# KNIME Newsletter

October 2011 - Vol. 1, No. 4

## Welcome to the Fourth KNIME Newsletter!

While the KNIME team is working hard on putting the finishing touches on KNIME 2.5 (with many cool new features), we are happy to present the latest KNIME Newsletter. This time we have an article from one of our KNIME partners who is using KNIME to perform customer sentiment analysis by applying (among others) the KNIME text analytics plugin. We can't wait to see what they will come up with once the network mining extensions are available, too. In addition, we report on our first KNIME Open Source Days—quite honestly, the energy and enthusiasm of the KNIME Community surprised even us! And last, but not least, make sure to block February 1+2, 2012 in your calendar for our Annual User Group Meeting in Zurich.

We look forward to seeing you there! The KNIME Team

KNIME news

**Mark Your Calendars!**
**The 2012 KNIME User Group Meeting**

Over 130 attendees enjoyed our 2011 UGM earlier this year — we are planning a similarly exciting program again on

**February, 1&2, 2012**
in **Zurich, Switzerland.**

Don't miss it!

As usual, there will also be accompanying training courses and workshops. For details see: www.knime.com/UGM2012

## "Social Media Intelligence" with KNIME

Social Media is a hot topic But the actual task of monitoring and understanding what customers are saying about an organization via social media touch points has not been easy until now. Dymatrix, a KNIME partner specializing in all forms of Customer Intelligence and with a very solid reference pedigree, has used KNIME to develop the first application capable of capturing not only a wide variety of social media touch point data, but also of combining that data with an organization's own customer intelligence.

**"...actually monitoring and understanding what customers are saying about an organization via social media touch points has not been an easy task until now."**

"When we talk about analyzing social media data, we are primarily talking about capturing and mining text data from a number of different sources and transforming it into something intelligible. The challenge is threefold and involves retrieving the social media data, mining it for textual insights, then surfacing it in a form suitable for marketing and management," says Stefan Weingärtner, management consultant with Dymatrix. Dymatrix utilized its own expertise to design social media engines that can pull all relevant information from social media networks such as Twitter. "KNIME is the perfect data mining platform for text mining, easily reading all the data we have collected. The KNIME text mining nodes are extremely powerful and – unlike other tools specifically for social media monitoring – the nodes can be customized to cover languages other than English."

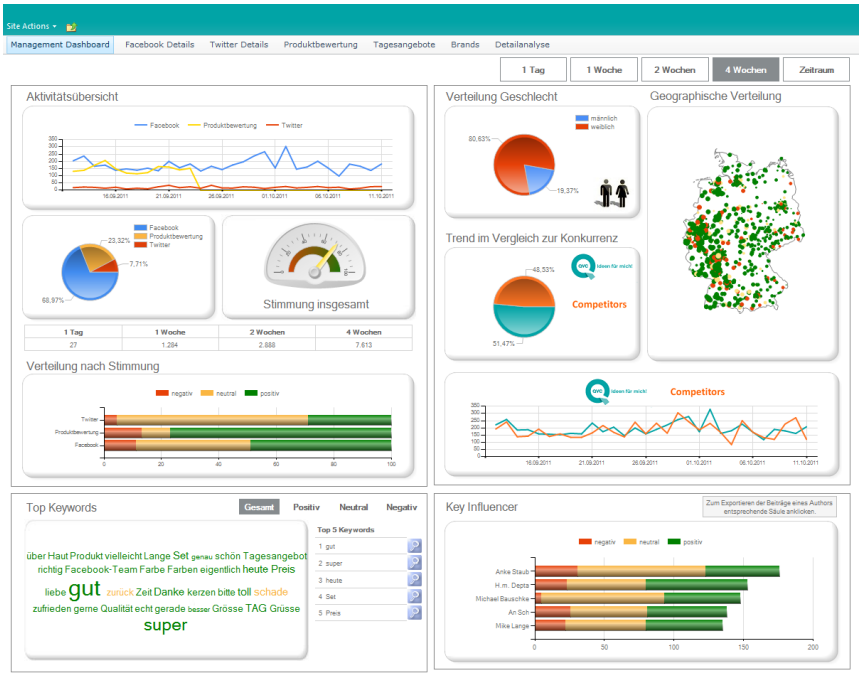
Understanding social media text mining means answering such questions as: which touch points are being used? By whom? Who is listening to whom? Who is talking positively (or negatively) about our products or services? What is the tone of the conversations? How do we compare in these conversations to our competition or others we wish to benchmark against? "Most importantly, using the KNIME platform means that we can combine social media data with other existing customer intelligence within an organization to develop truly new insights, which can be surfaced in multiple ways, depending on the target audience," says Weingärtner.

The DynaSocian application from Dymatrix packages the social media engines together with a series of KNIME workflows. In addition, the application has a full front end giving non-data mining experts the ability to not only set up, but to run and – most importantly – visualize information in a special front end. "A good example of DynaSocian is demonstrated here by QVC, the extremely successful television shopping organization. By using the software, QVC was able to get a good idea of the tone of, but also of the items being promoted. In this case, we did the example in German."

**Stefan Weingärtner**
Executive Board
Dymatrix Consulting

says Weingärtner.

KNIME is being used more and more as the data mining and analytics platform of choice. And KNIME does not stop at text mining for social media. In KNIME version 2.5, scheduled for release in early December, additional nodes have been added to conduct network analysis. You can refer to the www.KNIME.com website to see some examples of KNIME text mining. For help reading social media data, or for further information on the KNIME-based DynaSocian application, please contact Stefan at www.Dymatrix.de.




## New SAS to KNIME Guide

The newest addition to KNIME Press, a guide to KNIME for SAS™ users is now freely available:

www.knime.org/knimepress

Note that with KNIME2.5 powerful text processing functionality will be added to KNIME which is not yet covered in this guide—look out for an update in early 2012.

## The 1st KNIME Open Source Days

**Thorsten Meinl**
KNIME Team

The KNIME Open Source Days were hosted for the first time at the beginning of October in Konstanz. The idea of bringing together all of the community developers, who provide open-source extensions for KNIME, was born when we saw the number of projects in the community contributions quickly increasing after the first community release in December 2010. During the week around 35 developers from all over Europe met at the beautiful Villa Rheinburg on Lake Constance to present their projects and discuss issues with the core KNIME developers. A lot of fervent hacking also took place, further extending project functionality. "I especially liked solving problems together with people I had never met so far, merging our know-how to produce successful results," says Manuel Schwarze from Novartis, who is working on the RDKit extensions for KNIME.

The many projects in the community contributions already include a wide range of application areas such as image processing, chemoinformatics, next generation sequencing, and scripting extensions to name only a few (see below for a complete list). Other projects such as the Palladian toolkit for information retrieval from the University of Dresden, the BALL library for chemo- and bioinformatics from the University of Tübingen, or the new CDK integration by EBI in Cambridge made major steps towards inclusion in the next stable release. "Especially the mixture of presentations from the various groups during the week was beneficial for finding new cooperations," says Klemens Muthman from the Palladian project.

The KNIME developers themselves gave an outlook on the upcoming KNIME 2.5 release and presented a testing framework for KNIME nodes, which is now also available to community developers. The entire KNIME Team felt that the KNIME Open Source Days were a great success and look forward to organize a similar event next year!

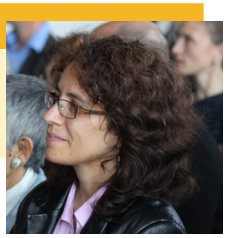
The community contributions are available at http://tech.knime.org/community and can be installed easily via an online update site. They currently include the following projects:

- RDKit - an open-source chemoinformatics toolkit which includes a collection of standard chemoinformatics functionality for molecule I/O, substructure searching, chemical reactions, coordinate generation (2D or 3D), fingerprinting, etc.
- Indigo - integration of the Indigo toolkit allowing users to create high-performance workflows for completing standard chemoinformatics tasks.
- Erl Wood Chemoinformatics - a huge collection of nodes generally geared towards pharmaceutical research with a focus on SAR data manipulation, interpretation, and viewing.
- CDK - integration of the well-known Java library for structural chemo– and bioinformatics.
- HCS Tools - contains readers for various microscopes, quality controls metrics, all common plate normalization methods, a very powerful plate heatmap viewer, library annotation tools, as well as barcode and plate layout utilities.
- Scripting integrations - a scripting integration framework for KNIME, which is based on RGG templates, its main purpose being to hide the script complexity behind a user-friendly graphical interface.
- NGS tools - a collection of example flows and utility nodes for processing next generation sequencing results.
- Image Processing - adds new image types to KNIME and the corresponding nodes to read more than 100 different kinds of images, to apply well known methods for preprocessing, and to perform image segmentation.
- Palladian - integration of the Palladian toolkit which provides the functionality to perform Internet Information Retrieval tasks such as crawling, classification, and extraction of various types of information.

## Tips & Tricks

### The Chi-Square Test of Independence and the Crosstab Node.

The "Crosstab" node has been introduced with KNIME 2.4 and can be found in the "Statistics" category. It performs a number of different tasks and calculates a number of statistically interesting variables opening the door to using KNIME for statistical tests.



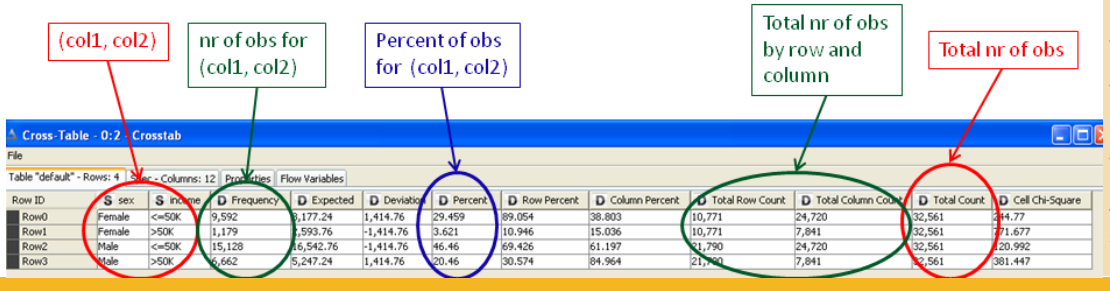
**Rosaria Silipo**
Data Mining Consultant
Zurich, Switzerland

#### The Chi-Square Test of Independence

The Chi-square test of independence assumes as a null hypothesis that two categorical variables (col1 and col2) are independent. Such a hypothesis of total independence expects random, equally distributed frequencies, *E(i,j)*, for each i and j value pair. The Chi -Square variable, which calculates how far the expected frequencies are from the observed frequencies, has an approximate chi-square distribution with *DF = (r-1) x (c-1)* degrees of freedom.

The "Crosstab" node selects two of the input data columns and produces two output tables. The cross table contains the contingency table with the expected and observed frequencies, while the statistics table contains the Chi-Square related values for this contingency table. In particular, in the statistics table a *"Chi Square (prop)"* cell contains the probability P(x >= X), where X is the value of the Chi-Square variable for this particular contingency table. If we fix an acceptance threshold at 5% (0.05), if P(x>=X) lies below this threshold, the null hypothesis of statistical independence between the two variables col1 and col2 can be rejected.

More details on the crosstab node are available on my blog:
http://dataminingreporting.weebly.com/blog.html.




#### The Contingency Table

The "Crosstab" node builds a contingency table on two selected input data columns, col1 and col2. A contingency table is an r x c matrix of observations, where r is the number of distinct values in col1 and c is the number of distinct values in col2. Each cell reports the number of observations for each pair of values from col1 and col2.

The contingency table is made available at the outport "cross table". The node also offers a view ("View Cross Tabulation" option in its context menu) in which all observations and totals are organized in a table representation.