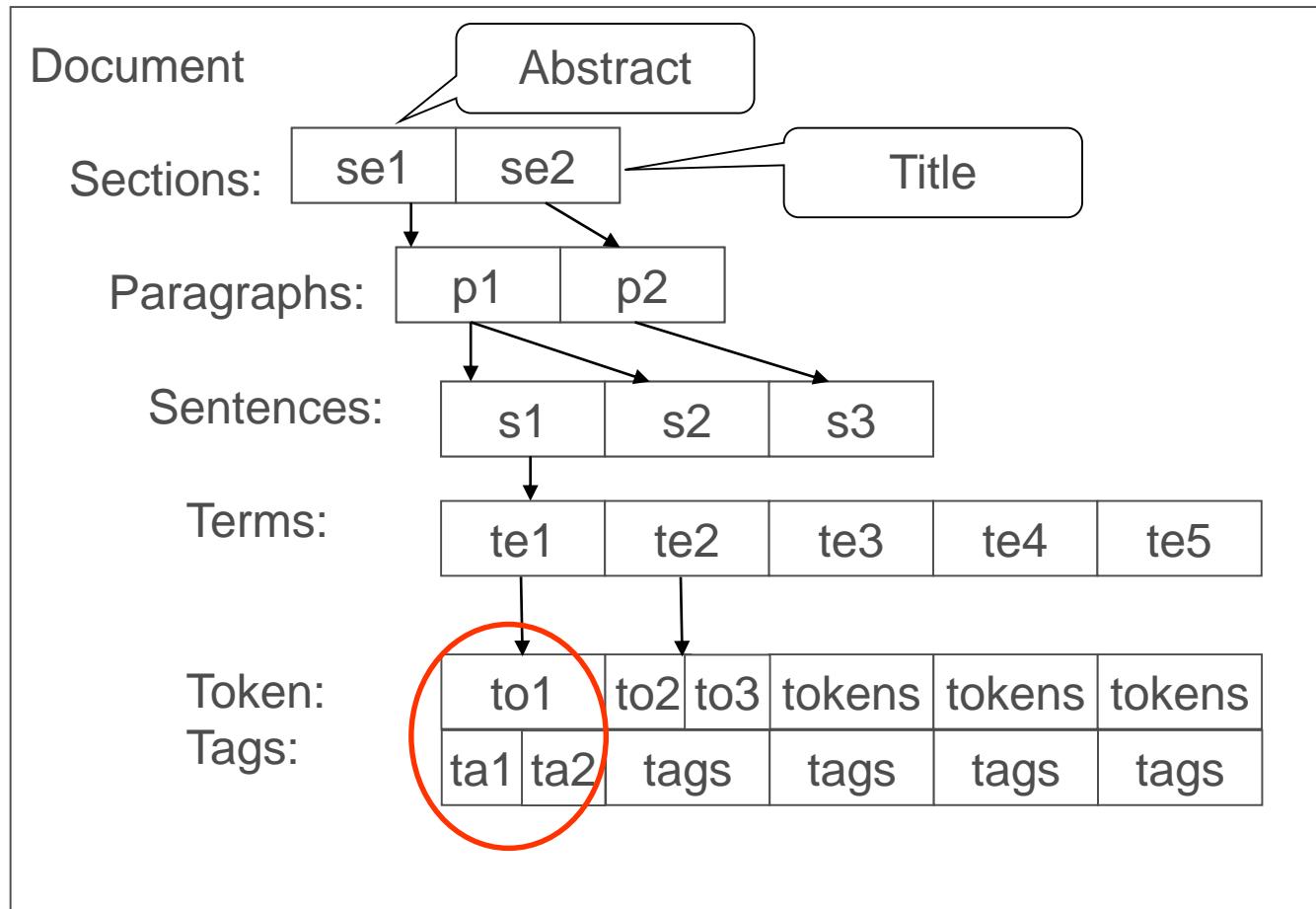
# Parsing / Tokenization

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.

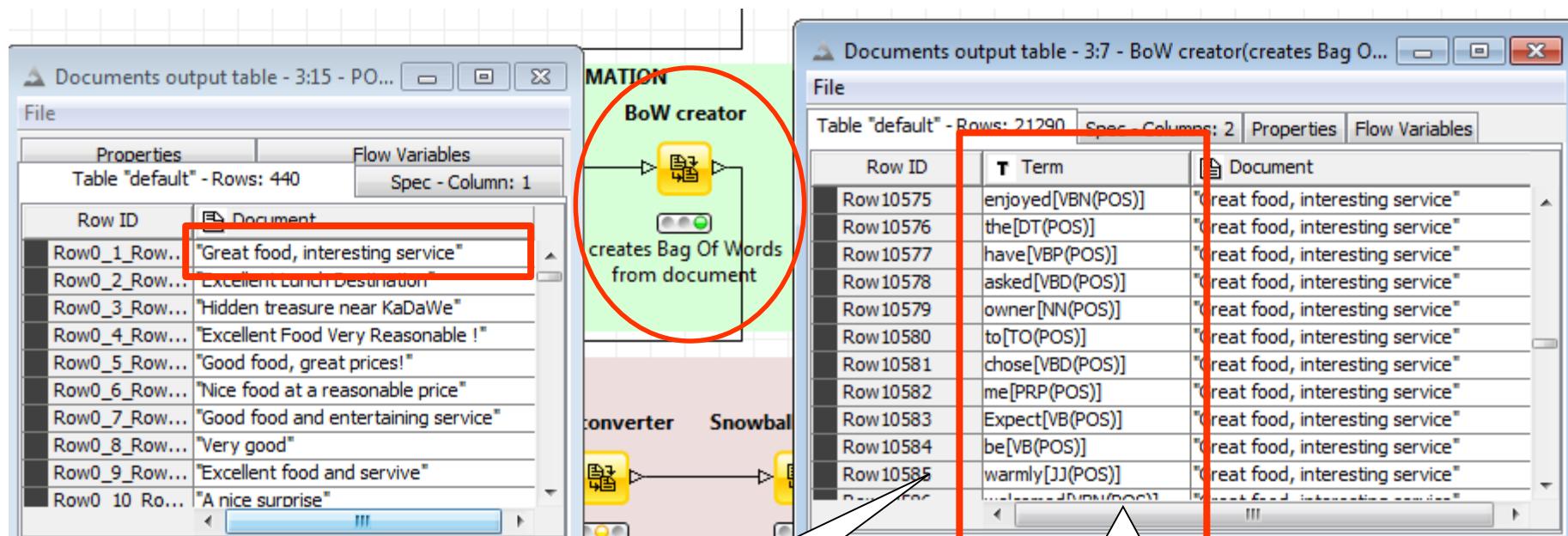
# Parsing



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

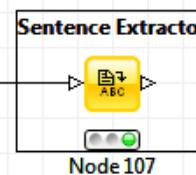
Each Term is extracted with Tags (NN, VB, ...)

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

# Data and Sentence Extractor

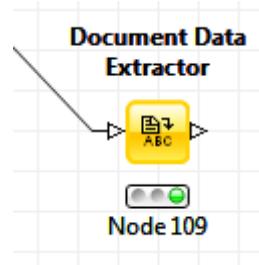
Filtered table - 3:5 - Column Filter(keep)

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

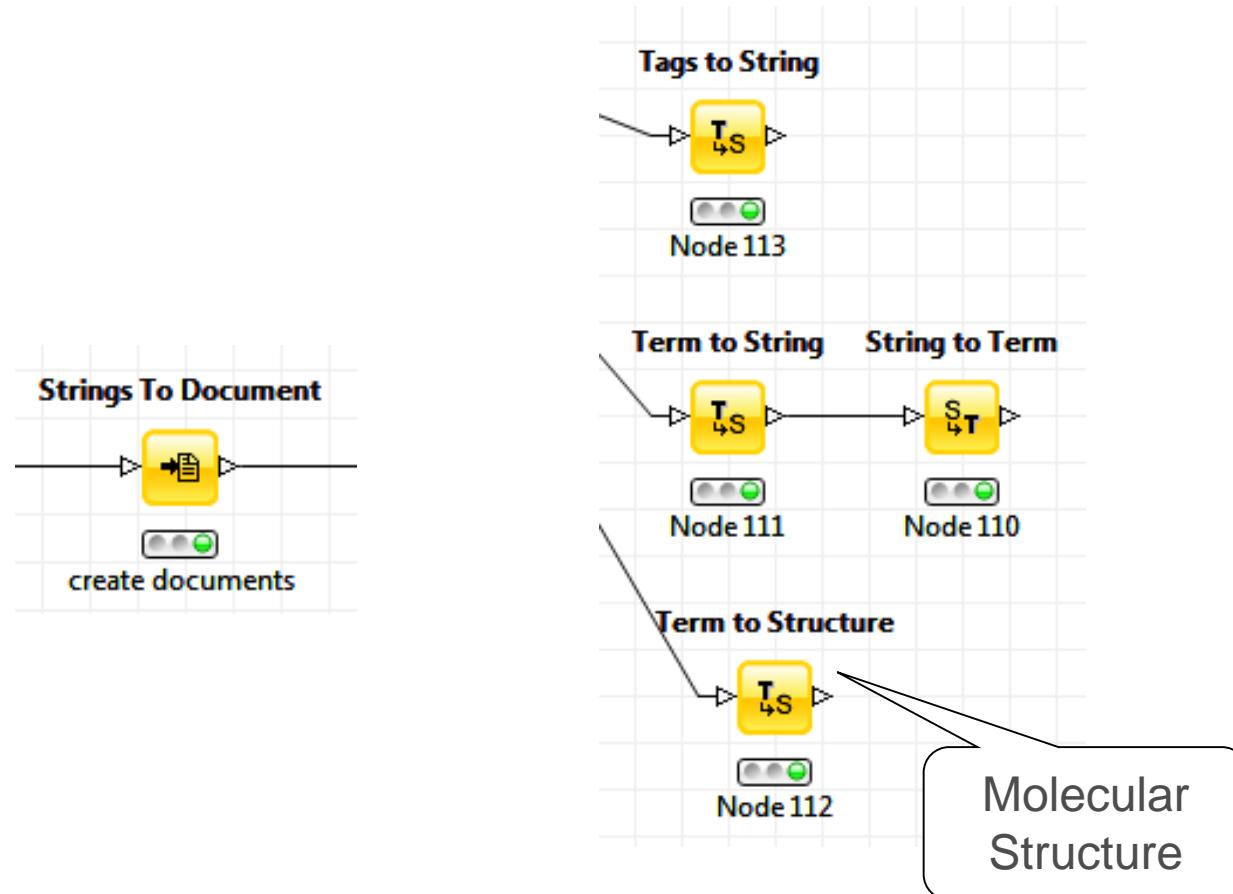
Row ID	Document	S Sentence
Row1	"Great food, interesting service"	Great food, interesting service
Row2	"Great food, interesting service"	this restaurant is seemingly known by Berliners fo...
Row3	"Great food, interesting service"	I enjoyed the food and I have asked the owner t...
Row4	"Great food, interesting 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...



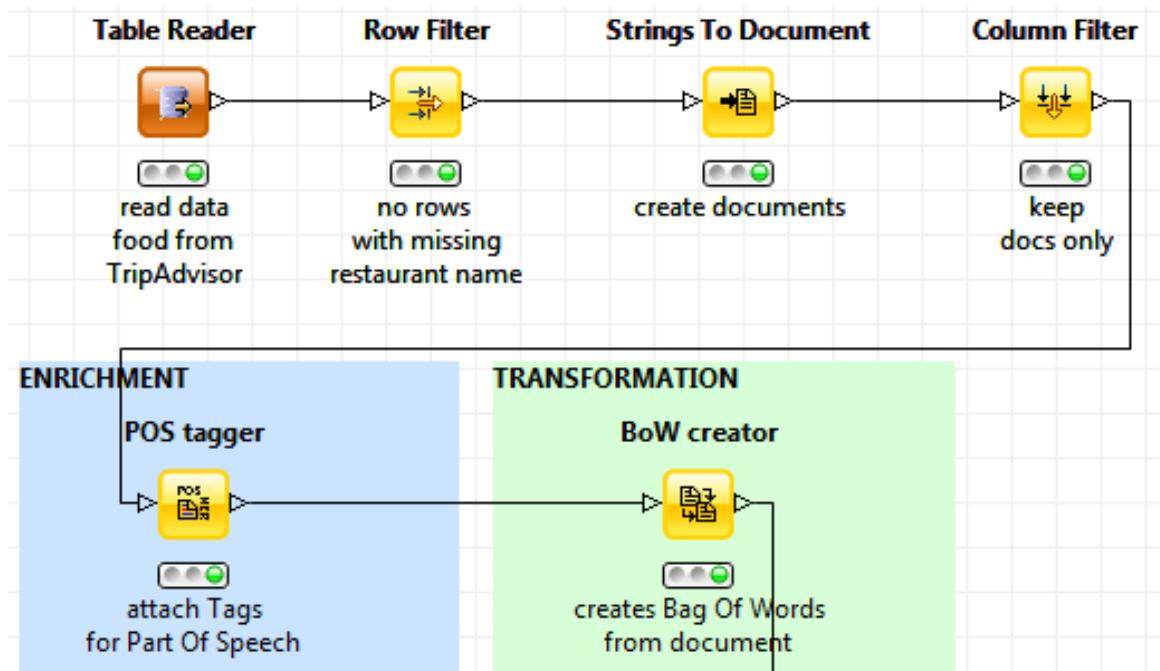
Dialog - 3:109 - Document Data Extractor

Options	Flow Variables	Memory Policy
Document column:	Document	
Data extractors:	<ul style="list-style-type: none"> <li>Title</li> <li>Abstract</li> <li>Text</li> <li>Document body text</li> <li>Author</li> <li>Author set</li> <li>Category</li> <li>Category set</li> </ul>	
<input type="button" value="OK"/> <input type="button" value="Apply"/> <input type="button" value="Cancel"/> <input type="button" value="?"/>		

# Conversions



# Workflow



# 4 - Preprocessing

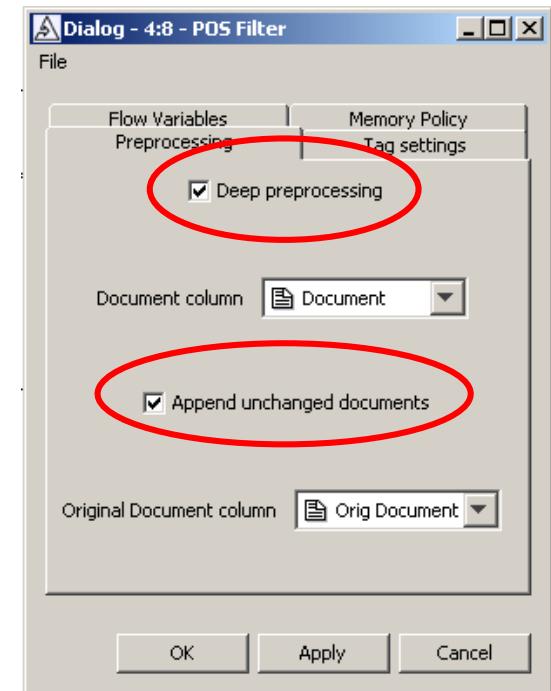
# Preprocessing Mode

## Normal

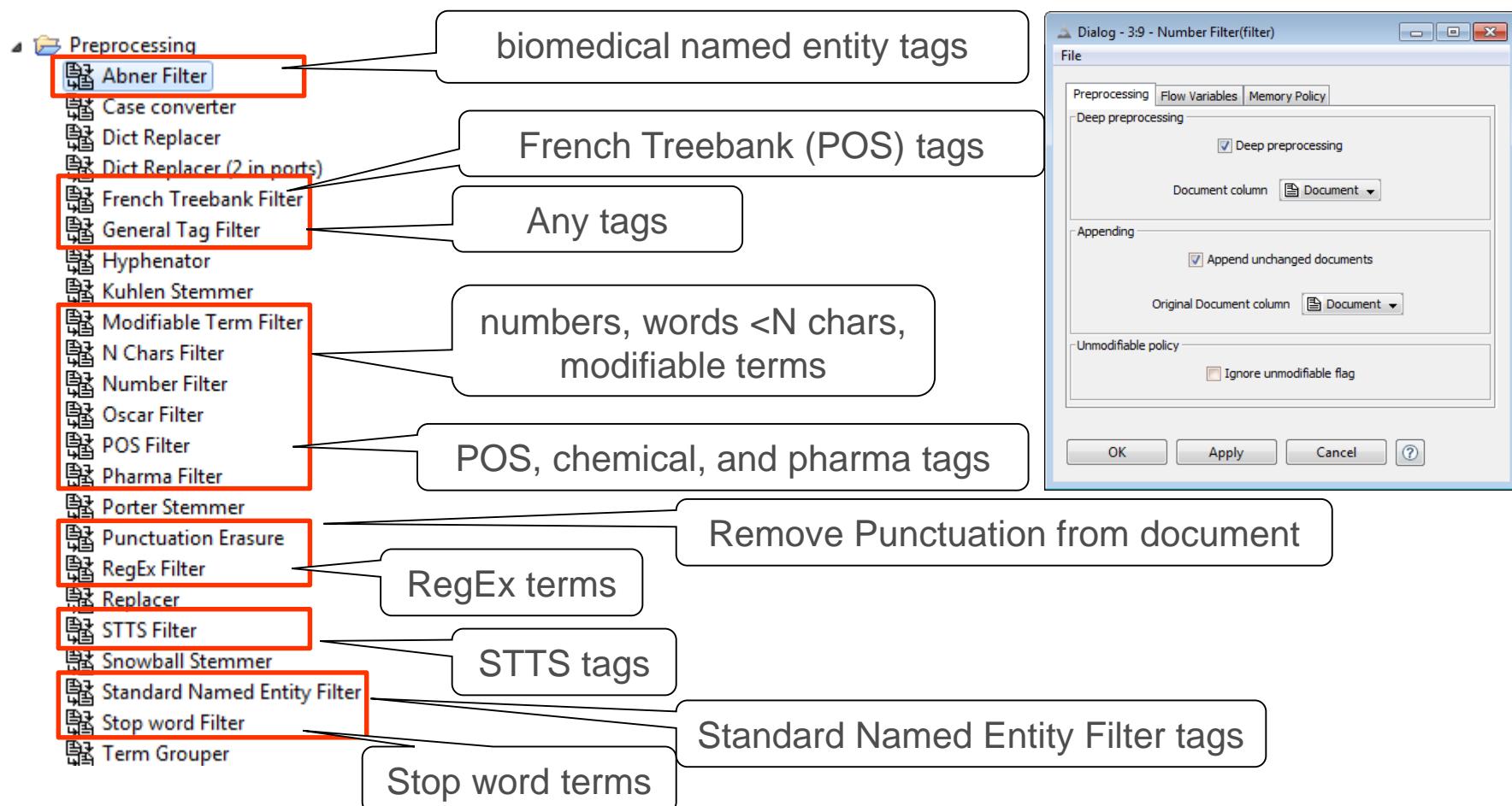
- Faster
- Preprocesses only terms of the term column
- Documents are not changed

## Deep

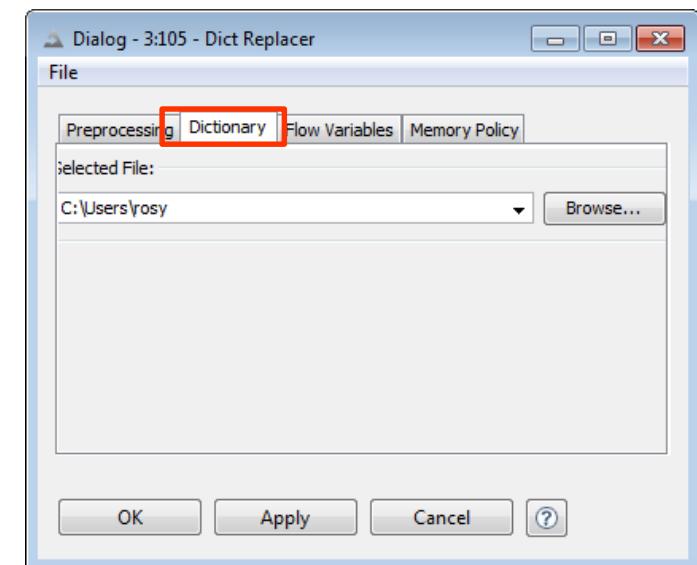
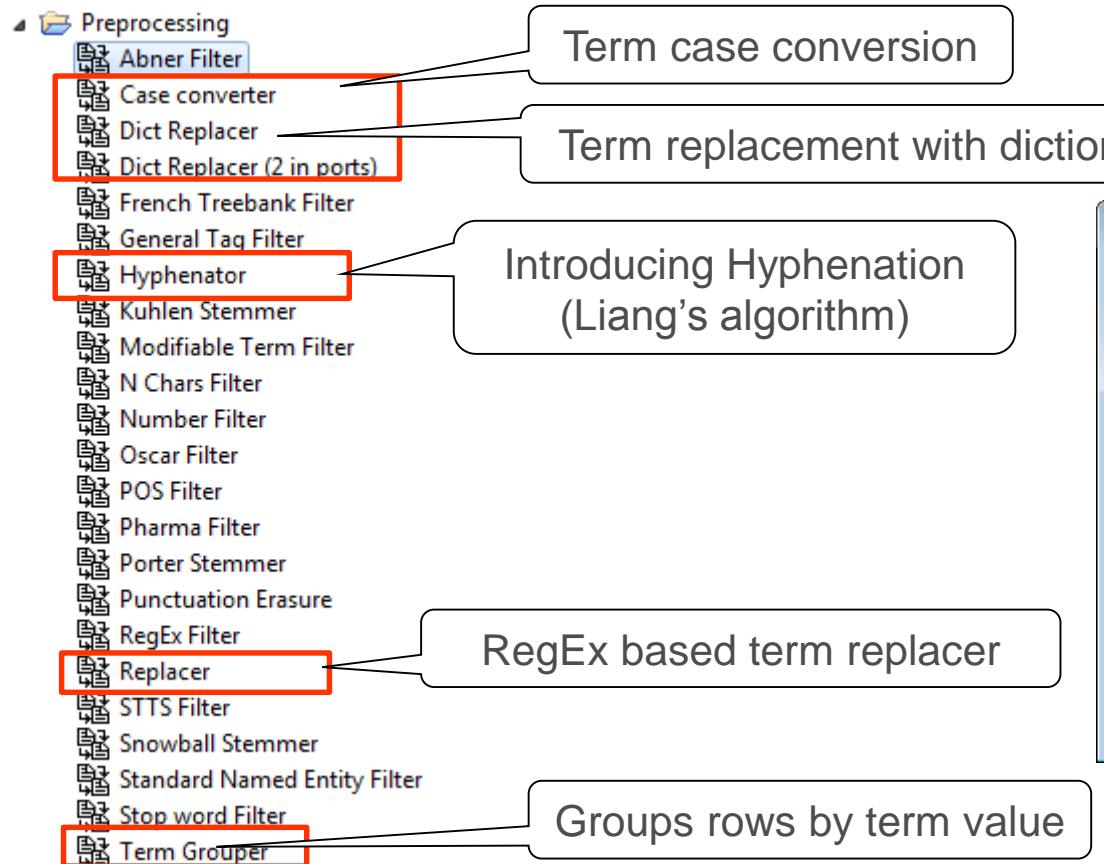
- Slower
- Preprocesses terms of the term column
- Terms in documents are changed as well
- Unchanged documents can be appended



# Filtering by Tags and Terms



# Converting and Replacing



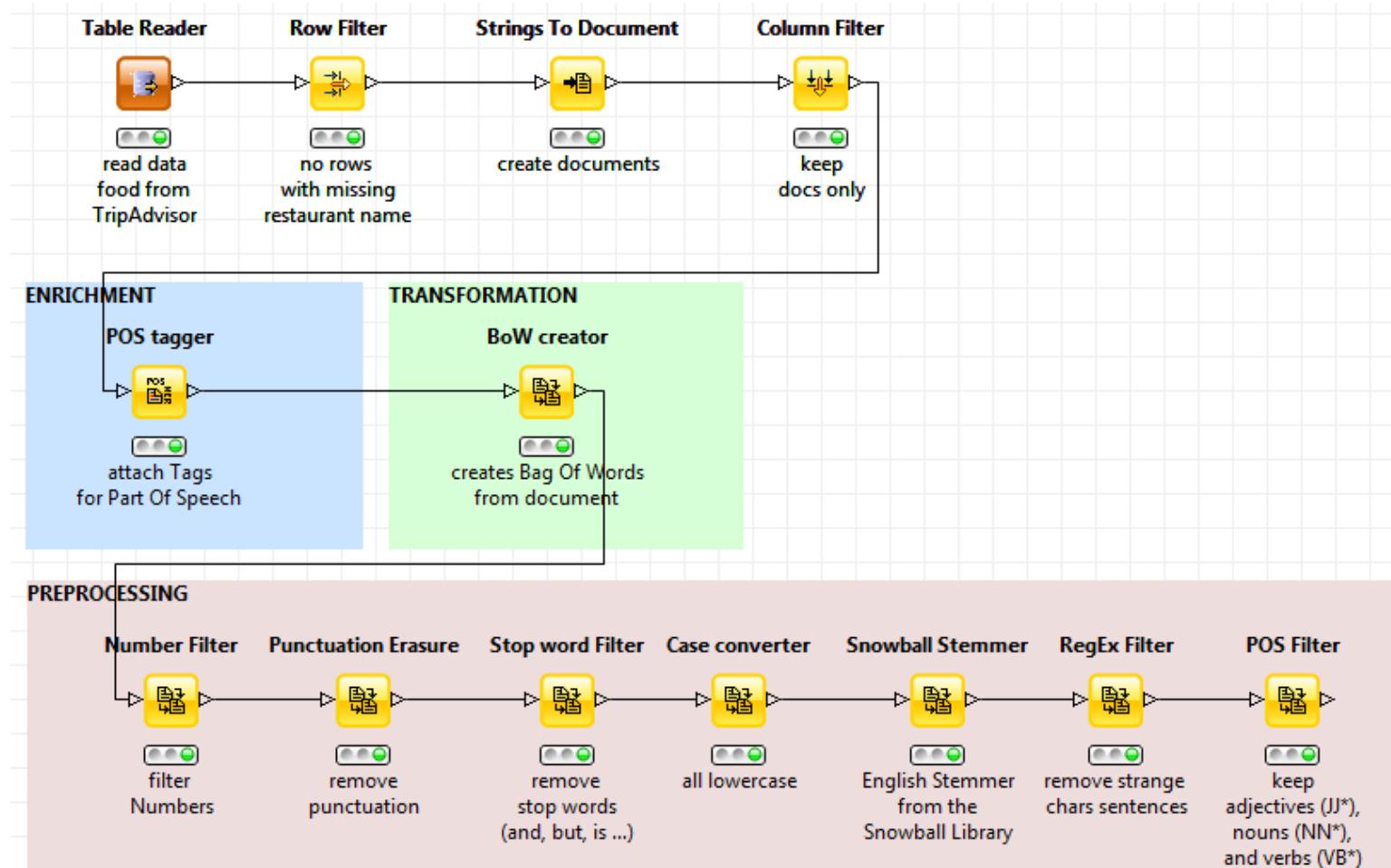
# Stemming



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

Stemmed term replaces original term!

# Workflow

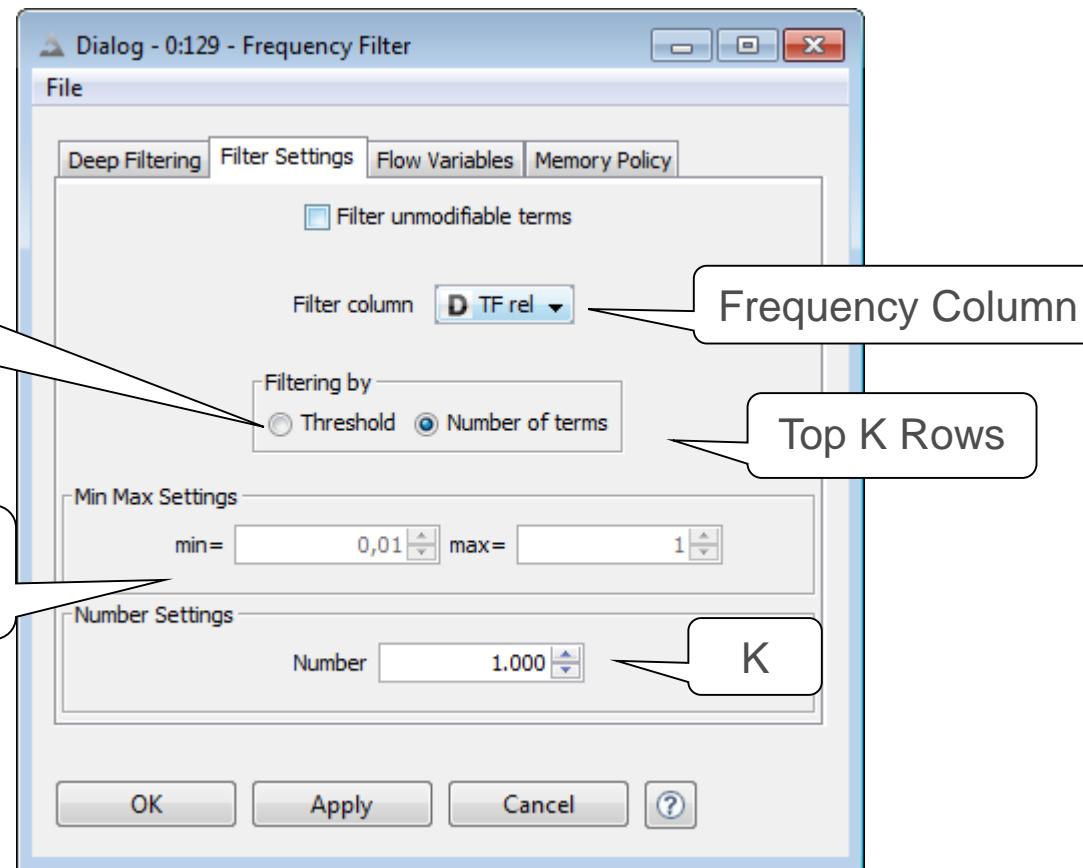


# 5 - Frequencies

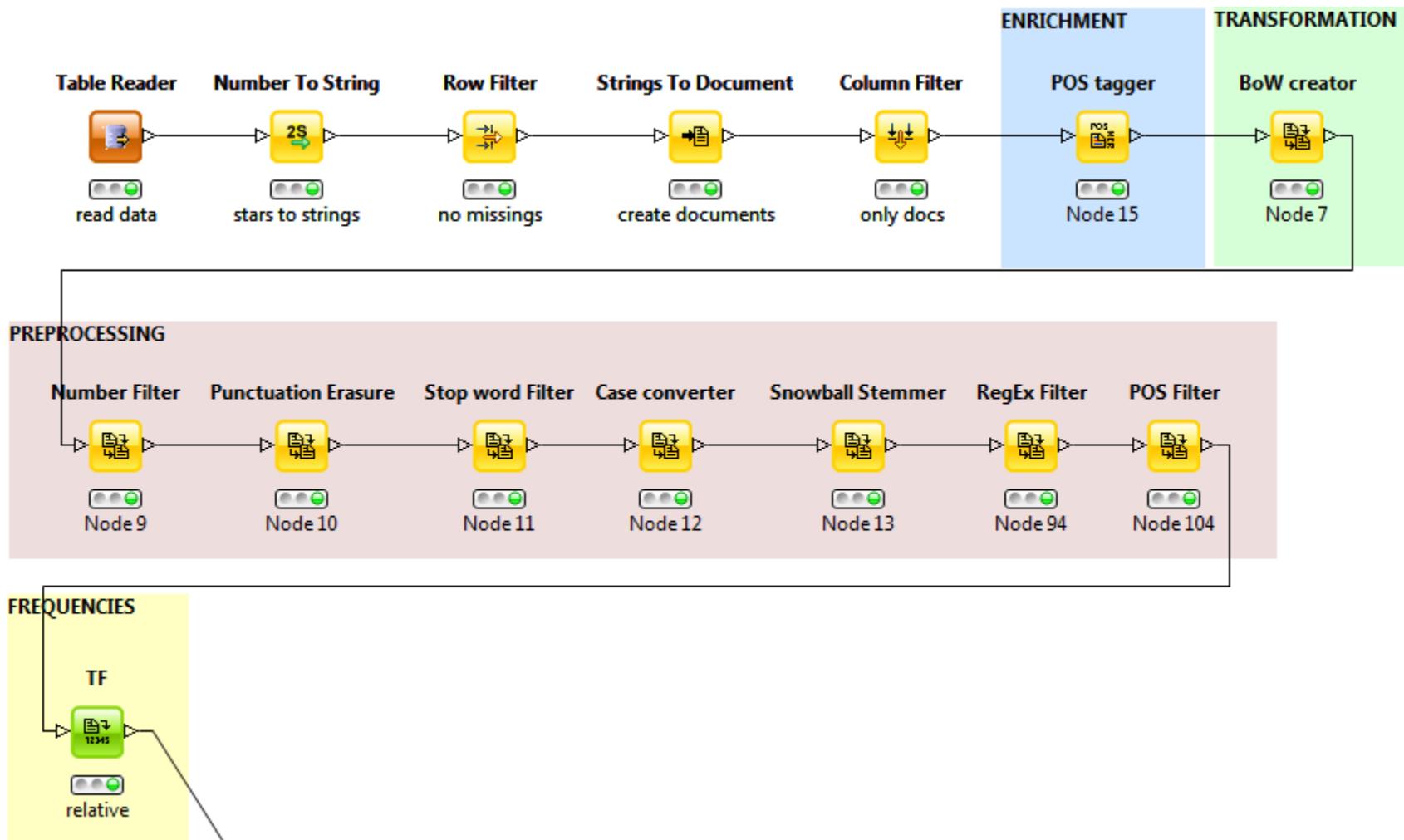
# Frequency Measures

- **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



# 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\Yosi\NoFileSpecified.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->Textprocessing->Search Engine Preferences

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

Double-click  
opens the  
document

# Document Viewer

Amazing burgers! Amazing!!!

style+

ABNER Link to: Google

**Previous and next document**

**TITLE**  
Amazing burgers! Amazing!!!

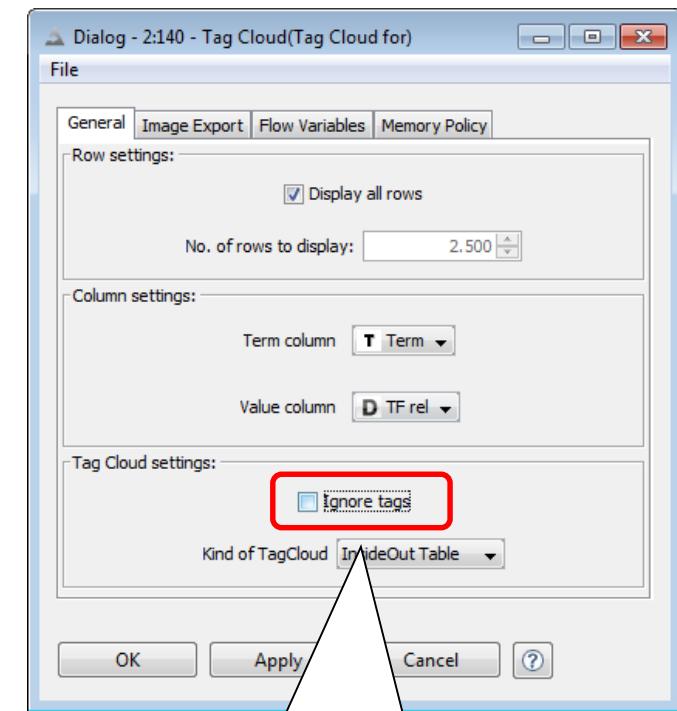
**UNKNOWN**  
If you like burgers you have to visit here! Very much USA style+ The food is stunning, great staff and all great value. I would highly recommend booking as it was packed at 10pm on a Monday evening. Seriously yummy!

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

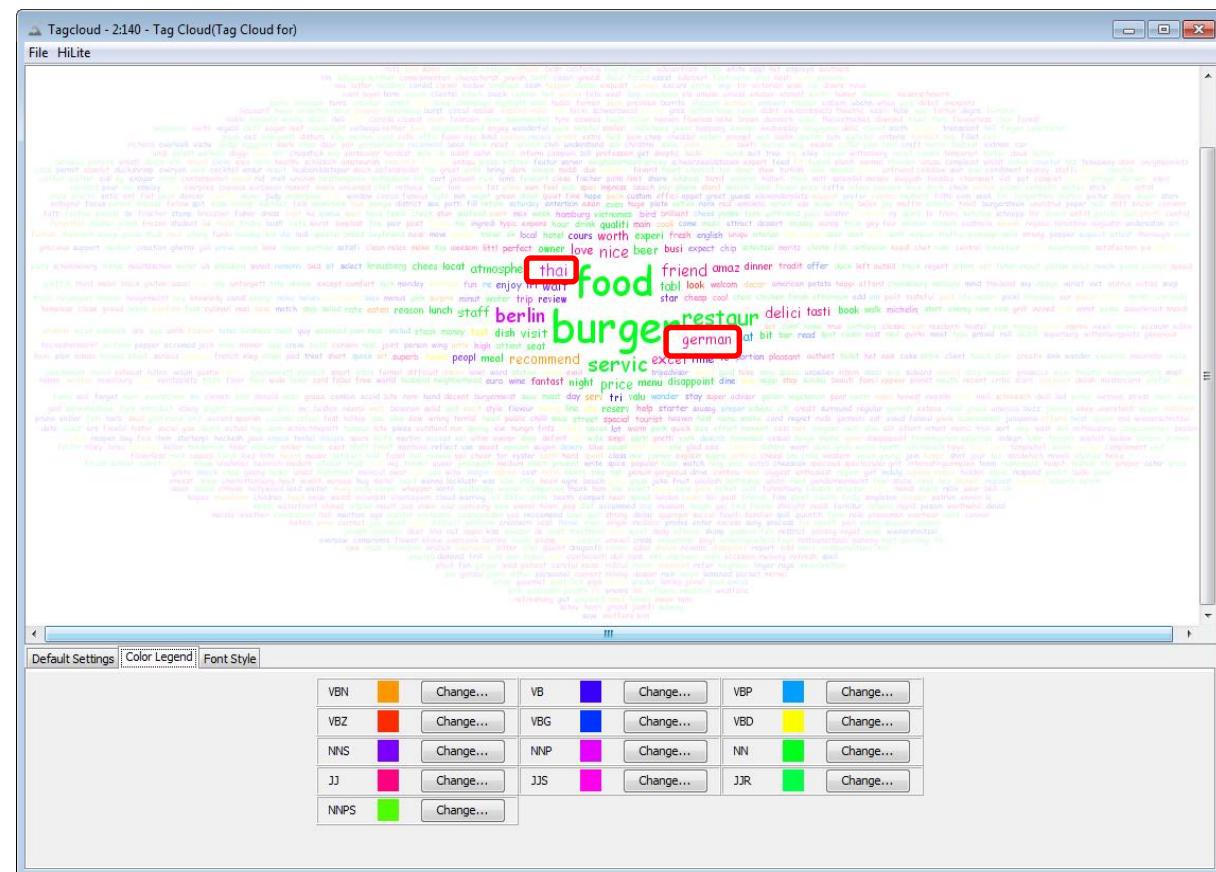
**Meta info**

Name	Value

# Tag Cloud



Adj, verbs and nouns  
as same word



# Asian Restaurants

nokki  
 pepperric tourist liqueurif  
 plum weather opinion top garlic  
 pork bland orang bustl one surround cant  
 richard chat pasion card combin arriv mint stag  
 tri engag person categori cousin candlelight chat appropri merit smile  
 ondi fan price clientel criteria care cut busi centr appmekong smack  
 base huge recogn close eleg check dear late companion block cafe apart leav practic  
 chop impress remind comfort engag chespstick enjoy low trick interior come batti bring alexand lodi permit  
 coconut inoffens restaur compar enjoy claim ensur miss unbelieve indegr treat impress close bohn bowl advis jovial park  
 cruis interv salad credit euro close eye neither unfriend intent month includ timid hit classi yummii basi woman issu overlook  
 decent issu select difficult expect cocktail fare own use jewel uk look veget girl terribl great circul worth averag type hype offer  
 dish justifi set dip extrea copious fellow par waitress monday mix min welcome leav fast friendl suspect glutam christma welcom appreci trueli honey middl  
 duck mediocr sincen discov follow dine globe pick waitress mourinbo main occas book plate januaris host mix follow surround fit choic version aircondit trip honest latter  
 flavorless plain spent dough quest do gracious popular walk origin dish reason energi ate lack left fantast famous drink fawti start fair check trip ading tri hilari langaug  
 froam previous squash duck haphazard enthusiast pretti warn peanut europ spici manag believ home tri season germani impress eat sunday expert servis facad check treasur accomod travel goodth interst  
 grate re standard eager kadem equip guess recommend wrong potato favorit top bear eve found cook alexanderplatz thailand read cook expect crispi sit even seat expert cashier thavietnames wost sydney ginger genuin  
 isaan restauranfand starter edg kind exhaust hockeshermarkt result awsom prepar gem visit hous lock match waiter ve atteni vietnames curri live shortl appear class prompt die seach experi califonia skeptic vibranc sugar frill fragrant  
 kapao slow stick eggplant kokossoup torged hilfot roman blanc rememe green help pack environ fanci perfect decor cours enjoy fantast welcom residenti pad monday week charact match cross salut euro bunal serf travel spot flower feel  
 kitchen squash surpris email love fork histori saturday begin reput includ roll dinc accid offer tri offer atttent portion walk disappoint visit train definic decid walk sunday decid afternoon colder reserv drink blend saturday touch smell finish exemplari  
 annon takeaway entr mark fresh hurri stamp brown reserv krapau meal fanci book water vietnames wine saigon wait tasti fresh host starter amaz seafood dine friend thai regular joy distanc basil relat didnt berlin rind told simpl fill establish  
 love berlin tea everyon miss garnish issu ta coladcola seek left stun phone finish stay read happy wife look love visit valu bit price ate interior live surpris disappoint surpris help franc beer address reflect denmark bathroom research templehof serv  
 tell noodl gleet line take cool select main enthusiasm visit mail german vegetarian includ come eaten reserv menu delici excel staff worth asian custom chang bathroom effort rate rank vitor fabul fri wow prosecco delic assist relax tabl salti  
 look thank critc smile recommend tripadvisor manag mail recommend minut sauc lot english dish qualiti excel **berlin** **thai** friend time price nice parti gem flavour sit wonder noodl told australia pom worri pretti comment arriv relax supris  
 make cater turkish exager overpr guid marker underst die surpris share juli ton mirell simpl red **Food** **food** **restaur owner** visit eat serv love german line due even octob faboul sampl menus trust plate brand  
 hashess meet uniq duck veggi wait recommend friend stumble station conveni averag rate special trip restaur price **niche** atmospher book eat soup day move phone unbeliev guy explain partner intrus total pay book agre pilzen spoil  
 extra overpres heat moder uniq english verg wonder starter left tri friend entertain spice delici tabl servio dish  
 papaya complimentari wise fail quirki inconsist motel world enjoy wine cook stay advisor tom car **review** vietnames **spice** **dark** fast am fast charact decid district generous team okay believ advic philosophi **spice**  
 insan museum arrin except qualiti speak street ka worst present cuisin neighborhood street warn vietnames right meal **servic** **tasti** reason main ingredi entertain starter seat son regret super charm birthday drive spice offend beauti adrianc paid  
 creat yummii fashion reach interact note arriv expect enjoy bill pack posit base menu rice return worth tri amaz dinner tabl recommend littl busi cours appet effici hard est standard charm tabl daili spanish mango basic abund mediocr select request  
 past eat accomod flat read joint notic artifact flavor fine busi fast get expect review chose expect **expens** found money experi peopl tasti visit helf sat day asia idea serv pay coriand solid make authent yam lucki see reliv explain enthusiast  
 phad fortun oppi flavor round knife ouir call forc indi tai fruit look outsid welcom quick minut chicken high duck curri cheap staff lunch joke person guess space hungry attract shi lustr attest worri lover say prefer earlier employe  
 pla fruit arriv focus server list pleas closer funni level hand humour welcom pad door superb fun time authent massag awesom tripadvisor wait start custom servic guess waitress remark light asia wittenberg kadaw satisi poor destin drive  
 prik heavi bloodi forgotten social liter pleasur cloud german locat take add knowlegd lose outstand cosi kreuzberg manag locat fabul disappoint spring coupl hand flavour like receiv join appear wait hidden salt paneng creation disservic  
 refreshung hype boy fusion starterpl mar proceed cold glad moni western clean pay tripadvisor girlfriend servid hot track yummii green repres crave wittenbergplatz gal pineappl head weekend viet handi role panang coffe describ  
 region insid brillant herb station miner quaint companion glutam prais walk corner proud flair entre eat soups rest owner advisor glass word pepper greet vietnam usual greet repli offer cifi cool  
 river local chianfi hour style mkt real conceing greet site min husband unlik nearbi live beef bit inflm arriv track pace gorgeous typic underestim finish regular occasion chines chanc  
 satay lot chicken kinda ta novemb return confirm healthi sauc cheer serv waiter fast attent quiet spring stay monday fuss truth thirsti explain railway live charlottenburg car  
 save orderd classi leisur various pack riesd de help smart disappoint recent disappoint left simpl prepar mobil fusion travel fast euro pronis krubuk chap bergmannkitz  
 som polit com like veggi passion run dim kitchen unanim follow sit servic disagre mift feel tradit fast euro produc king carrot beerbk  
 sweet recept comfort local wonder plate sampl dragonfl kreuzber sidewalk main add lie expect suggest top discoveri platz khiew capabl beaten  
 tam satisi complementari manag zoo plentl shake dream lodt accept level excess stay suit disappoint pick kha blond add  
 tripadvisor sunni continu manner advisor pop simlar dull leav email start spoon deserv personnel juic bill twice  
 absenc terra cozi meat awesom restaura situat edibl spoil sophist deserf perfect itali bike tree  
 omblanc tourist deco memor bad run spill serv current perfect gaeng beetroot select  
 center town decor menus bottl sat close pan forgot antic recommend  
 charg vegabli duckshrimp met careful opt flexibi week prize  
 doubl vibe ent met firm walkabl mongo  
 half viet ertugrul visiti litr

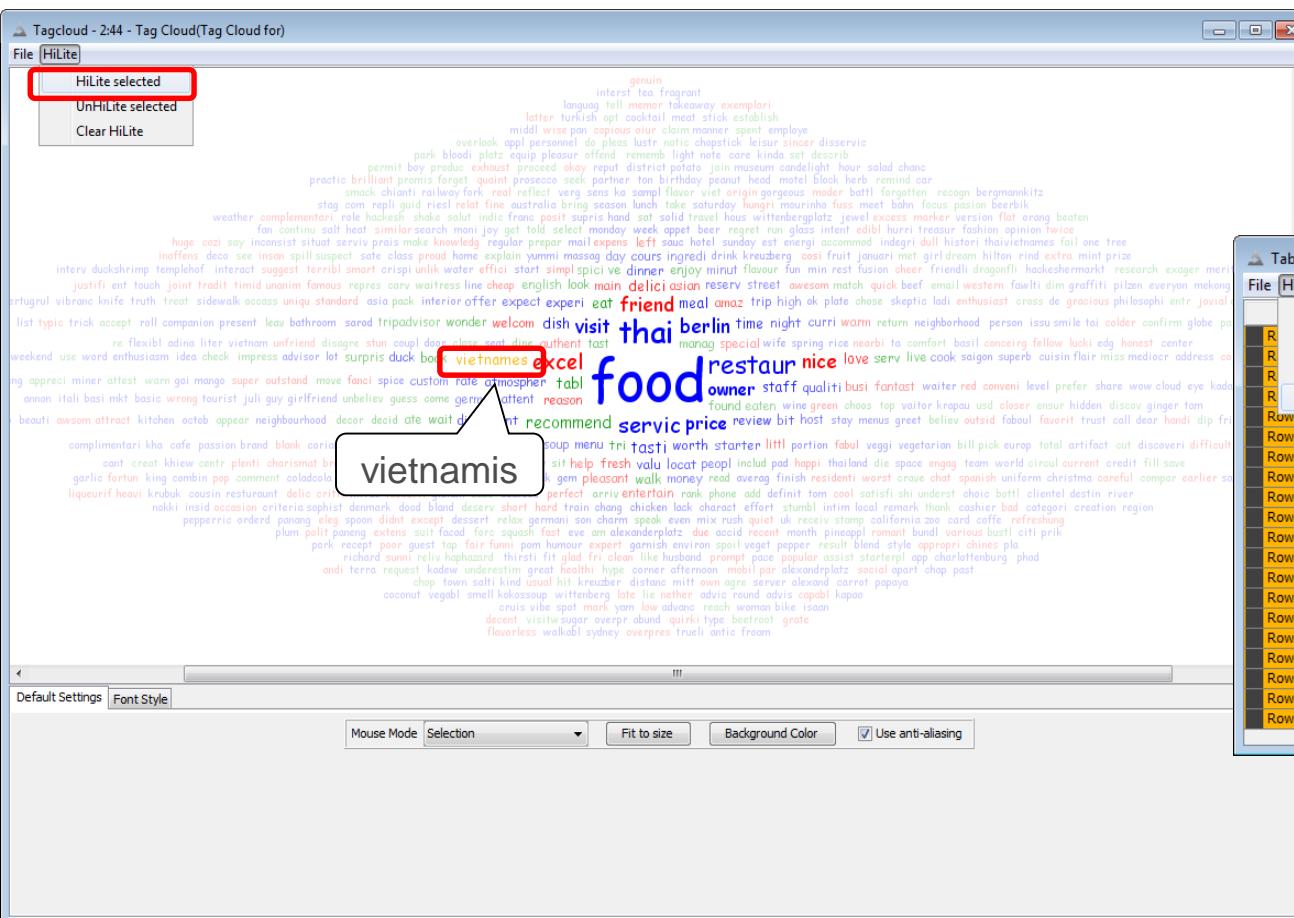
# German Food Restaurants

tab  
 tip plush spell  
 understand red email place sauerbraten  
 wienershitzel regal employ rang dough pineappl marvel  
 charg remark explain reachi furnitur rain doggi pile excus  
 connor shrimp foie refer hospit irrit form quirki doggi net westfalia  
 la stig followid region intrus jean indule introduct fix premis disturb nasti subway  
 notch slip foodi research januari modern interestingcomplex treatment hand interv feb pipe deft lost station  
 panel solid friday restaurantpub jour oberbayern mango unavail am reflect fun glass fay patient dedic imagin pride  
 unit stodgi get restrict jump orchestra profan vacat app moll witw oil fowl frischer dont overheard cracker ice owe  
 water strip haut rotkohl lack perform provid king polit trendi candi gem veg lentil fleisch express dissapointi overcook couch humor nation  
 zealand supermarket hood say map plump realis present head ultim favorit eel travel dance till futur film european dissapoint ny chang gulp irresist  
 artist synthet intoler singleton mid quintecenti share reput boil door opinion decid oyster feel classic regret terribi fabol fennel equal desper middi catalogu flavourless famous  
 aspic tendenc lard son notic shown snap centr guy rate bottl averag address authent bavarian opt amus use stuff entir familiar eay dewdx mention bring finger dice  
 bean term lit starsit patato song squeal bust run mind intern refin bouch sunday special music list formal crowd wall schlachteplatt darm extra dollar continu mault believ fillet date  
 blend toilet lmao stove pictur straight surround match spacious rude overpr eve comfort arriv name potato select credit fill terrif fine wiener philosophi berlini excit cool check mosa band fanci centuri  
 butter underton lol tax pot stube tarantino salt grub cold gorgeous take genuin profession stay set busi review guid card ve unforgott cart front pain suggest entranc champagn boyfriend low bag extrem caus  
 illa london tender pup teen think standard flammkuchen italian outsid ectect read sauc attent uniqui tri night max locat class citi advisor pricye chose choos noisi black embarrass btw bitter levelit australian exempl baco  
 ett mine usual relief tipyc thr wood abl schwarzwald satisi definit relax home meat trip even wine staff dine book fritz ate wife short planet antiqui lucki moritzplatz fisher cranberri arrive berlino kitchen appropri  
 remind tonight till lamb sit agre ambienc tango prepar hope spot kreuzberg drink excel lunch wait tabl enjoy qualiti expect welcom pretti surpris maultaschen touch fresh haugemacht cream cosmopolitan  
 istablish chanc tour real flavour rememb top re pork sausag michelin cours beer german delici tradit look decor cuisin waitress lot monday lone christma fare gluten exquisit combi  
 detail aisl pari accommod stand twist bougignon dream chop french piec advanc lobster worth euro walk pleasant husband spaetzl effici vegetarian apart discreet bake capit war alloc idea  
 t mushroom rest smoke neighborhood fri high bit tourist fantast friend berlin atmospher price hotel moritz cheap plate deserv miss simplest p  
 born antler phone art theatic vocal close interplay wander posit moment see world visit choic day regent pack tripadvisor cozi honest call chef spoke atleast ver  
 jke difficult glad spent size schwartzaldstuben accompani found waiter fast dinner menu superb eaten option cake trap saturday money bank apfelstrudel schnizel andor upstair visitor hoppi thr  
 rack duck meet sauerbraten floor steal cosi beauti like valu main local amaz recommend servic nice meal eat reason overr start huge poor super stew heavi amateurish schnapp omous unobtrus  
 chhose respond delight advis augustin germanberlin obtrus splendid true stori mix featur littl afford  
 ein anticip comment husbanddapear piano cost wrong pasta english seat free fish typic experi love star time schnitzel tasti fischer offer limit discov pizza menus kept compet regenc actual unfilt strong gesc  
 scene entertain barrestaur daytim invit salti build sweet finish favourit except und expens peopl serv dish perfect disappoint cook attract ok dark inform follow window simpl pay accept tire speak germani style br  
 deer situat exuber bike distanc kase satisfatori histori goulash describ chees own wonder reserv help hour portion stop warm live impress told awesom version vari parti accent thorough skull furnish stuffi bowl horvat  
 pour differ skeptic fir brew earmark month strang noodl pm eleg heart week dessert bar salad respect minut culinari seafood wish coupl rothaus appreci our wuerstchen theater similar fair struggl bell hartin  
 prune district slow folk bundl els neighbor thank server ignor central errat pair write stumbli pub ambianc includ person spirit come center newli third surviv serverhostess etc street worthwhile ff  
 quick earlier smile foot busl excurs plan varieti venu bellic decent bad bread play step quiet allow stun paid base mustard support sort rheinisch due space wide facil  
 salmon friendlier speed fussi contemporari flight prove ingredi afternoon avoid starter one birthday complain treat neighbourhood greet entre mastercard studi requir relat crispi rothausbrau succul creation  
 snow glanic starti hall dancer greedt steak convers charm sole word soup search ham type gras hearti student realiz plaster coriand ravioli sequenc assort  
 wrap goat stiff hard energi happen tag leav extens creativ level sauerkraut veget flawles gendarmenmarkt spoil realiti pianist commit proper sacrileg adlon  
 ceremoni goulasch swept inferior fat hot test appet return pass uk yum fit spit quantiti norm circus profit rage whip  
 compromis grand tad intimid hit incred wanna left facki corner enter snowi proceed museum cheer plenti pocket tomato  
 crust hungry temporari jewish holiday late avail appear custom road prais mount characterist persuad plus synagog  
 demand influen terrin joint hug mask celeb rains milong mit brother pear pleas strawberri  
 flavor lemonad translat kid kiss offic manag loaf bland patron mound southern  
 forest gut light truth kitsch kooki land appeal overheat mean pancak  
 furnitur insid line violin klop occur onion massiv neue  
 inexpens jewel linger white mitt juici kaiserschmarrn

# Fast Food Restaurants

verd  
 add tap uk  
 ars tomato unab surfac salsa  
 barbecu bog transplant arriv heard ubahnloc superb refresh  
 factor condescend tube bean heinrich accoutr hasir toto silenc mountain  
 meati negat ubahn capac highstreet bahn hardest wish guid tenderloin sell kitchen  
 meista pleas vegetarian dish hung basic heat poor glad unprenti great sweet screen kind  
 oomph smart vibrant folk institut bathroom involy recomend da nyc flavour tender goe sad sandwich harri  
 origin standard victorian french ladi beaten judge schlesich attract mc former neighborhood febriari tempt glass prove rib guacamol  
 overpow starv afternoondecid fridg marag burgersteak kid slack bestburgerev mind provid donald west london expand suppos girlfrend period recent gourmet  
 veget tastiest bump hey messi burgerthen med ton choic moment queer cheesi ppl stomach real hubbi enen structur gem peak receiv gonza  
 advis wintri call mistaken orderingcollect chain medium unconsid crispi relax attent averag middl live montana finish beef feta enclos stroll free overpr rapid waterfront  
 divers appet clever morn plain citi monday underpass dine short booth burguer head star help terribl meet build normal buy custom spici entir nois prefer transform  
 burgerheaven deli oberbaumbruck platz coleslaw nicer toast dress futur diet napalm http set grill clean compani tire martini addit hand burrito convinc skimp email month planet succul  
 differ process complaint obvious accomod fashion pleasant bbq lover move bite cheeseburg awesom coupl saturday reason limit hilltop thick lack bettercheap cold singl eleph machin phenomin  
 drive readi cooler pain adequ funk track expens hostel reloc neighborhood choos crowd line hip york boyfriend husband greasi boston wonderful muffin wan coffe say door leav pay  
 footbal peper regular crave postiv chill ghetto favorit sauc half stay joint come sat fresh reserv book fantast lot chicken afternoon fat anticip tyre ketchup tie circus river district knife  
 cultur rate clemati homeless ny chat understand birdhous littl spot public perfect review visit qualiti disappoint day wonder miss smoke stop serious total hous thunder care ridicul pace  
 rosenburg connoisseur isra cover centr diner popular grab expect street trip time price worth atmospher cool toilet cut waiter juici station break slow speed main strong canal railway  
 cucumb jack battleship classic feed homemad decent kreuzberg bla ral bar eat friend servic staff tri beer chilli read fine advanc even lamb word seen left sluggish cafe molecul odd  
 luncheonett biggest rang world incred top drool veggi look busi **chees** food  
 season destin park cheescak wide chili enjoy noisi yum super baconchip wait  
 god expert slighti exot pint coke life spent town true make galleri cheap bird  
 get tantai fell prici disappoint ambien watch outsid dinner german east experi  
 plate highlight think goo pub dive design girlfriend happy local style nom re hamburg love **delici** fri recommend american wing serv week famili yummi queue stun high drag potato alley  
 rent tripadvisor guy refuel greas breakfast casual speak altern hope meat heaven tasti restaur **excel** brilliant drink hot simpl quick share bun pricey heavan delish popper vibe mine  
 hilari play salti hype turkish hard shake mail traffic melt hungri found eaten money tabl locat menu special easi bad stand europ parti due hamberg buzz pie tuesday light discuss hipster  
 florida holiday pop satisti imbiiss type lacklustr spars scarc usa start client meal availbeach lunch huge english weekend interior uniu varieti corner eateri businessman outsid speedi island  
 collect kiez red shot opinion whopper manner surround sideord treat prepar offer hour step palm music excit dont mouth effici overr beefi bill memor shock hnnim compet goodish mound  
 gooey comfort list refus skip particular worring met wss sink effort tor cook ok portion patti chang charm server convert filthi beauti measur road hairi comment extra mayo  
 ground command load roll squar posit yes one yesterday space tourist lost walk pack venu option rememb cheddar ultim bark like picnic combin easili massiv sound  
 home consid mood signatur strass rid accid overal heart superburg minut medium sit welcom orient advisor pickl atlant lad manic environ chao downsid lonley rock  
 hop convernut similar terrif truck boot person ticket teenag chilicches near helpful quiet schlesisch spectacular journey lunchsnack elev cent dollar lick ribey  
 intern cost previous sooth zum typic cake polit union variat fault includ america base jalapeno imagin difficult canada dip lettuc reopen  
 jazzi doubt report suggest ambient volum check pretti decor belov match thrive grungi euro cute burst counter lead quest  
 laid eg revisit trap bigger warn dodg reccomend fill stag grit decid critc bottl continu hollywood paradis  
 loyal entertain snack unless boy written feel salad fuhrerbung close cours birdhous condiment han ooz  
 minimum exempl supermarket adalbert calori wrong find chanc colleagushash complet friday onion  
 mustard eye system advisorfrom cash xberg closest bartend competit fiend north  
 news fair take alcohol chose apart blare fatti nonsens

# Hiliting in Tag Clouds



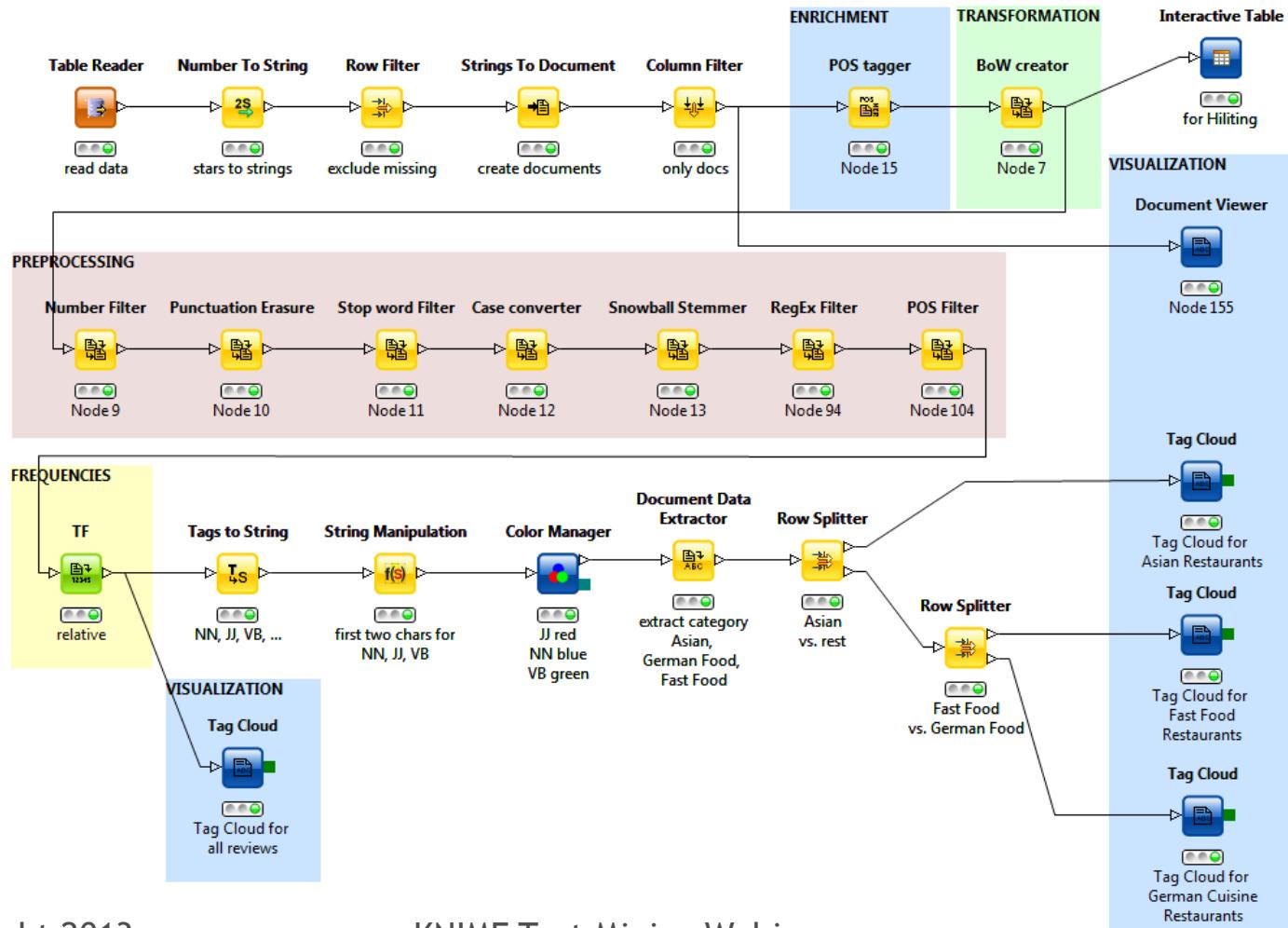
Interactive Table

**Table View - 2:158 - Interactive Table**

**File [HiLite] Navigation View Output**

	Hilit Selected	Document
R	Vietnamese[N(POS)]	"lived up to the hype"
R	Vietnamese[VP(POS)]	"Vietnamese fusion at its best."
R	Vietnamese[J(POS)]	"Vietnamese fusion at its best."
R	<b>Show All</b>	
Row3551	vietnamese	
Row4456	Vietnamese	<input checked="" type="radio"/> Show Hilit Only
Row5723	Vietnamese	Show Un-Hilit Only
Row7006	Vietnamese[J](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[J](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[J](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[J](POS)]	"Best asian in a long time"
Row20857	Vietnamese[NNP](POS)]	"A great experience"
Row21070	Vietnamese[J](POS)]	"A Unique Place to Visit"

# Workflow



# 7 - Topic Classification

# Document Vector

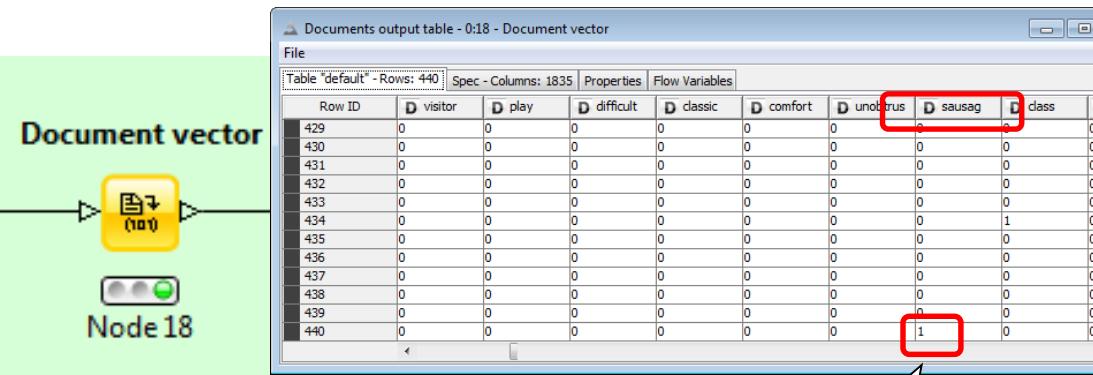
Terms and documents output table - 0:16 - TF(TF relative)

Table 'default' - Rows: 8150 Spec - Columns: 4 Properties Flow Variables

Row ID	Term	Document	Orig Document	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	atmospher[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[NNS(P...]	"german pub"	"what a German pub should ...	0.118
Row19	potato[NNS(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

Table "default" - Rows: 440 Spec - Columns: 1835 Properties Flow Variables

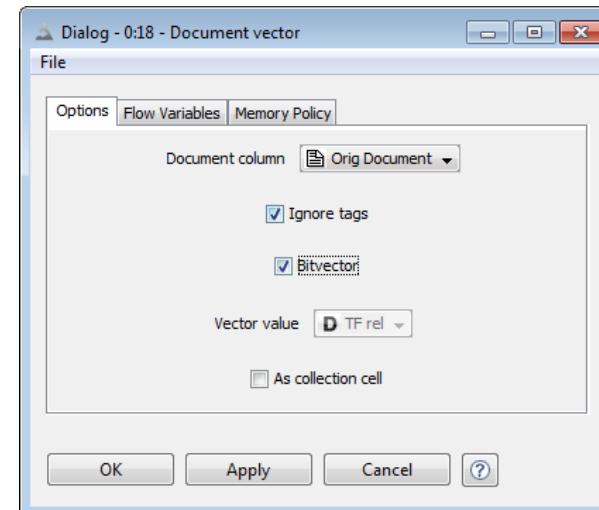


```

graph LR
    A["Terms and documents output table - 0:16 - TF(TF relative)"] --> B["Document vector"]
    B --> C["Documents output table - 0:18 - Document vector"]
  
```

Row ID	D visitor	D play	D difficult	D classic	D comfort	D unicolor	D sausag	...
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

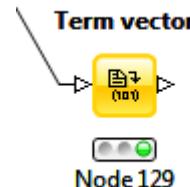


Bitvector or frequency measure

# Term Vector

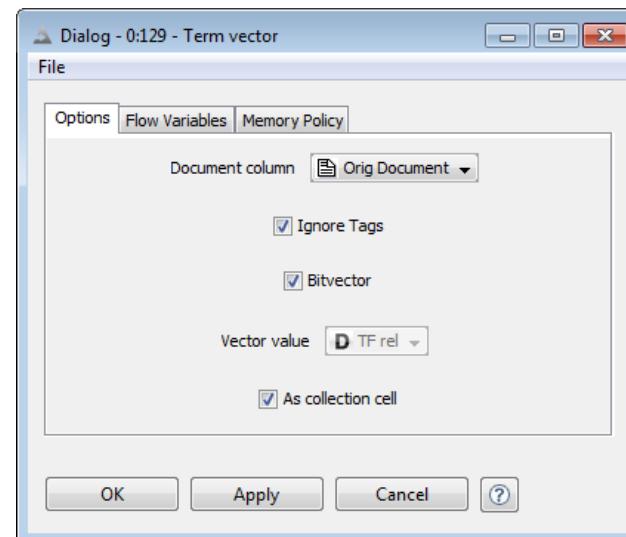
Terms and documents output table - 0:16 - TF(TF relative)

Row ID	T Term	Document	Orig Document	D TF rel
Row3	german[J(P...]	"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	atmospher[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[NNS(...]	"german pub"	"what a German pub should ...	0.118
Row19	potato[NNS(P...]	"german pub"	"what a German pub should ...	0.059
Row21	mustard[NNP(...]	"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,...]
128	beef[NN(POS)]	[0.0,0.0,0.0,...]
129	beef[JJ(POS)]	[0.0,0.0,0.0,...]
130	beer[NN(POS)]	[0.0,0.0,0.0,...]
131	beerdik[NN(...]	[0.0,0.0,0.0,...]
132	beetroot[VB...]	[0.0,0.0,0.0,...]
133	believ[VBP(P...]	[0.0,0.0,0.0,...]
134	bell[NNP(POS)]	[0.0,0.0,0.0,...]
135	hell[NN(POS)]	[0.0,0.0,0.0,...]



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!

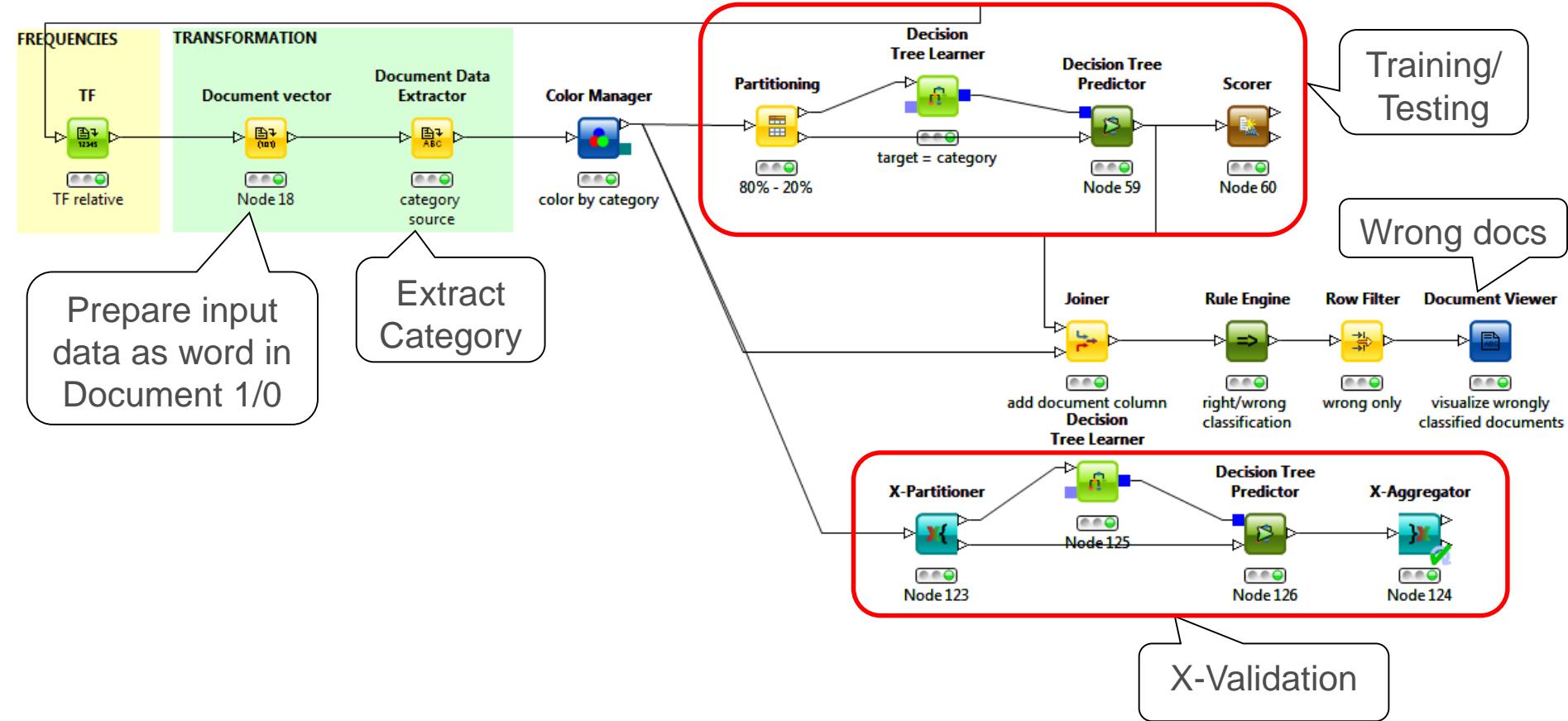
Target Topic is in Document Category.

# 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

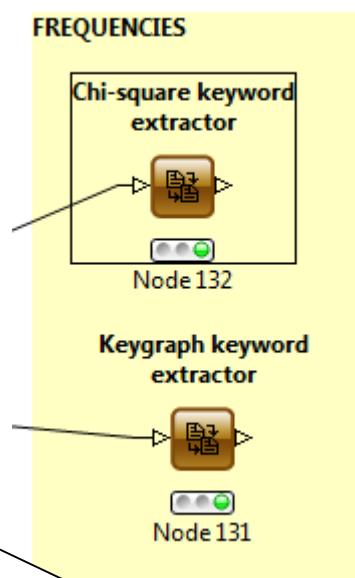
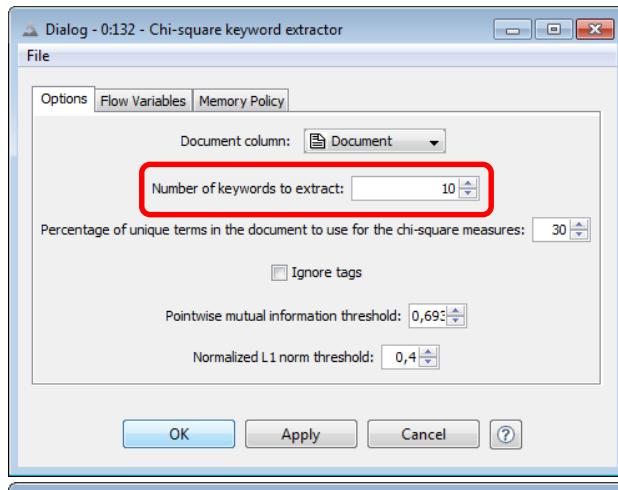
## Problem 2

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

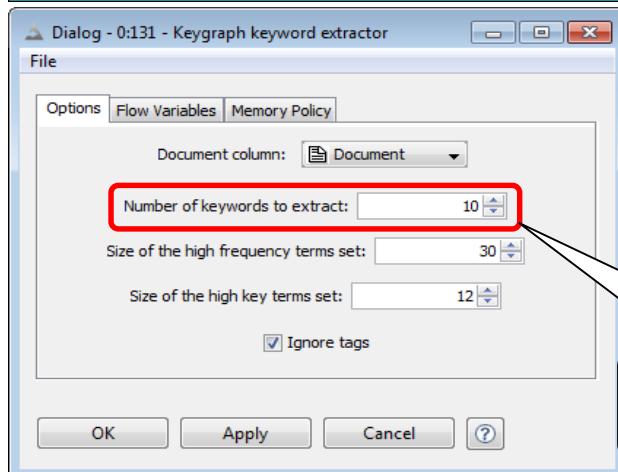
To reduce the input space dimensionality you can:

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

# Keywords Extractor Nodes



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



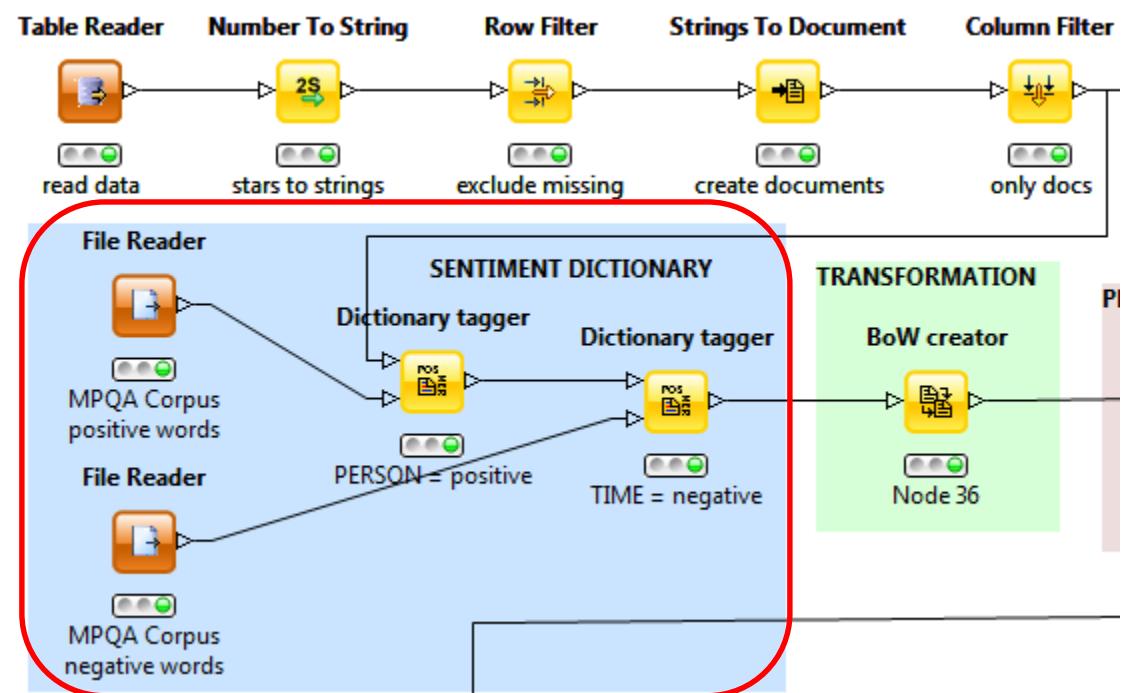
Max. # keywords per doc

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

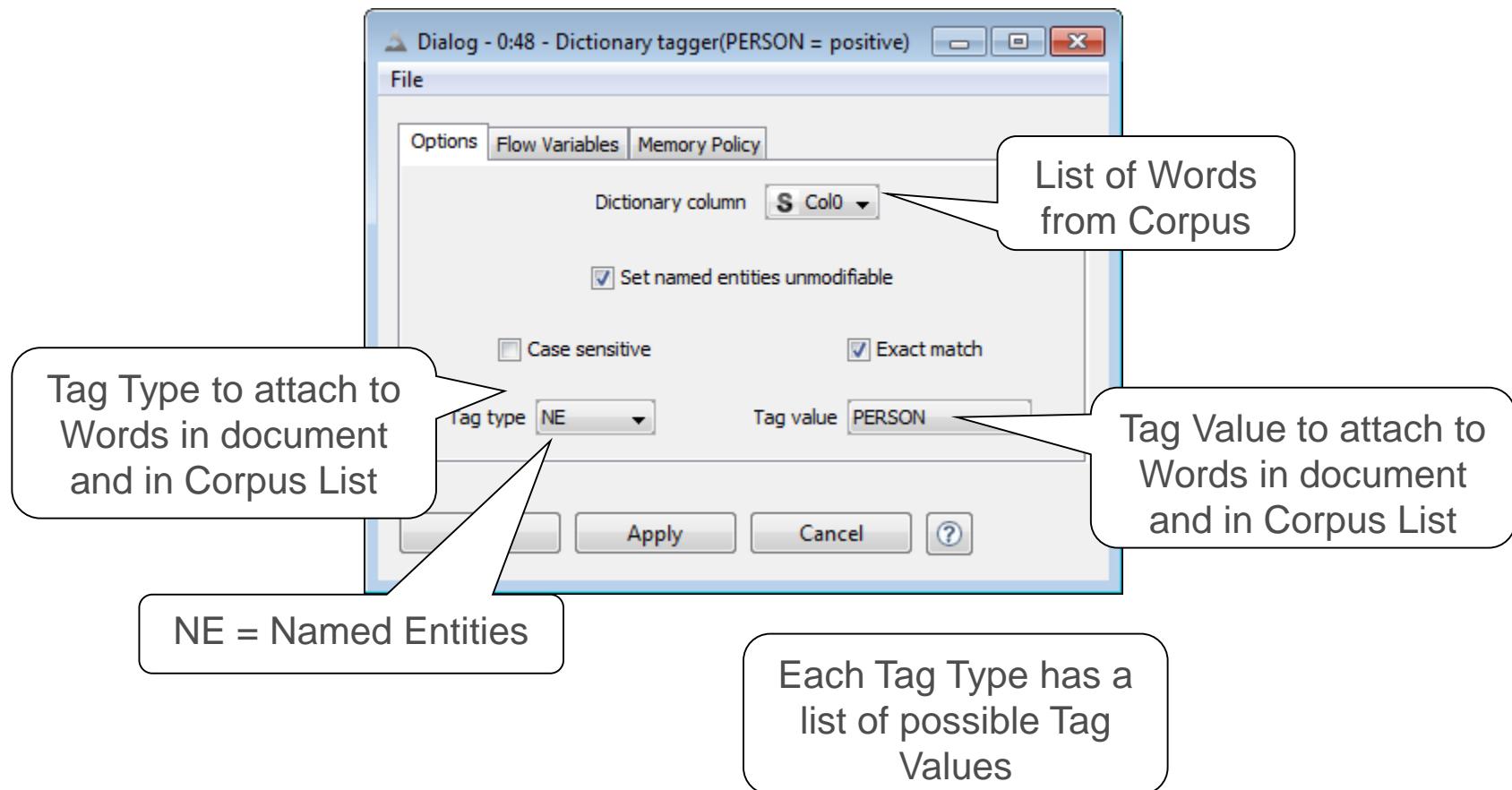
# 8 - Sentiment Analysis

# Sentiment Corpus

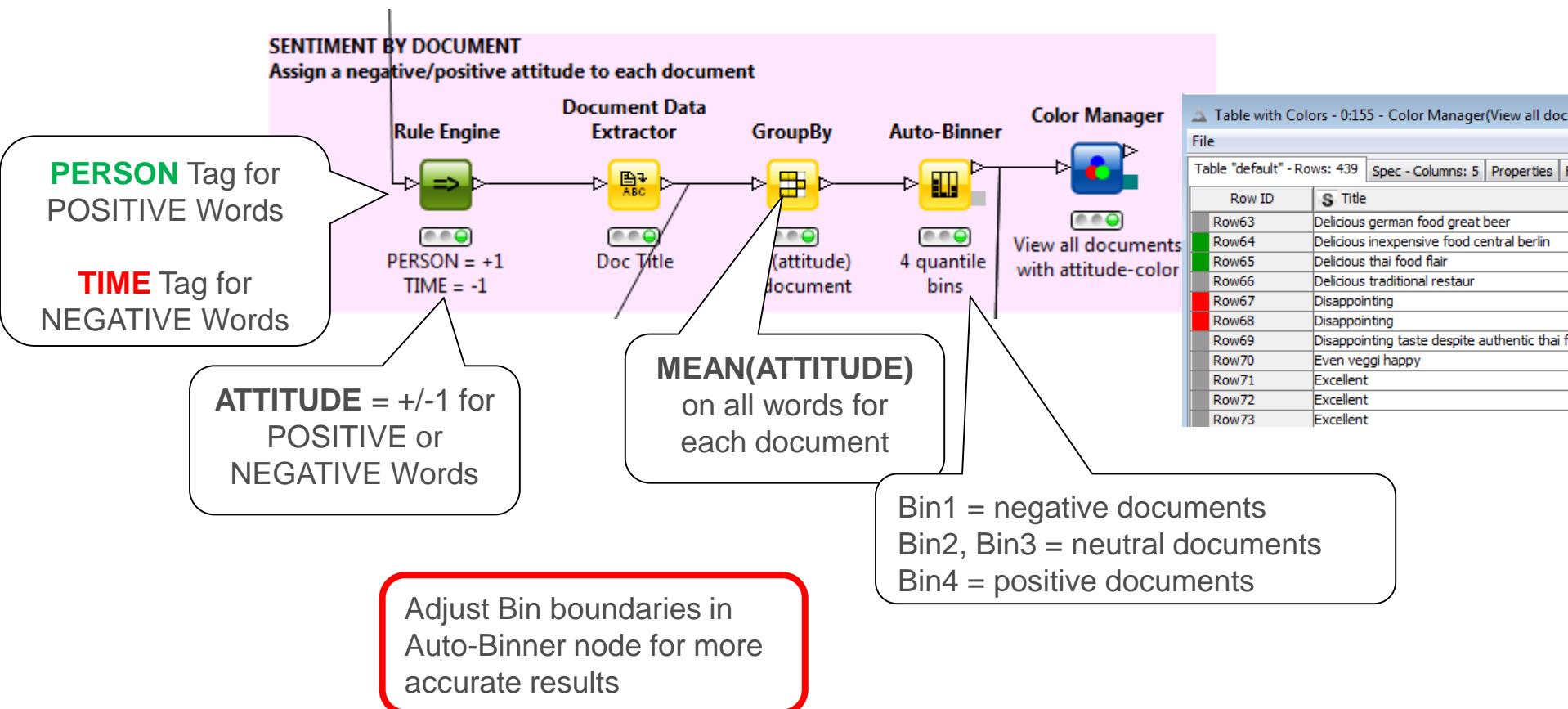
- MPQA Corpus with negative and positive words
- Tag Words according to Corpus with a Dictionary Tagger Node



# Dictionary Tagger Node

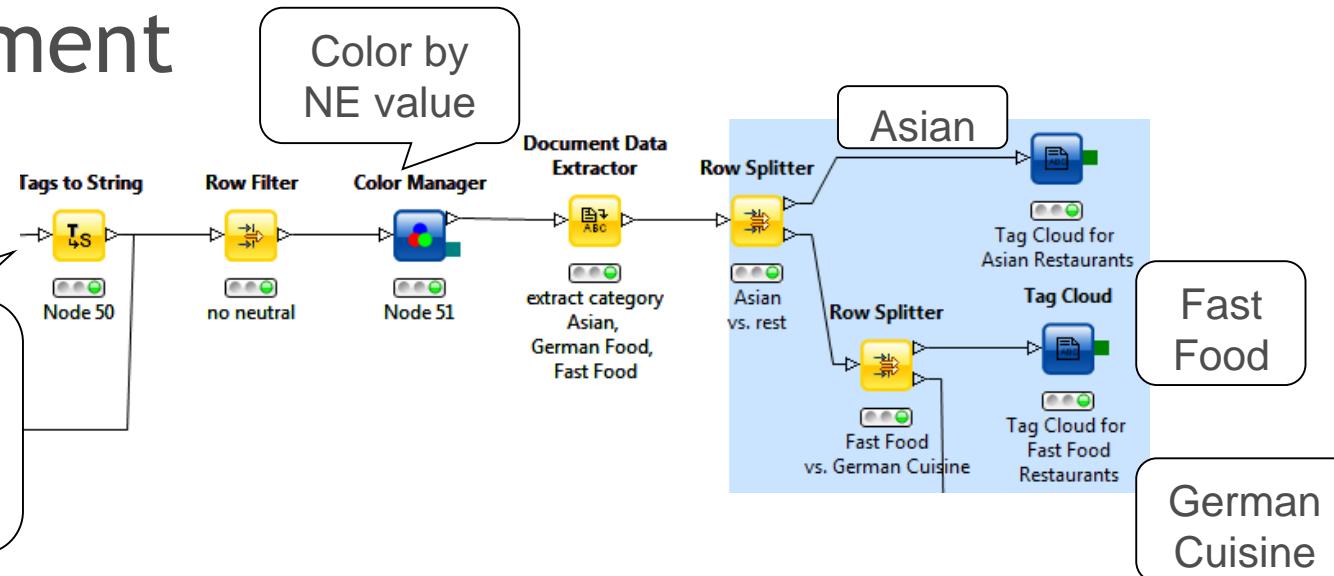


# Sentiment By Document



# 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. Rich Beautifully  
 Enthusiastic Sophisticated healthy cold wrong shy want Worth exemplary Awesome  
 Thank insanely complimentary Better skeptical poor romantic understated Will excessive truth  
 Unbelievably lucky correctly Fabulous slow prefer unfortunately pleasure regret tap Unusually exceptional sincere  
 Fine memorable cozy Fantastic spot prompt worst cross stunning pleasing quaint timid Solid enthusiastic passion  
 deserved Inconsistent traditional Real Amazing friends snack Well sullen perfectly properly superb Simple ensure to  
 disagree trick recommendation Blend happy helpful cheerful sense Delicious problem miss problems still 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 event taste expert fun unique open  
 live recommend excellent unanimous hard intimate ment skeptical elegant  
 Word prepared close will welcome back good friendly value efficient  
 delicious reasonable Nice super  
 smart Flexible pretty right gem busy fresh great best authentic disappointed lies down  
 play drive advice warn sunny Gorgeous righty  
 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 Medicore Superb funny less Fabulous unfriendly praise fair Free jovial relax Perfect  
 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 crowning blend disappointing haphazard plenty worries clearly disappoint  
 accept Waste suspect impress forget deserve genuine pwonder charming comfortable  
 agree attractive sweet inoffensive generous original unbelievably charismatic core  
 amazingly battle terrible moderate treasure brilliant hot

## Fast Food

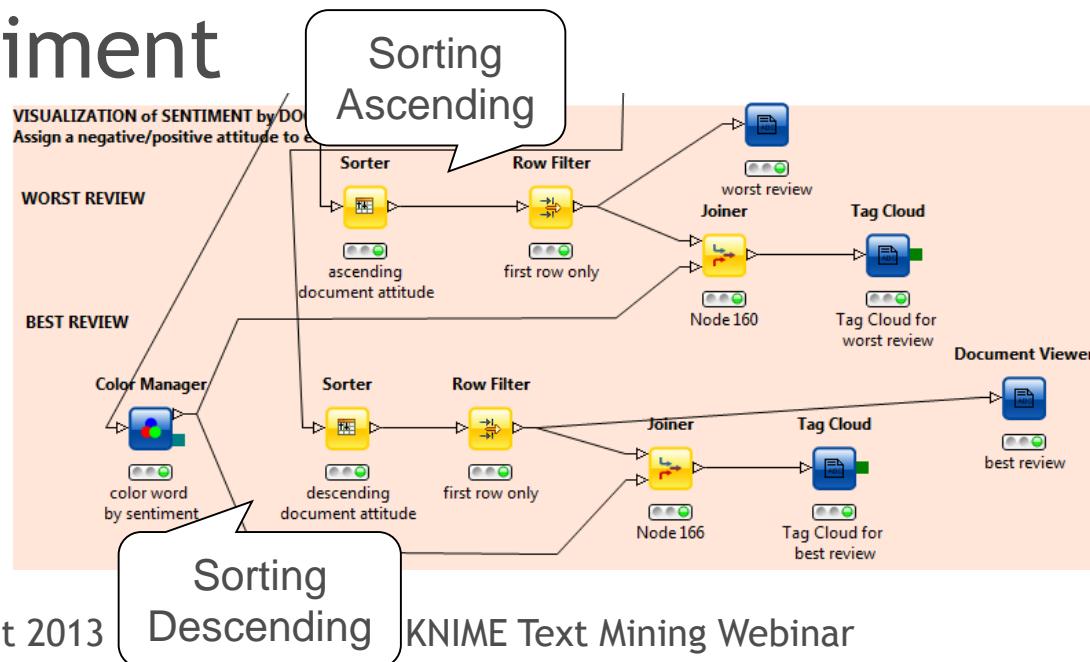
delicious  
 important Large STAKS  
 Back Lively AMAZING Hope  
 Disappointing Score BEST 2nd whatever Fine  
 Fatty Smokey Burnt Greasy Too Expensive  
 LOVE Ultimately Classic helpful comfort pain Queer extraordinarily truly Finally  
 Love beautiful Clearly hilarious criticism poor easily lively HOT excitement subject Cheap  
 Sadly cold DELICIOUS hot difficult real fell impressive complaints highlight Filthy elevated strong  
 disappoint Fantasy back enjoy reason loyal less Nice heart clever despite word downside stars Although  
 locking frozen supposed mind need below recommended Just decent clear close utterly disappointing  
 knife understars miss polite grill reservations value happy terrible clearly although chose 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  
 lower perfectly Perfect positive stunning might super help popular wonderfully sticky  
 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 resonance retreating Vile sluggish hungry seas Fun kind Amazing extremely manic Easy unpretentious  
 upper Funny and advanced unable impression hand wide Please low want simple serious serious clock  
 disappointed GOOD bad strong sound Fat incredible trying basic break running Ridiculous  
 Wall entertainment Irreducible plainness-focuser classic Even doubt  
 above especially Lovely problems stand condescending thrive sweet Wonderful odd original  
 disappointment excited Moving reasonable render speed Unique importantly memorable  
 Favorite fault Open refuse Simple hope light  
 True Friend Phenomenal did best

## German

energetic  
 tender fuzzy popular  
 trendy wild Beautiful fun incredibly  
 true greedy Best smile Admittedly Famous honest  
 ultimate hard Busy sturdy spirit slow worse Fair goes  
 mouthwash hairbrush toothbrush mouthwash toothbrush  
 affordable hungry Unfortunately struggle welcome reason friend hospitable playful pricy unavoidable erratic discreet  
 Attentive ignore Unique super word skeptical genuine exception flawless desperate last polite fire enjoyable complain  
 Awesome impressive Want surely Genuine treat heat strew 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 Cosy charming agree heavily painful stumble  
 twist profit interesting extremely Nice recommended Good warm difficult sure truth Stars Traditional fat kooky  
 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 right definitely live real relaxed Lovely exquisite  
 Perfect tradition able want kind will traditional Great friendly long disappointment cheap large  
 warmly trying commitment affordable miss authentic lovely worth bad Delicious fantastic least reasonably GREAT  
 wrong cold Star Amazing disturbed attractive recommendeven amazing taste efficient gorgeous rage useful deserved  
 Friendly competent beautifully Delicately enjoy expensive try busky dark electric problem respect delighted irritation stuffy  
 Bed Warm Lonely disappointing Super recommendation fresh try down poor junctions  
 Erratic lonely amorous Pretty and attractive respect cathartic  
 Grand might celebration reasonable fine cost  
 Help noisy correct Straight kid incredible happy appealing happy remarkable cheerful Spot  
 Poor obtrusive dance Top lack onus gladly relief bitter Perfection  
 Cheap Splendid perfect easy above entrance proper appreciate Expensive  
 disappoint Wonder fat pleading embarrased  
 sometimes accept plus 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

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

## Most Negative

drink  
**expenspoor** burger  
**pricei** tasti nyc  
service unconsid staff  
**fineslack** london  
head  
better cheap

# 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](mailto:education@knime.com)