

Open for Innovation

KNIME

KNIME Text Processing Overview

Scott Fincher, *Data Science Community Manager*

Elisabeth Richter, *Data Science Publisher*



Before we start...

- Agenda
 - Webinar – 45 minutes – 5 PM (Berlin) / 10 AM (UTC -6)
 - Q&A – 15 minutes – 5:45 PM (Berlin) / 10:45 AM (UTC -6)
- Ask your questions in the Q&A
- Session is recorded and will be available on YouTube
- Slides will be available as well on the KNIME Forum
- Example workflows are available on the KNIME Hub

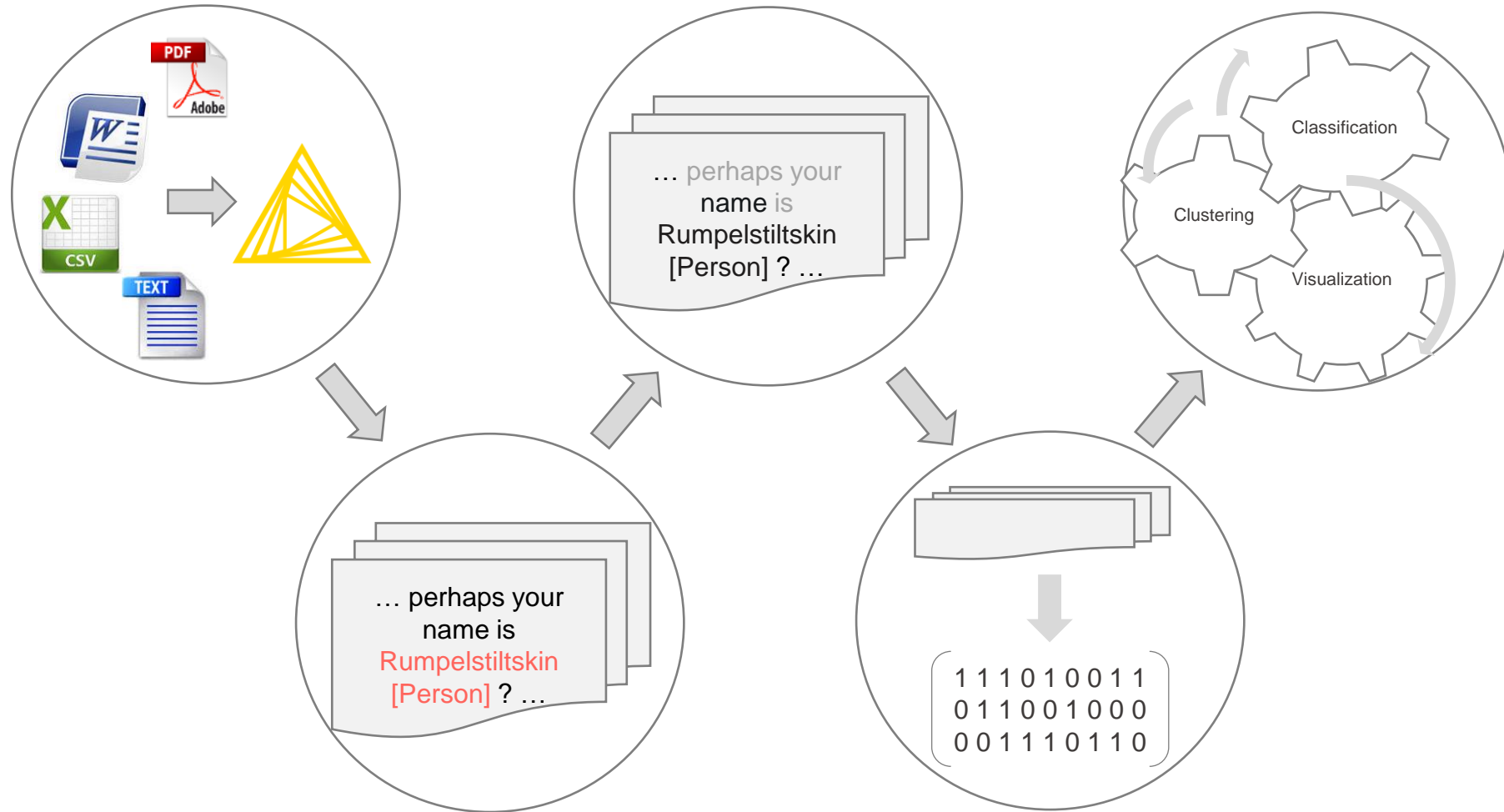
Agenda

- KNIME Text Processing Philosophy and Specifics
- Sentiment Analysis Example
- Topic Analysis Example
- Social Media Analysis Example
- Resources / Wrapup
- Q&A

Other Text Processing Use Cases


- Examples on the KNIME Hub for...
 - PDF Parsing / Tabular Extraction
 - Optical Character Recognition
 - Free Text Generation
 - Active Learning for NLP
 - Deep Learning for NLP

Philosophy



Additional Data Types: Document

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

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

Additional Data Types: **Term**

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

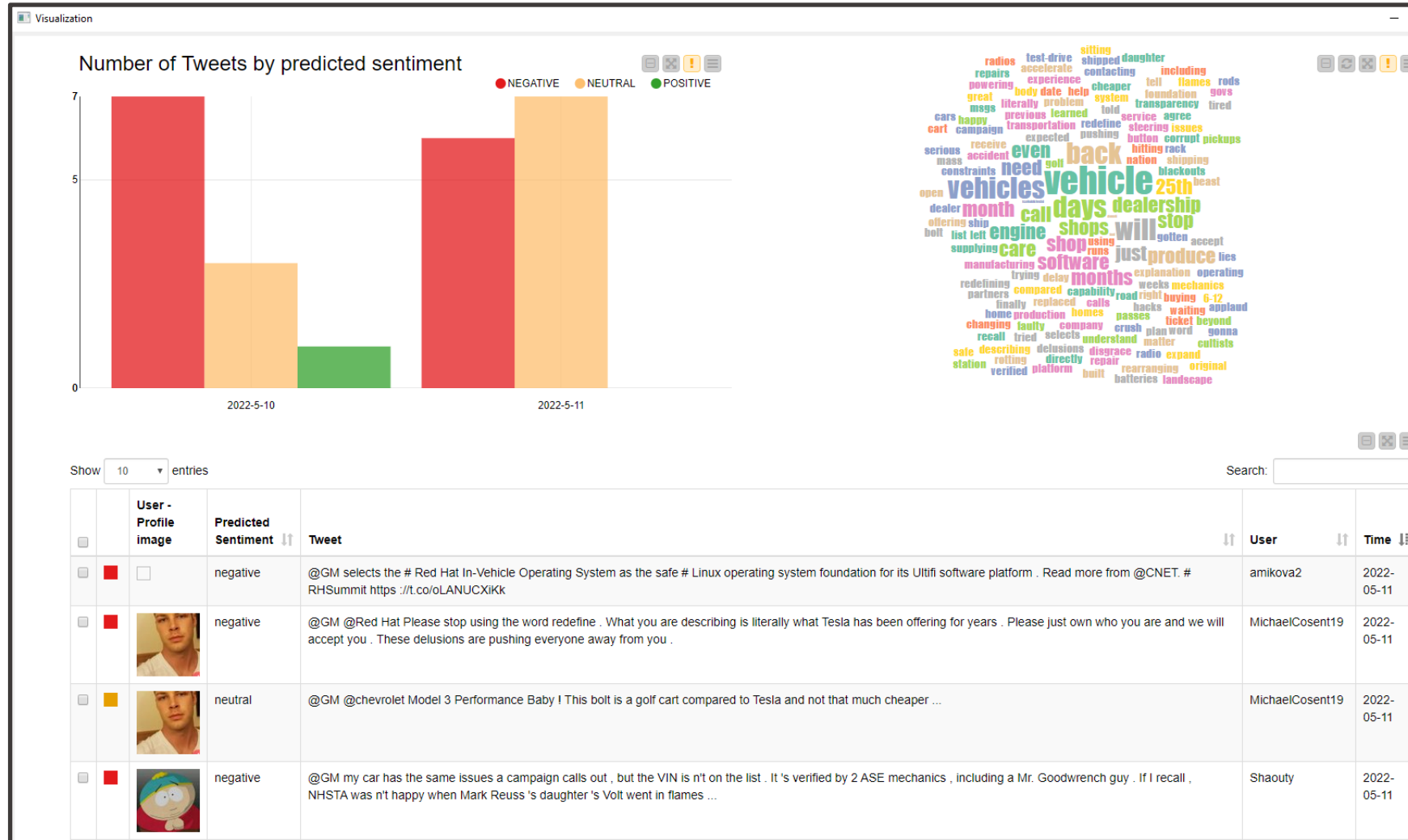
token → I → Pronoun (Part Of Speech)

token → love → Positive (Sentiment)
token → love → Verb (Part Of Speech)

token → Sevilla → Noun (Part Of Speech)
token → Sevilla → City (Named Entity)

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

Sentiment Analysis – A Teaser!



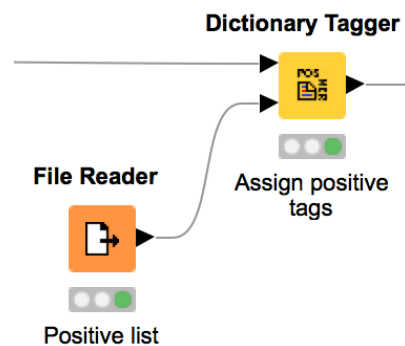
Sentiment Analysis – Movie Reviews Example

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

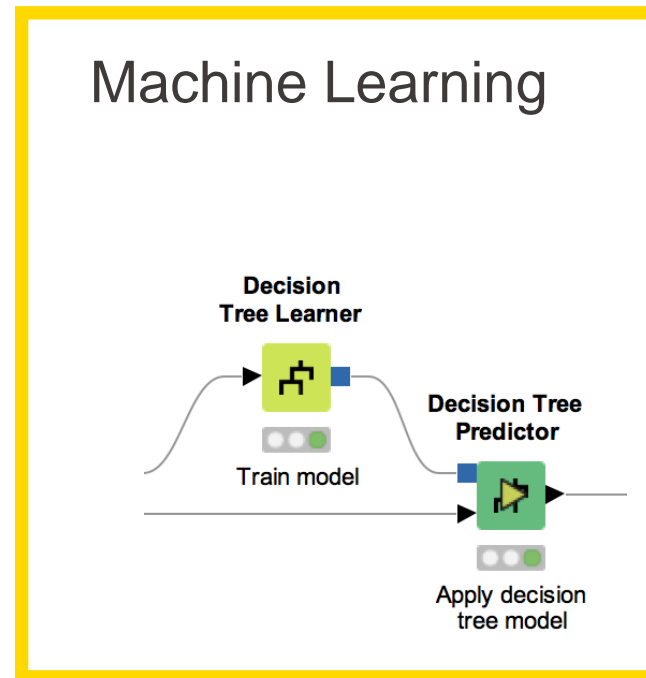
Our first example: *IMDB movie reviews*

Sentiment Analysis = Opinion Mining = Emotion AI

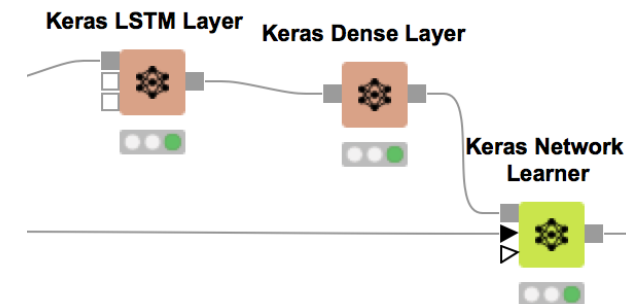
Lexicon Based



Machine Learning



Deep Learning

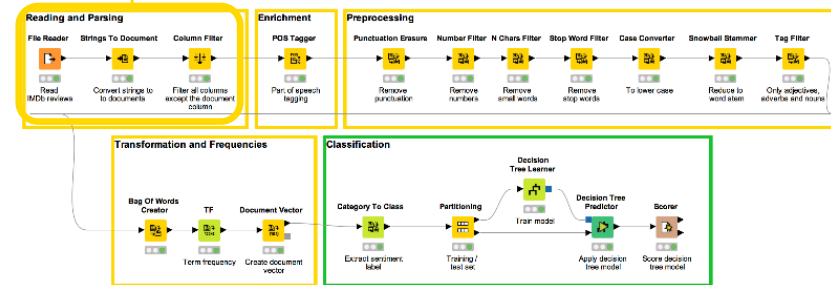
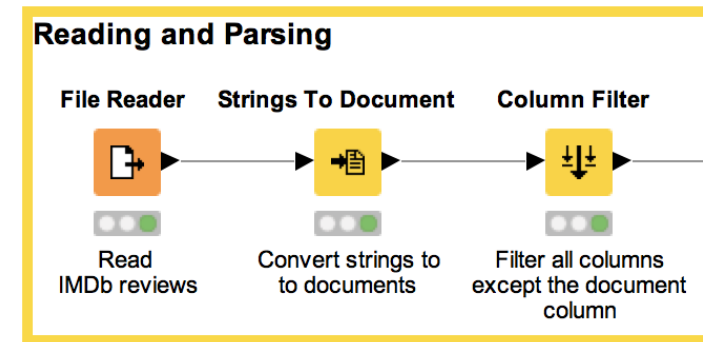
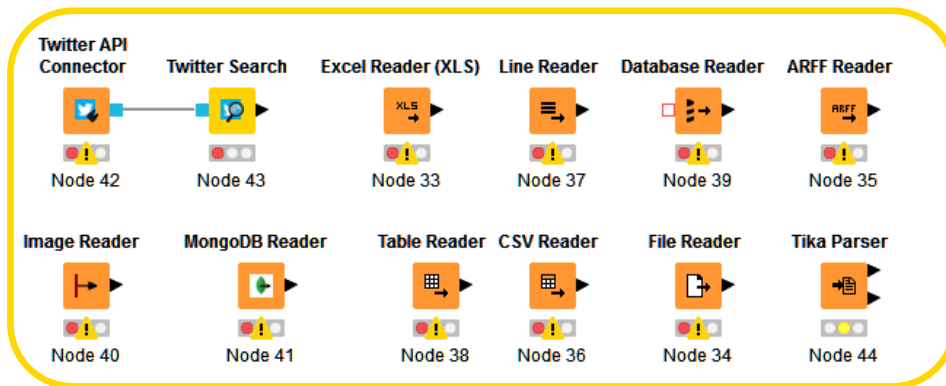


Part 1: Reading and Parsing Data

Read/Parse textual data

- KNIME Labs
 - Text Processing
 - IO
 - Dml Document Parser
 - Document Grabber
 - Flat File Document Parser
 - PDF Parser
 - PubMed Document Parser
 - RSS Feed Reader
 - Sdml Document Parser
 - Tika Parser
 - Word Parser

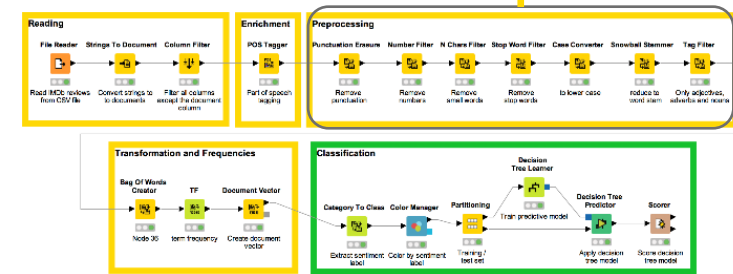
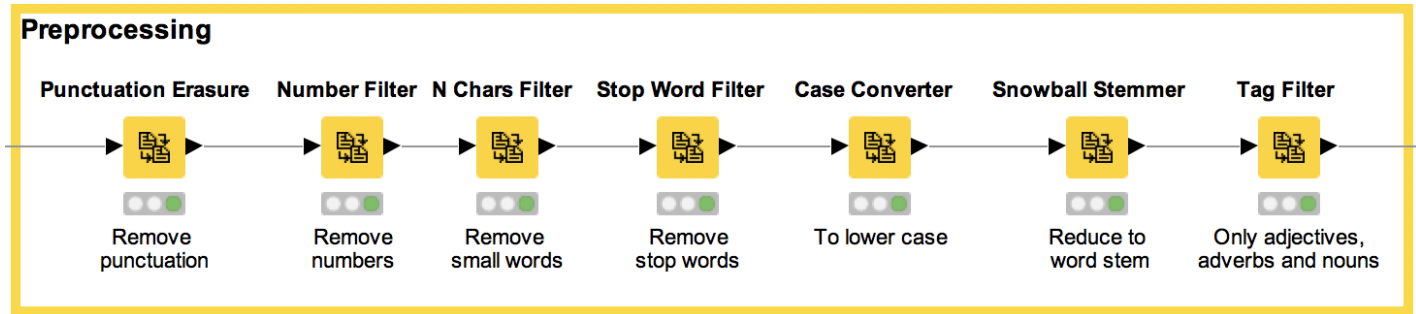
Other Reader nodes



Part 3: Preprocessing

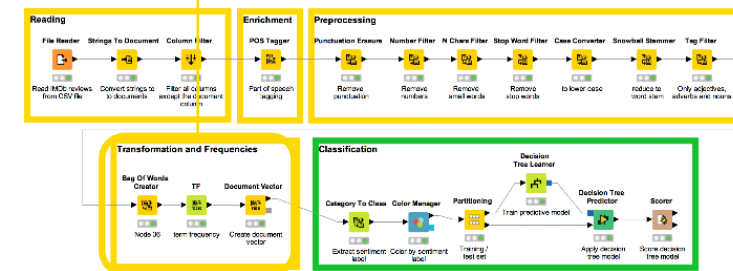
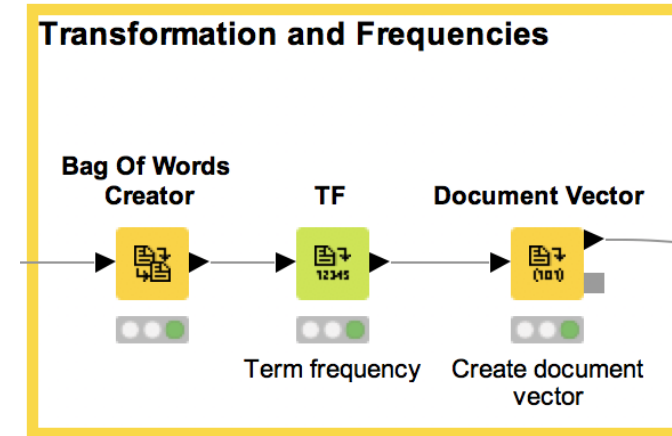
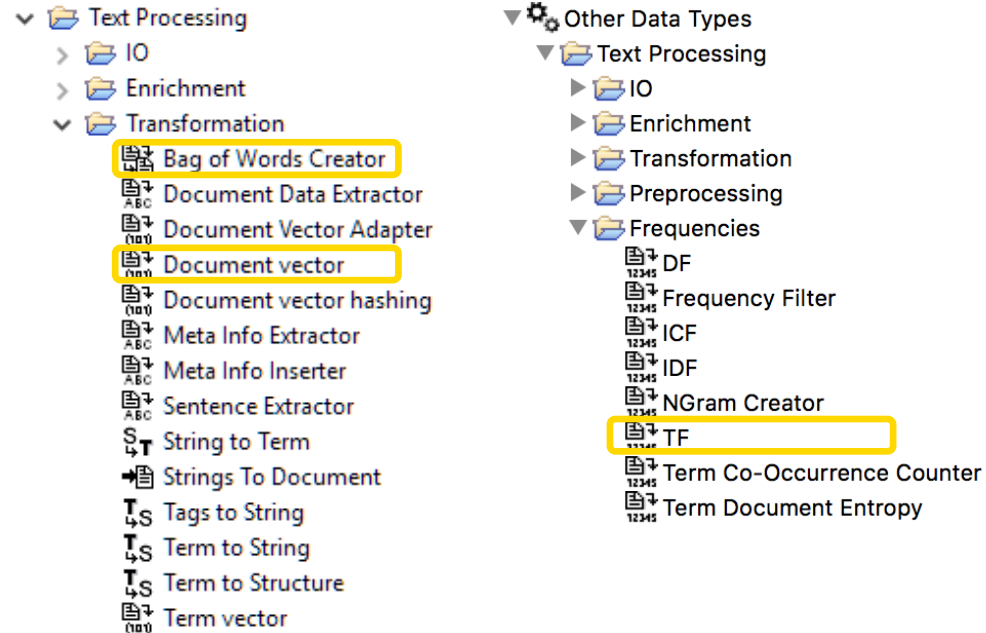
Preprocess documents and filter words

- KNIME Labs
 - Text Processing
 - IO
 - Enrichment
 - Transformation
 - Preprocessing
 - Case converter
 - Diacritic Remover
 - Dict Replacer
 - Dict Replacer (2 in ports)
 - Dictionary Filter
 - Hyphenator
 - Kuhlen Stemmer
 - Modifiable Term Filter
 - N Chars Filter
 - Number Filter
 - Porter Stemmer
 - Punctuation Erasure
 - RegEx Filter
 - Replacer
 - Snowball Stemmer
 - Stanford Lemmatizer
 - Stop word Filter
 - Tag Filter
 - Tag Stripper



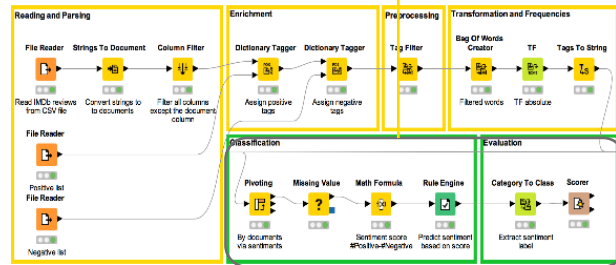
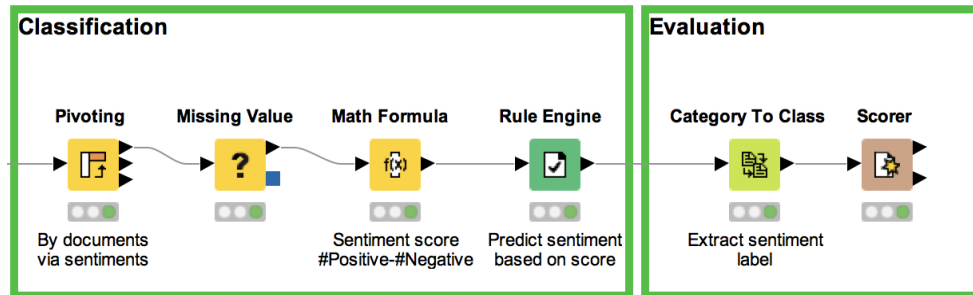
Part 4: Transformation and Frequencies' Computation

Preprocess documents

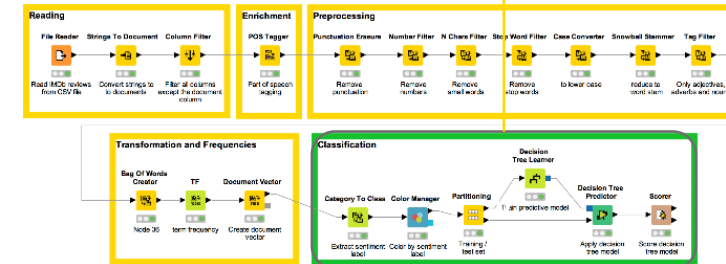
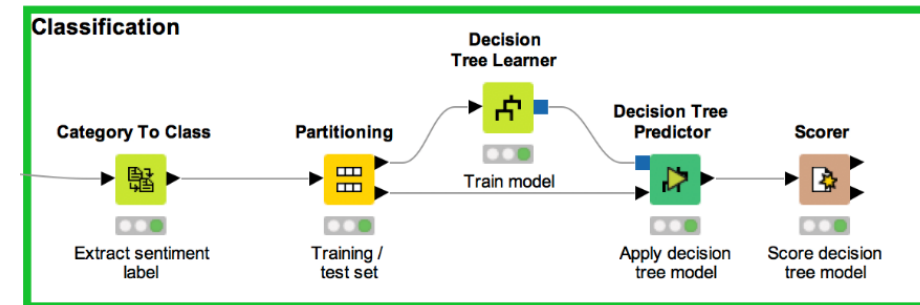


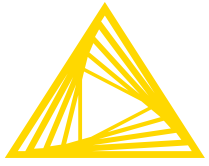
Part 5: Classification

Lexicon based



Machine learning

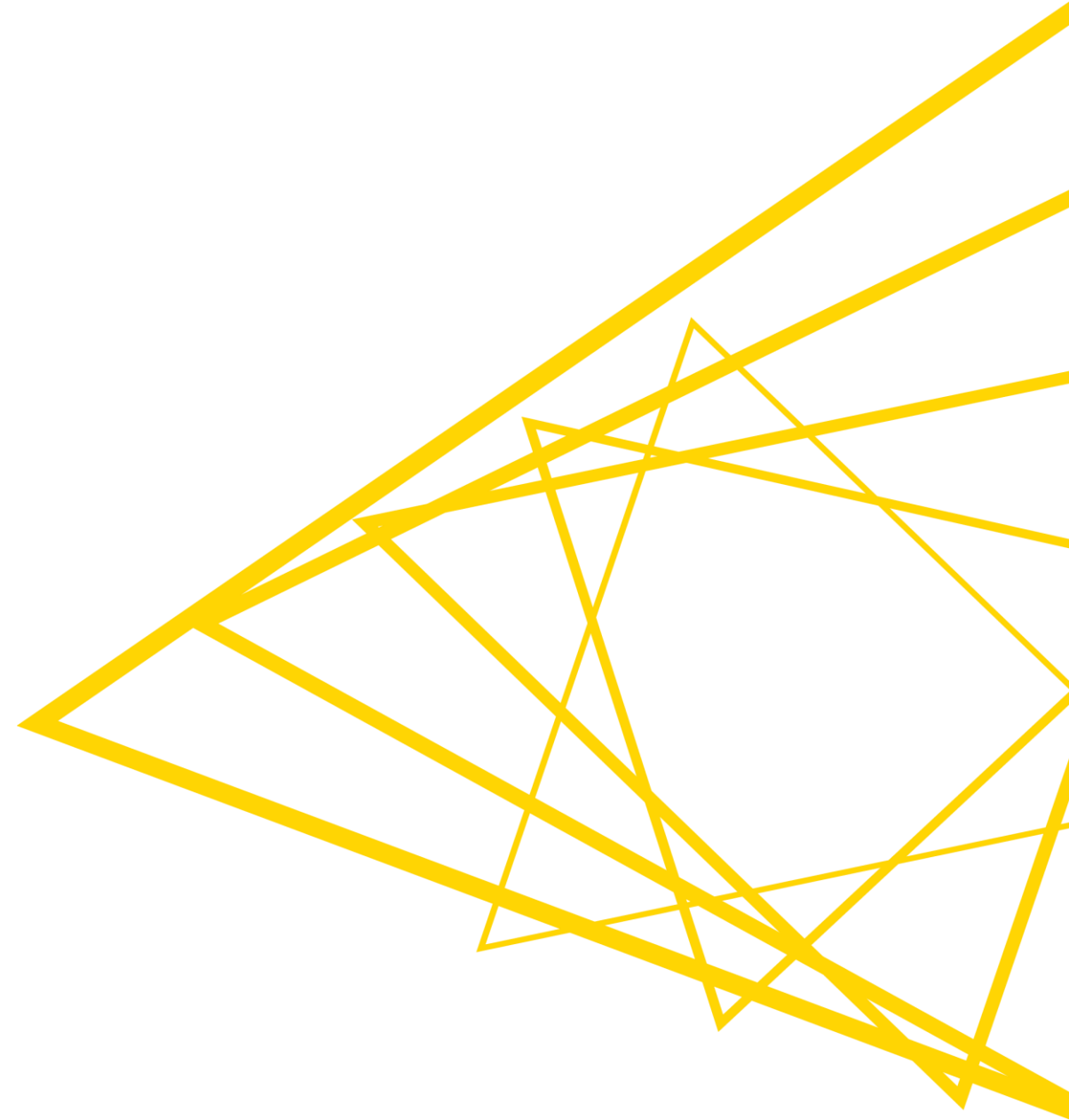


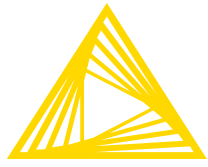


Open for Innovation

KNIME

Sentiment Analysis Demo

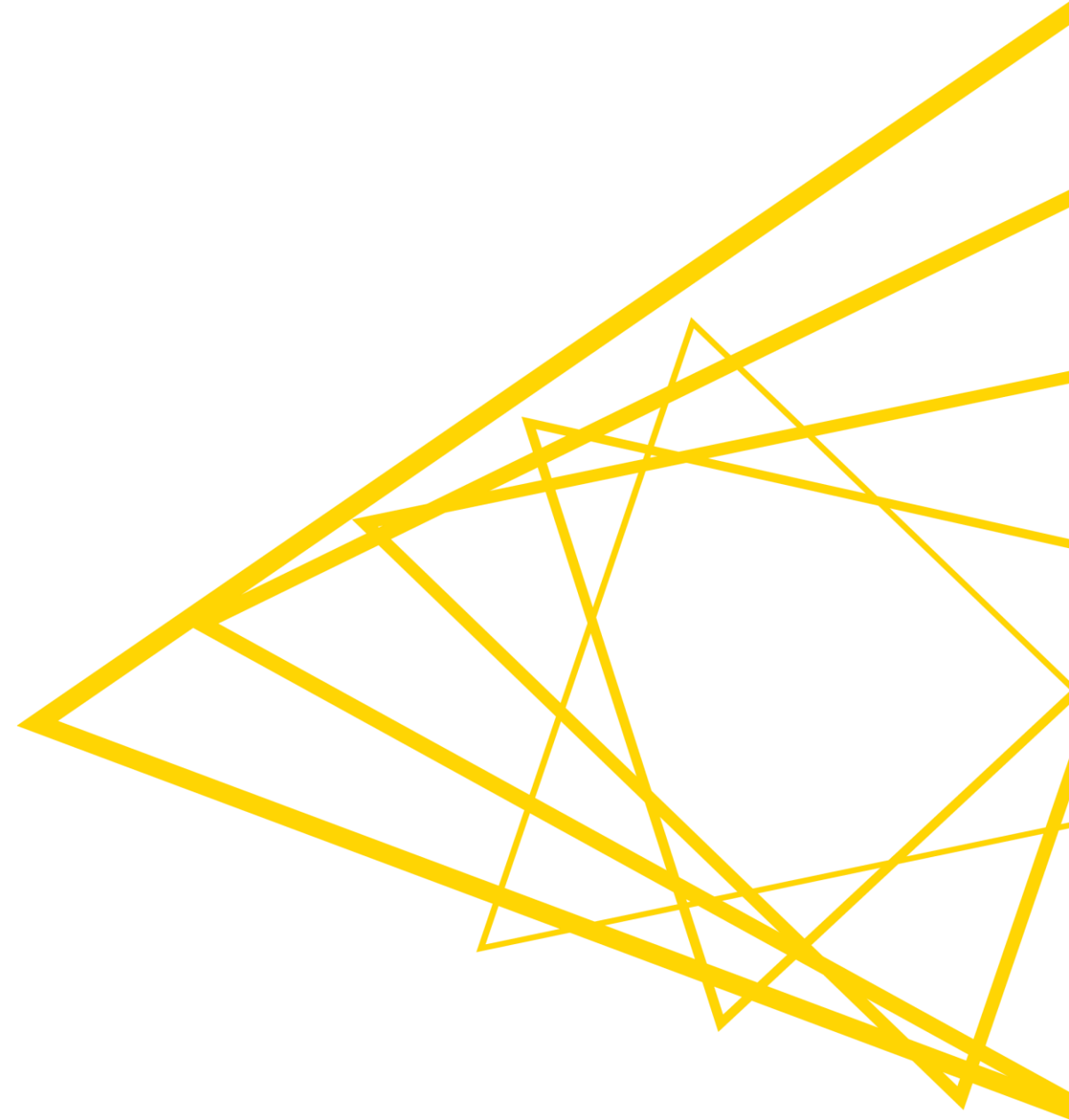




Open for Innovation

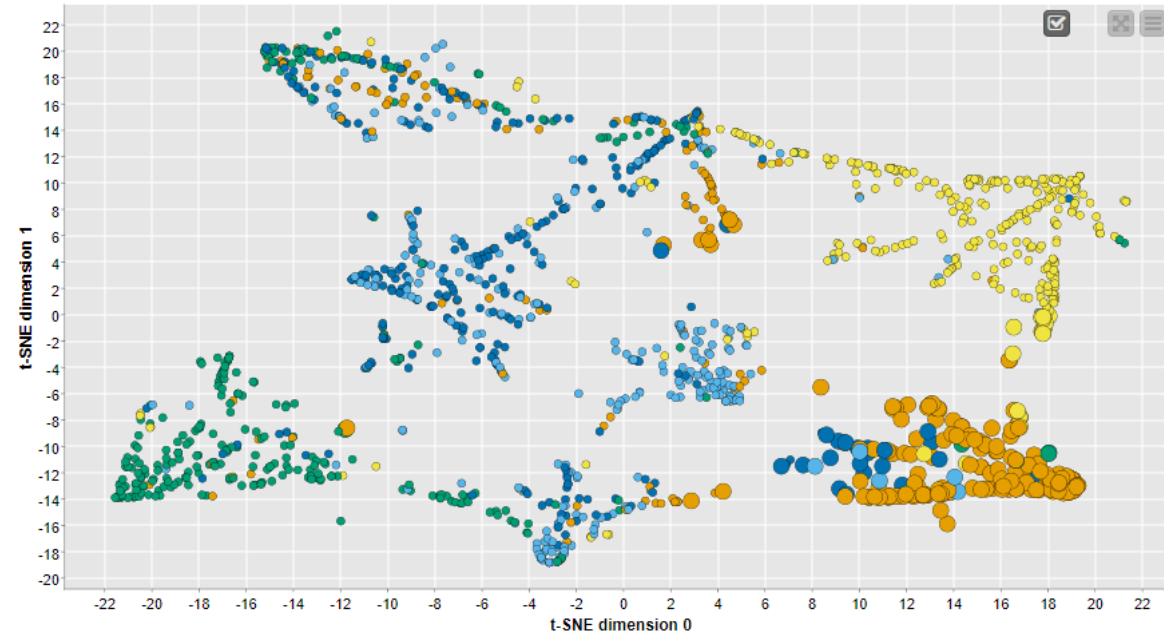
KNIME

Clustering and LDA



Clustering

- Find groups (clusters) of **similar** documents
 - Topic detection
 - Exploration
- **Unsupervised learning**
- We can use **standard KNIME nodes** to cluster the numerical document vectors.



List of topics

Clear Sorting

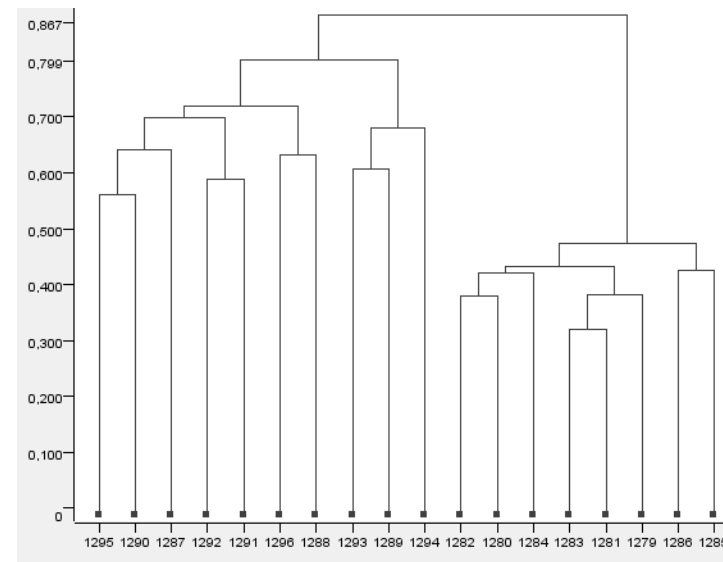
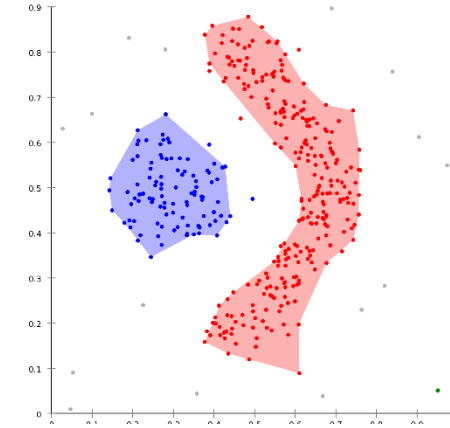
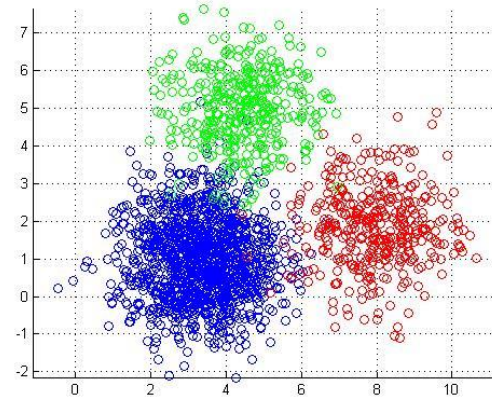
	Assigned topic	
<input type="checkbox"/>	topic_0	
<input type="checkbox"/>	topic_1	
<input type="checkbox"/>	topic_2	
<input type="checkbox"/>	topic_3	
<input type="checkbox"/>	topic_4	
<input type="checkbox"/>	topic_5	
<input type="checkbox"/>	topic_6	
<input checked="" type="checkbox"/>	topic_7	

Showing 1 to 8 of 8 ent

Visualizing Clusters

Methods:

- Hierarchical clustering
- K-Means / Medoids
- Density based
- t-SNE for embedding of high dimensional data
- ...



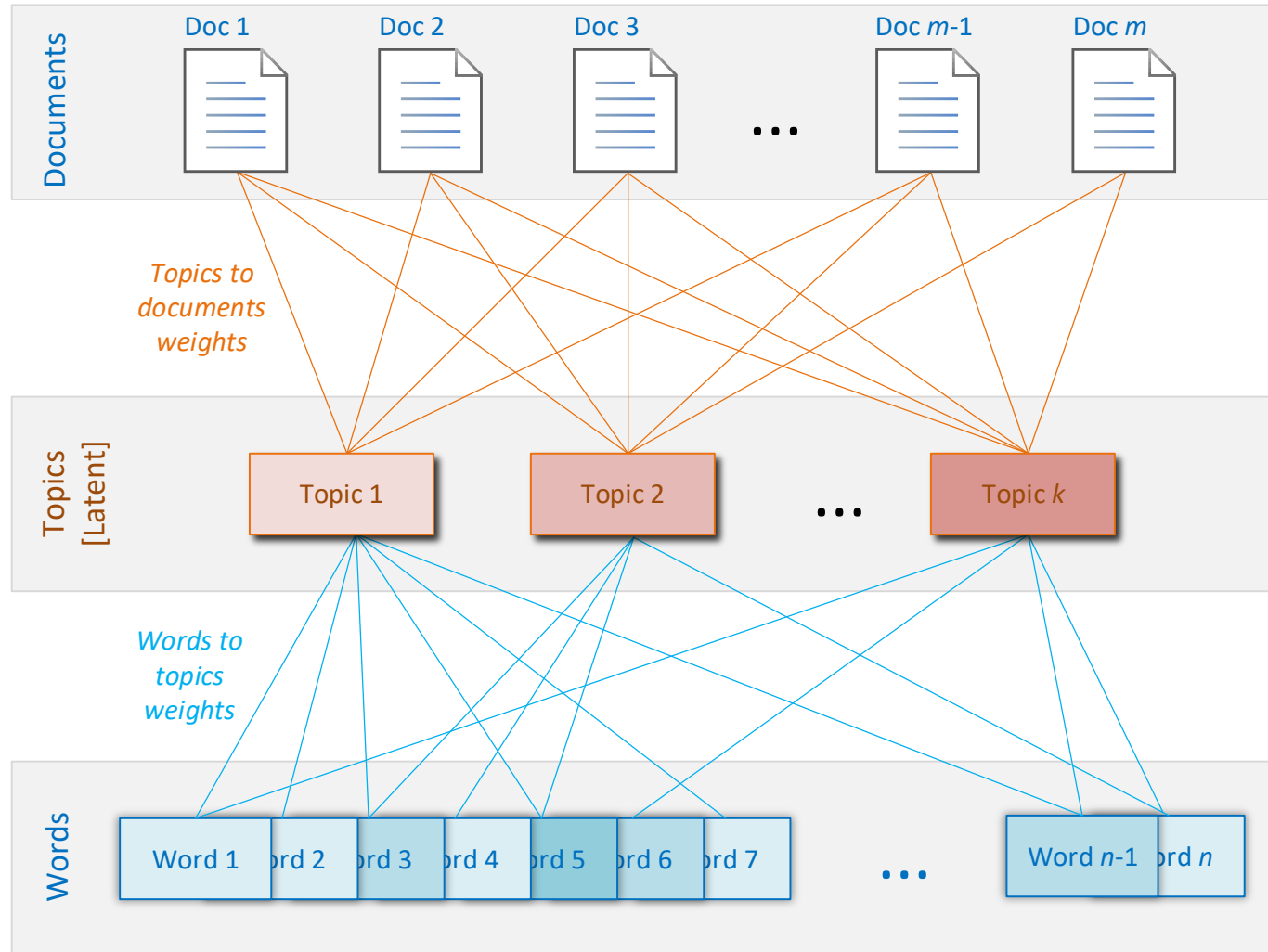
Topic Modeling in Text Mining

- The process of discovering (learning, identifying, extracting) topics across a collection of documents (corpus)
- Common assumptions for all topic modeling models:
 - Each document consists of a mix of topics
 - Each topic consists of a collection of words/terms
- Topics are “hidden” or “latent” constructs in between documents and words
- The goal of topic modeling is to discover these latent variables (i.e., topics) that shape the meaning/semantics in the document collection

Latent Dirichlet Allocation (LDA)

- LDA uses Dirichlet priors/distributions for the document-to-topic and topic-to-word associations/allocations
- LDA is a generative statistical model
 - It is an unsupervised learning process
 - Given a set of training data the goal is to identify the underlying distribution by generating samples from the same distribution
- Dirichlet distribution - $\text{Dir}(\alpha)$
 - It is a family of continuous multivariate probability distributions parameterized by a vector α of positive reals.
 - It is a multivariate generalization of the beta distribution
 - Hence, it is also called multivariate beta distribution (MBD)

Latent Dirichlet Allocation (LDA)



Latent Dirichlet Allocation (LDA) in KNIME

- Other Data Types
 - Text Processing
 - IO
 - Enrichment
 - Transformation
 - Preprocessing
 - Frequencies
 - Mining
 - Chi-Square Keyword Extractor
 - Keygraph Keyword Extractor
 - StanfordNLP Open Information Extra
 - StanfordNLP Relation Extractor
 - Topic Extractor (Parallel LDA)**
 - Misc

Document table with topics - 2:296 - Topic Extractor (Parallel LDA) (Extract topics from)

File Hilite Navigation View

Table "default" - Rows: 901 | Spec - Columns: 15 | Properties | Flow Variables

Row ID	S ABSTRACT	Document	D t...	D ...	D ...	D ...	D ...	D ...	D ...	D ...	D ...	D ...	S Assigned
Row0	The need for c...	"pid001"	0.377	0.001	0	0.618	0.001	0	0	0.001	0.001	0.001	topic_3
Row1	Although much ...	"pid002"	0.002	0.001	0.001	0.729	0.262	0.001	0.001	0.002	0.002	0.002	topic_3
Row2	When producer...	"pid003"	0.195	0.002	0.699	0.005	0.003	0.001	0.076	0.017	0.002	0.002	topic_2
Row3	Preservation of...	"pid004"	0.002	0.001	0.001	0.713	0.14	0.001	0.022	0.097	0.023	0.023	topic_3

Topic Extractor (Parallel LDA)



Node 78

Output

- Document table with topics
- Topic terms
- Iteration statistics

Topic terms - 2:296 - Topic Extractor (Parallel LDA) (Extract

File Hilite Navigation View

Table "default" - Rows: 90 | Spec - Columns: 3 | Properties | Flow Variables

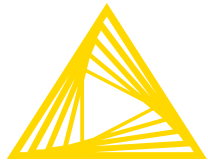
Row ID	S Topic id	S Term	D Weight
Row0	topic_0	firm	410
Row1	topic_0	investment	217
Row2	topic_0	value	193
Row3	topic_0	performance	140
Row4	topic_0	information	135
Row5	topic_0	market	132
Row6	topic_0	network	117
Row7	topic_0	technology	114
Row8	topic_0	benefit	109
Row9	topic_0	cost	105
Row10	topic_1	project	281
Row11	topic_1	software	185

Iteration statistics - 2:296 - Topic Extractor (Parallel LDA) (Extract

File Hilite Navigation View

Table "default" - Rows: 100 | Spec - Columns: 2 | Properties | Flow Variables

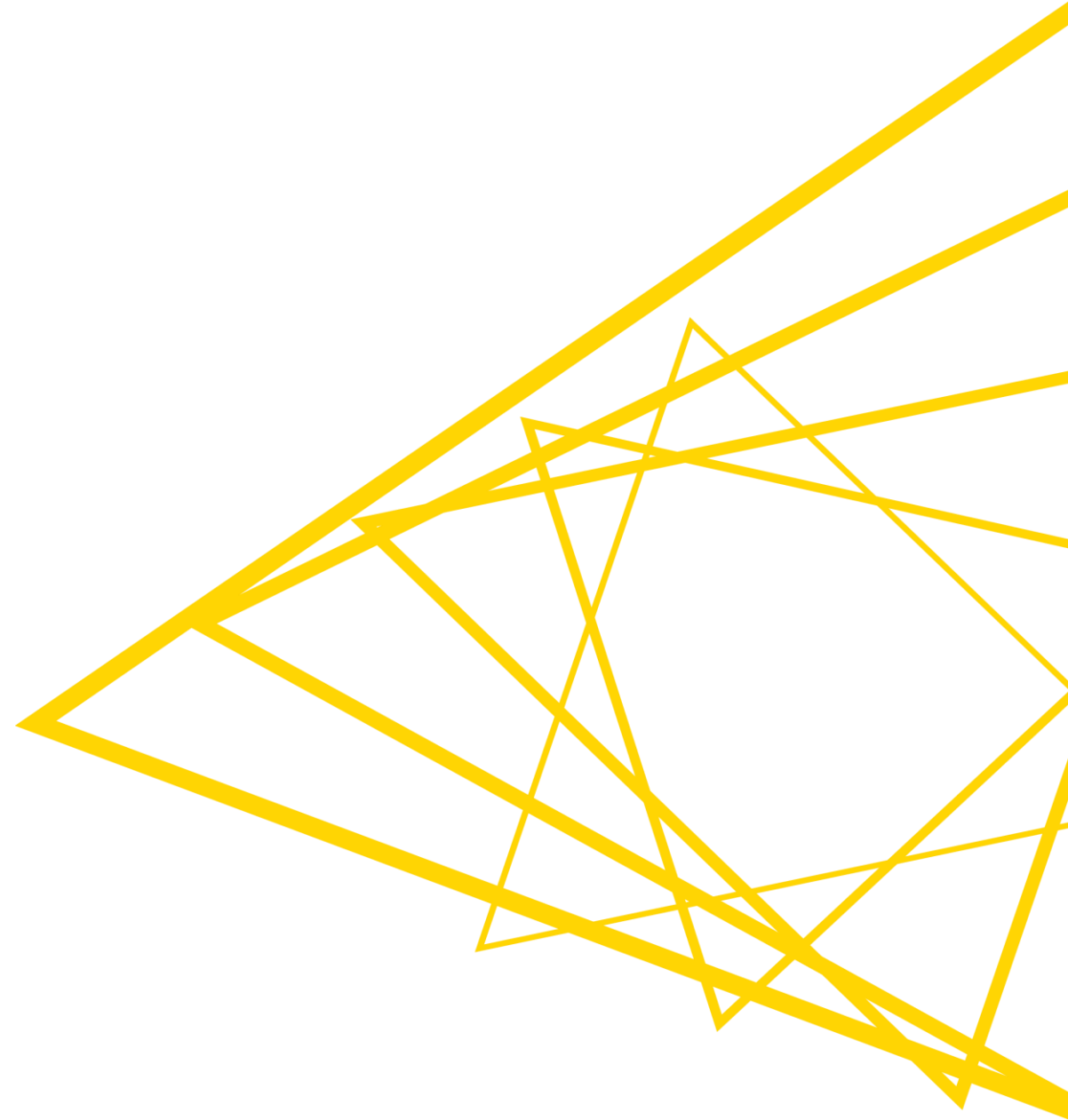
Row ID	I Iteration	D Log likel...
Row0	10	-8.722
Row1	20	-8.422
Row2	30	-8.314
Row3	40	-8.25
Row4	50	-8.211
Row5	60	-8.179
Row6	70	-8.166
Row7	80	-8.141
Row8	90	-8.121
Row9	100	-8.109



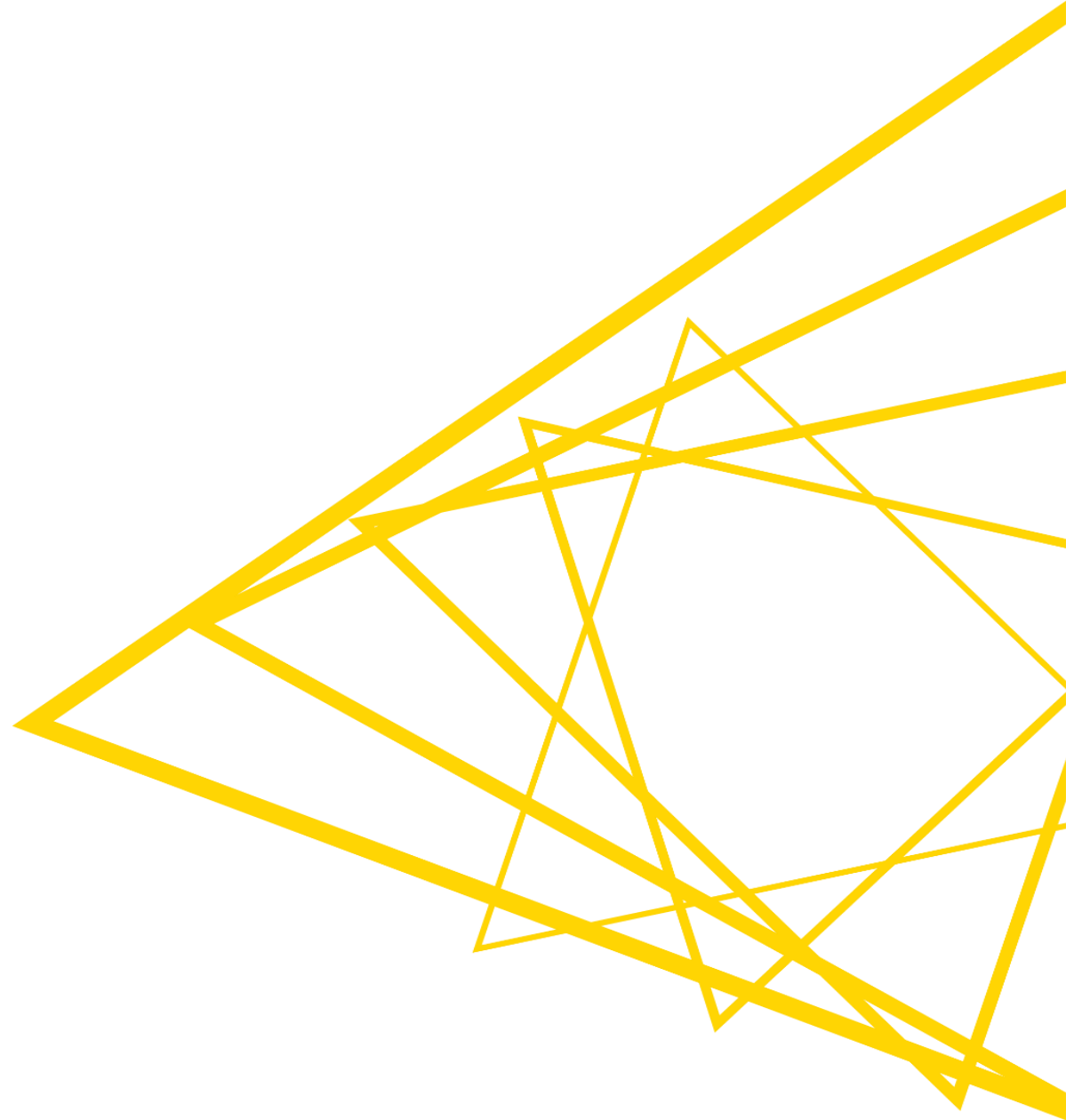
Open for Innovation

KNIME

LDA Demo



Twitter Sentiment Demo



From Words to Wisdom Book

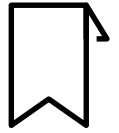
Free Copy of “From Words to Wisdom” Book from KNIME Press

<https://www.knime.com/knimepress>

with code: **INTRO-TP-1122**



Stay connected with KNIME



Blog: knime.com/blog



Forum:
forum.knime.com

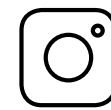


KNIME Hub:
hub.knime.com



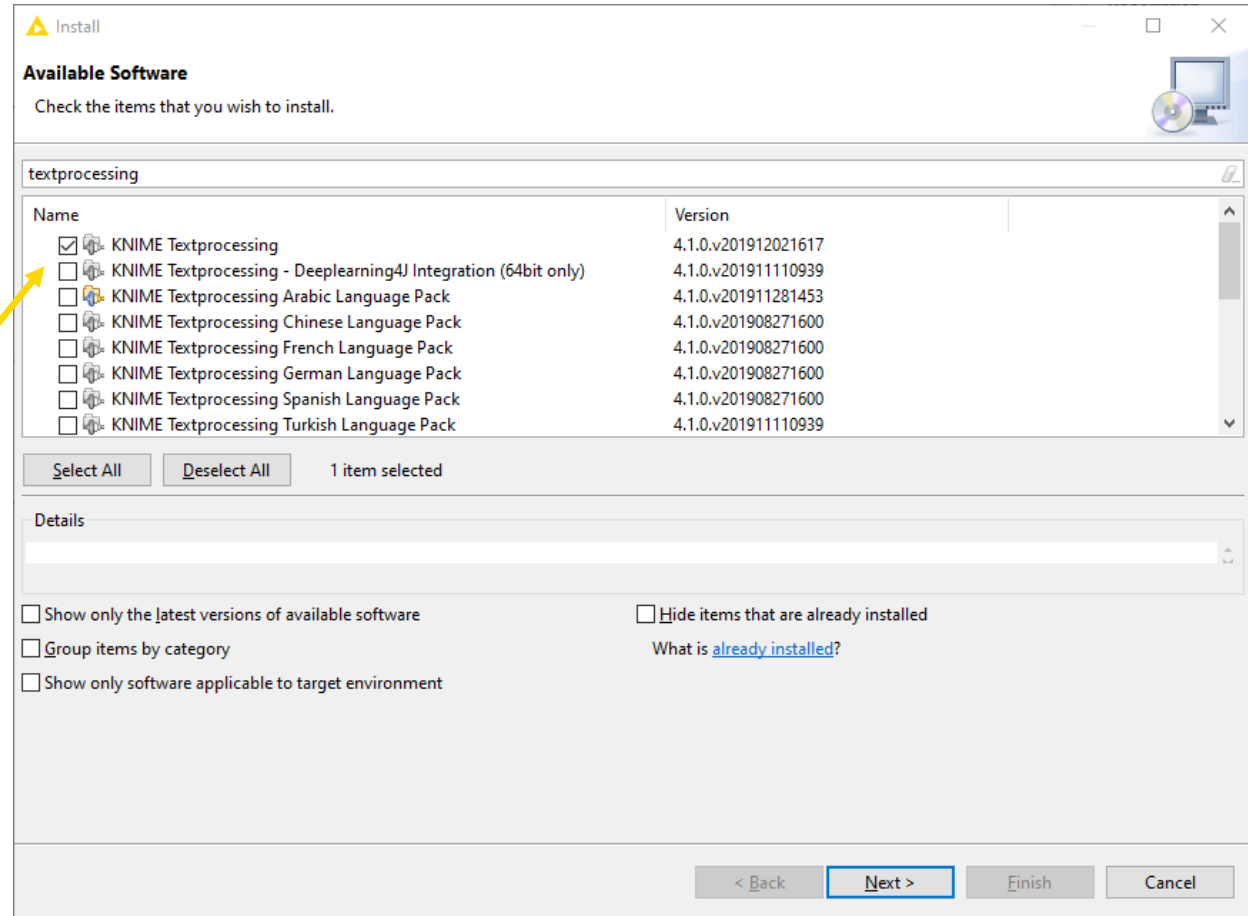
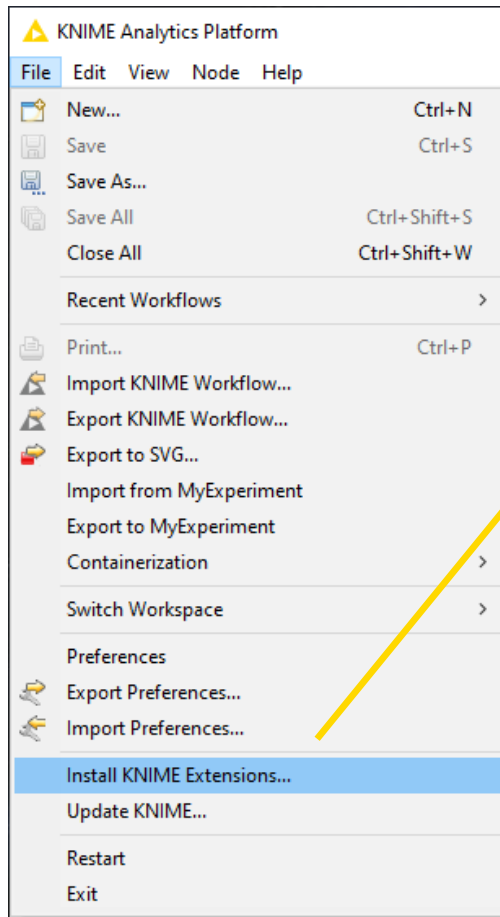
KNIME E-Learning Course:
www.knime.com/e-learning-course

Follow us on social media:



Installation

- Install KNIME Textprocessing Extension from menu...



Installation

- ...or drag-and-drop from KNIME Hub

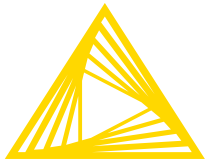
The screenshot shows the KNIME Hub interface for the 'KNIME Textprocessing' extension (v4.1.0). The page is titled 'KNIME Textprocessing' and includes a search bar at the top. Below the title, there are two tabs: 'Included nodes' and 'Related workflows'. The 'Included nodes' tab is active, displaying three nodes:

- Abner Tagger** (streamable): This node recognizes biomedical named entities, such as genes, proteins or cells and assigns tags to the corresponding ...
Other Data Types > Text Processing > Enrichment
Manipulator
- Bag Of Words Creator**: This node creates a bag of words (BoW) of a set of documents. A BoW consists of at least one column containing the term...
Other Data Types > Text Processing > Transformation
Manipulator
- Brat Document Writer**: This node takes the documents in the selected column and writes them, each as two files (.txt and .ann), into the selec...
Other Data Types > Text Processing > IO
Sink

On the right side of the page, there is a sidebar with the following sections:

- KNIME**
- Add to KNIME Analytics Platform**: Drag extension into the workbench of KNIME Analytics Platform 4.x
- Legal**: Copyright by KNIME AG, Zurich, Switzerland
License
- Short link**: https://kni.me/e/PH_ptBLLdL1Mich2

A large yellow arrow points to the 'Add to KNIME Analytics Platform' button.



Open for Innovation

KNIME

Thank you – Questions?

