KNIME User Group Meeting 2015
Text Mining Workshop

Kilian Thiel
Kilian.Thiel@knime.com
@KNIME
Agenda

• Introduction
• Sentiment Analysis
  – Single word features
  – Ngram features
  – Dictionary based
Resources

- **The KNIME Website** ([www.knime.org](http://www.knime.org))
  - **Textprocessing** ([tech.knime.org/knime-text-processing](http://tech.knime.org/knime-text-processing))
  - **Forum** ([tech.knime.org/forum/knime-textprocessing](http://tech.knime.org/forum/knime-textprocessing))

- **Blog** for news, tips and tricks ([www.knime.org/blog](http://www.knime.org/blog))

- **KNIME TV** channel on [YouTube](http://www.youtube.com)
- **KNIME** on [Twitter](http://www.twitter.com) [@KNIME](http://www.twitter.com/KNIME)
Introduction to KNIME Textprocessing
Installation

1.)

2.)
Tip

- Increase maximum memory for KNIME
- Edit knime.ini
  - Add “-Xmx3G” as last line of knime.ini file
  - Or change existing setting

- Useful additional extensions
  - Palladian (community extension)
    - Web crawling, Text Mining
  - XML-Processing (KNIME extension)
    - Parsing and processing of XML documents
Philosophy

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

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

111010011
011001000
001110110
Data Import

• Node Repository: KNIME Labs/Text Processing/IO

• Available Parser Nodes
  – Flat File Document Parser
  – PDF Parser
  – Word Parser
  – Document Grabber
  – ...

Text Processing
  ▼ IO
    ▼ Document Grabber
    ▼ Flat File Document Parser
    ▼ PDF Parser
    ▼ PubMed Document Parser
    ▼ Sdm1 Document Parser
    ▼ Word Parser
Enrichment

- Node Repository:
  KNIME Labs/Text Processing/Enrichment

- Available Tagger Nodes
  - Standford tagger
  - Dictionary (& Wildcard) tagger
  - OpenNLP tagger
  - Abner tagger
  - ...
Preprocessing

• Node Repository:
  KNIME Labs/Text Processing/Preprocessing

• Available Preprocessing Nodes
  – Stop word Filter
  – Snowball Stemmer
  – General Tag Filter
  – Case converter
  – RegEx Filter
  – ...

...
Transformation

• Node Repository:

KNIME Labs/Text Processing/Transformation

• Available Transformation Nodes
  – BoW Creator
  – Document Vector
  – Strings to Document
  – Sentence Extractor
  – Document Data Extractor
  – ...

Text Processing
  • IO
  • Enrichment
  • Transformation
    • BoW creator
    • Document Data Extractor
    • Document vector
    • Meta Info Extractor
    • Meta Info Inserter
    • Sentence Extractor
    • String to Term
    • Strings To Document
    • Tags to String
    • Term to String
    • Term to Structure
    • Term vector
Additional Data Types

• **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
Data Table Structures

- **Document table**
  - List of documents

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

- **Document vectors**
  - Numerical representations of documents
Import Export Workflows

[Image: A screenshot of the KNIME interface showing the 'Import' window with options to select archive files and browse for the workflow to import.]

[Image: A screenshot of the workflow import selection window with various directories and files listed, including 'IntroductoryCourseDay1.zip' and related items.]
Today’s example

- **Classification of free-text documents** is a common task in the field of text mining.

- It is used to categorize documents, i.e. assigning pre-defined topics or for **sentiment analysis**.

- Today we construct a workflow that reads and preprocesses text documents, transforms them into a numerical representation and **build a predictive model to assign pre-defined sentiment labels to documents**.
Today’s example

• Subset(s) of the Large Movie Review Dataset v1.0
• The LMR Dataset v1.0 contains 50,000 English movie reviews with their associated sentiment label “positive” or “negative”
Today’s example

Goal:

• Build **classifier** to distinguish between positive and negative movie reviews.

• „**Absolute masterpiece of a film! Goodnight Mr.Tom has swiftly become one of my favourite films of all time. Nobody should miss out on seeing this film, it's just too good! ...“**
Hands On

- Adjust Xmx setting in knime.ini
- Start KNIME
- Import Workflows
Sentiment Analysis
Classification (Recap)

Data Partitioning

- Original Data Set
- Training Set
- Test Set

Training and Applying Models

- Train Model
- Apply Model

Scoring Strategies

- Score Model
Classification (Recap)

- All data mining models use a **Learner-Predictor** motif.
- The Learner node trains the model with its input data.
- The Predictor node applies the model to a different subset of data.

![Diagram of Decision Tree Learner and Predictor nodes with Training and Test sets](image-url)
From Strings to Numbers
Hands On

• Read movie review data
• Convert to documents
• Preprocess documents
• Filter features based on frequency
• Transform documents into numerical representation
• Train and score predictive model
• View decision tree view
Negations

• What about negations?
  – Not good
  – Not bad
• Can not be handled with single word features
• Ngram features (1- and 2grams)
NGram Creator

• Allows for creation of
  – Character Ngrams
  – Word Ngrams

• As
  – Bag of words
  – With frequencies

<table>
<thead>
<tr>
<th>S</th>
<th>Ngram</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>special effects</td>
<td>318</td>
<td></td>
</tr>
<tr>
<td>that is</td>
<td>312</td>
<td></td>
</tr>
<tr>
<td>and it</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>that it</td>
<td>284</td>
<td></td>
</tr>
<tr>
<td>to watch</td>
<td>279</td>
<td></td>
</tr>
<tr>
<td>this one</td>
<td>279</td>
<td></td>
</tr>
<tr>
<td>the end</td>
<td>275</td>
<td></td>
</tr>
<tr>
<td>but the</td>
<td>274</td>
<td></td>
</tr>
<tr>
<td>the same</td>
<td>272</td>
<td></td>
</tr>
<tr>
<td>to make</td>
<td>272</td>
<td></td>
</tr>
</tbody>
</table>
Hands On

• Read movie review data
• Convert to documents
• Preprocess documents
• Create 1- and 2-gram features
• Filter features based on frequency
• Transform documents into numerical representation
• Train and score predictive model
• View decision tree view
**Dictionaries**

- Custom dictionaries for sentiment classification
  - Dictionary Tagger
  - Wildcard Tagger (regular expressions)
MPQA Subjectivity Lexicon

• Dictionary with words and related subjectivity clues
  – 2718 words with positive polarity
  – 4901 words with negative polarity

• http://mpqa.cs.pitt.edu/lexicons/subj_lexicon
Dictionary Based
Hands On

- Read movie review data
- Convert to documents
- Tag documents
- Create pivot features
- Specify score and rule for label assignment
- Score prediction
Appendix

• Palladian and XML nodes to extract move titles from IMDB web pages
• First dataset contains IMDB URLs of related movies

<table>
<thead>
<tr>
<th>Index</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>3617</td>
<td><a href="http://www.imdb.com/tt0210075/usercomments">http://www.imdb.com/tt0210075/usercomments</a></td>
</tr>
<tr>
<td>3671</td>
<td><a href="http://www.imdb.com/tt0337640/usercomments">http://www.imdb.com/tt0337640/usercomments</a></td>
</tr>
<tr>
<td>3157</td>
<td><a href="http://www.imdb.com/tt0303549/usercomments">http://www.imdb.com/tt0303549/usercomments</a></td>
</tr>
<tr>
<td>660</td>
<td><a href="http://www.imdb.com/tt0716825/usercomments">http://www.imdb.com/tt0716825/usercomments</a></td>
</tr>
<tr>
<td>265</td>
<td><a href="http://www.imdb.com/tt0182225/usercomments">http://www.imdb.com/tt0182225/usercomments</a></td>
</tr>
<tr>
<td>4027</td>
<td><a href="http://www.imdb.com/tt0347779/usercomments">http://www.imdb.com/tt0347779/usercomments</a></td>
</tr>
<tr>
<td>5820</td>
<td><a href="http://www.imdb.com/tt0095655/usercomments">http://www.imdb.com/tt0095655/usercomments</a></td>
</tr>
<tr>
<td>1574</td>
<td><a href="http://www.imdb.com/tt0298131/usercomments">http://www.imdb.com/tt0298131/usercomments</a></td>
</tr>
<tr>
<td>10668</td>
<td><a href="http://www.imdb.com/tt0088915/usercomments">http://www.imdb.com/tt0088915/usercomments</a></td>
</tr>
<tr>
<td>1473</td>
<td><a href="http://www.imdb.com/tt0828154/usercomments">http://www.imdb.com/tt0828154/usercomments</a></td>
</tr>
<tr>
<td>8337</td>
<td><a href="http://www.imdb.com/tt0110099/usercomments">http://www.imdb.com/tt0110099/usercomments</a></td>
</tr>
<tr>
<td>11217</td>
<td><a href="http://www.imdb.com/tt0098492/usercomments">http://www.imdb.com/tt0098492/usercomments</a></td>
</tr>
</tbody>
</table>
Web Data Extraction

File Reader
Read IMDb reviews from CSV file

Row Filter
First 50 only

HttpRetriever
Download website

HtmlParser
Node 292

XPath
Title extraction

String Manipulation
Cut off last 26 characters
Hands On

• Read movie review data
• Load webpage content
• Parse HTML
• Extract title field
The End

Kilian Thiel
Kilian.Thiel@knime.com
@KNIME