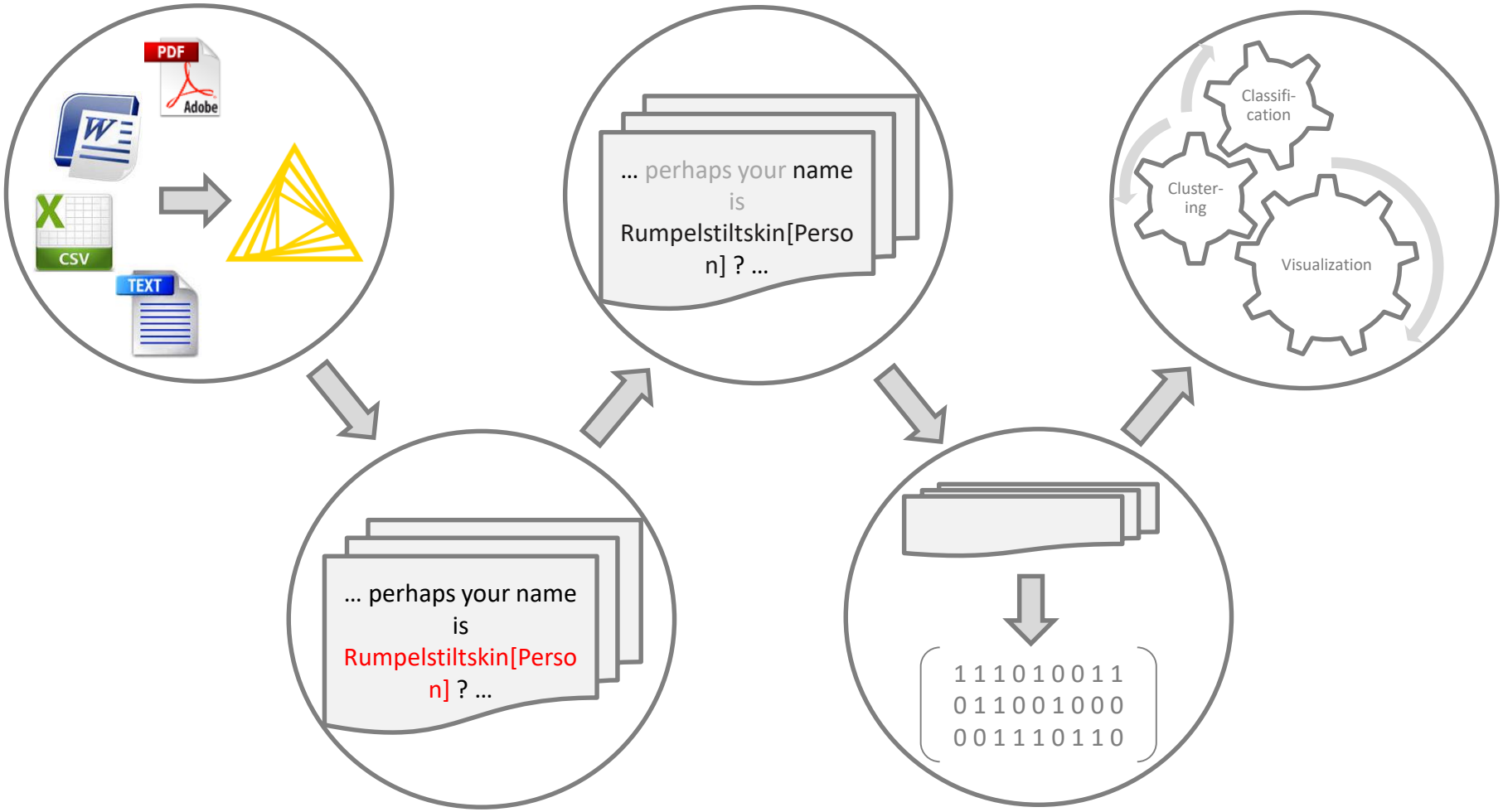
KNIME Text Mining Workshop

KNIME Spring Summit 2017
Kilian Thiel
KNIME.com AG

Agenda

- Textprocessing introduction
- News Recommender - a classification approach
- Topic Extraction using the elbow method
- Game of Thrones character interaction network

Textprocessing introduction



Textprocessing introduction

- Document Cell
  - Encapsulates a document
    - Title, sentences, terms, words
    - Authors, category, source
    - Generic meta data (key, value pairs)
- Term Cell
  - Encapsulates a term
    - Words, tags

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"

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)]

Textprocessing introduction

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

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

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

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

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

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

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

News Recommender - a classification approach

- Collect news content from RSS feeds.
- Take user input to label news item (interesting and not interesting).
- Build predictive model on labeled data.
- Apply model on unlabeled data.
- Webinterface to visualize news item recommendations (communicates via REST with the KNIME Server).

Topic Extraction using the elbow method

- Similar to clustering in topic extraction the initial. question about the number of topics/clusters in the data is not easy to answer.
- Topic Extractor node needs a number of topics (k) defined in the dialog.
- Idea: using the elbow method to find out the number of clusters k in the documents and use k as number of topics.

Game of Thrones character interaction network

- Visualization of character interaction as network.
- Tag character names in text.
- Compute co-occurrences in sentences.
- Create network with a node for each character and an edge for each co-occurrence.