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
... perhaps your name is Rumpelstiltskin[Person]?

111010011
011001000
001110110
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
Textprocessing introduction

- **Document table**
  - List of documents

- **Bag of words**
  - Tuples of documents and terms

- **Document vectors**
  - Numerical representations of documents
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.