KNIME & Teacher Bots:
From Workflows to Micro Services

Kathrin Melcher
Vincenzo Tursi
Rosaria Silipo
Phil Winters
The History of Bots

1950
A.L.I.C.E.
Artificial Linguistic Internet Computer Entity.

Alan Turing
Bots

“A bot is software designed to automate the kinds of tasks you would usually do on your own (or another human would do for you).”

Search Bots
Teaching Bots
Communication Bots
Personal Assistant Bots
Data & Developer Bots
Team Bots
The Human Internet Search Process

1. Ask Question
2. Translate to Keyword(s)
3. Index

- Keywords
- Categories

- "Best" Question Answer?
  - Yes
  - Not Yet
  - None
The Search Challenge: Context

A Jewish Holiday

Paul Shapiro
Professional Search marketer
(and huge KNIME fan…)

Optimizing for Hanukkah: Sometimes it’s still strings, not things

Hannukah
Chanukah
Hanukah
Channukah
Chanuka
Chanukkah
Hanukka
Chanukka
Hannukkah
Hanukkah
Channuka
Festival of Lights
Feast of Dedication
A Data Science Project

Data Preparation

Training Model

Apply Model

Scoring

classes

Training set

Test set

1 2 1 0 4 2 1 0
2 3 1 5 6 2 0 0
6 2 0 0 2 3 1 1
1 5 6 2 3 3 0 3

1 2 1 0 4 2 1 0
2 3 1 5 6 2 0 0
6 2 0 0 2 3 1 1
1 5 6 2 3 3 0 3
Reality Check

We need a class ontology and class labels.
Ontology Definition

• **Ontology** is the *philosophical study* of the nature of *being*, *becoming*, *existence*, or *reality*, as well as the basic *categories of being* and their relations

• Introduced by Greek philosophers (Parmenides)

• **Parmenides** was among the first to propose an ontological characterization of the fundamental nature of reality

• In *computer science* and *information science*, an *ontology* is a formal naming and definition of the types, properties, and interrelationships of *entities*
Ontology Example

Relationship of major animal lineages with indication of how long ago these animals shared a common ancestor. On the left, important organs are shown, which allows us to determine how long ago these may have evolved.
The Specialist Topic Search Process

Ask Question

- Translate to Keyword(s)
  - Not Yet
  - "Best" Question Answer?
    - Yes
    - None

Keywords

Special Ontology

- Keywords
- Categories

Medicine
  - symptoms
  - diseases
  - treatments

Pharmaceutical
  - drugs
  - dosages
  - allergies
Creating an Ontology: Simple! Build Context

1. Ask Question
   - Translate to Keyword(s)
     - Not Yet
       - "Best" Question Answer?
         - Yes
         - None

2. Keywords
3. Categories

Ontology
A Real World Ontology Need: I want to learn KNIME

• I have a question…..
  – Terms
  – Concepts
  – Context
  – Background
  – Depth
  – Breadth
  – Language
Our Own Ontology (20 Classes)

From e-Learning Course
- Installation
- Data Access
- ETL
- Mining
- Control
- Deployment
- DataViz

From other Resources
- Use Cases
- Text Processing
- Big Data
- Server
- Image Processing
- Reporting

From Experience
- Development
- Integration
- Optimizing KNIME
- Life Science
- Announcement
- Bug
- Legal
Active Learning Cycle

1st attempt Class Labels → Model Training → Extract most uncertain predictions → Re-labeling → Class Label Extension

Training Set [Forum Questions]
The Human “Learn KNIME” Process

1. Ask Question

2. Translate to Keyword(s)

3. Index

Keywords

Categories

“Best” Question Answer?

Not Yet

None

Yes
Teacher Bot “Emil”: A Bot to help Learn KNIME

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

“Best” Question Answer?

Yes

Email

Keywords

Categories

KNIME Ontology Model
Teacher Bot “Emil”

Emil, our Teacher Bot!

Question

Hello there! I'm Emil, what can I help you find today?

My Summary:
trouble in connecting to sql server

My Detailed Question:
Hi am using database node reader to connect to sql server data base. I have loaded the sql driver and configures the node but i got the following error: The JDB protocol stream is not valid. what should i do to solve this error?any help would be appreciated
Teacher Bot “Emil”

HI!

Based in your question

"hil am using database node reader to connect to sql server data base. I have loaded the jdbc driver and configures the node but i got the following error: The TDS protocol stream is not valid. what should i do to solve this error? any help would be appreciated"

select the category that you want to inspect and click "Next".

<table>
<thead>
<tr>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Access</td>
</tr>
<tr>
<td>ETL</td>
</tr>
<tr>
<td>DataViz</td>
</tr>
<tr>
<td>Something Else</td>
</tr>
</tbody>
</table>

Showing 1 to 4 of 4 entries
Here are some links based on your question:

- [https://www.knime.com/blog/GoogleBigQuery_meets_SQLite](https://www.knime.com/blog/GoogleBigQuery_meets_SQLite)

Question answered? Then click "Next". Otherwise click "Back" to choose another category.
Teacher Bot “Emil”

Will They Blend? Experiments in Data & Tool Blending. Today: Blending Databases. A Database Jam Session

In this blog series we’ll be experimenting with the most interesting blends of data and tools. Whether it’s mixing traditional sources with modern data lakes, open-source devops on the cloud with protected Internal legacy tools, SQL with noSQL, web-wisdom-of-the-crowd with in-house handwritten notes, or IoT sensor data with idle chatting, we’re curious to find out: will they blend? Want to find out what happens when IBM Watson meets Google News, Hadoop Hive meets Excel, R meets Python, or MS Word meets MongoDB?

Follow us here and send us your ideas for the next data blending challenge you’d like to see at willtheyblend@knime.com.

Today: Blending Databases. A Database Jam Session

The Challenge

Today we will push the limits by attempting to blend data from not just 2 or 3, but 6 databases!

These 6 SQL and noSQL databases are among the top 10 most used databases, as listed in most database comparative web sites (see DB-Engines Ranking, The 10 most popular DB Engines, Top 5 best databases). Whatever database you are using in your current data science project, there is a very high probability that it will be in our list today. So, have nothing!
Teacher Bot “Emil”: A Bot to help Learn KNIME

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

“Best” Question Answer?

None

Yes

Email

Keywords

Categories

KNIME Ontology Model
Teacher Bot “Emil”

Goodbye!

It looks like I have solved your problem. Till next time.

Cheers,

Emil and the KNIME Team
Teacher Bot “Emil”
Creation of an Initial Ontology

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

“Best” Question Answer?

Email

None

No

KNIME Ontology Model

Keywords

Categories

Model Training

Training Set

Initial Labelling
Web Crawling of the KNIME Resources

This workflow reads in the URLs of the different resources available on the KNIME website and extracts the content into a document cell. In case of the node guide only the URL of the first page is given. The metanode extracts all links on the webpage.

Only three nodes
Step 0 – Initial Labeling

Labeling a Training Set based on Distance (and no Clue)
Step 0 – Initial Labeling

Active Learning 0: Class Labels based on Distance Matching between Resource Documents and Forum Questions
- Extract keywords from Resource Documents.
- Extract keywords from Forum questions.
- Run a similarity search between keywords vectors from resources and forum questions and find the minimum distance association.
- Assign class to question based on minimum distance association.

Labeling a Training Set based on Distance (and no Clue)
Step 1 - Model Training

Train Document Vector and Random Forest Model

This workflow trains two models. A Document Vector Model depending on the keywords of the training set for the PreProcessing and a Random Forest model to make the prediction of the document_class.
Creation of an Initial Ontology

1. Ask Question
2. Teaching bot
   - Translate to Keyword(s)
   - Not Yet
   - "Best" Question Answer?
   - Yes
   - Email
   - None

3. Keywords
4. Categories

5. KNIME Ontology Model
6. Model Training
7. Training Set
8. Initial Labelling
Adding Active Learning to the Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

"Best" Question Answer?

Yes

Not Yet

None

Email

Keywords

 KNIME Ontology Model

Model Training

Training Set

Initial Labelling

Active Learning Cycle
Active Learning

Random Forest

10% most uncertain classes → Diff. between three top probabilities for each predicted class

Labeling → Predicted Classes or “Something Else”

Labeling manually all “Something Else”

Active Learning Cycle

Model Training

Training Set

Initial Labelling

Based on Distance

k-NN (k=1)

Category Assign

Category Define

Labeling

Subset chosen to be labeled

Extend
Adding Active Learning to the Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

"Best" Question Answer?

Not Yet

None

Not Yet

Email

Keywords

KNIME Ontology Model

Model Training

Initial Labelling

Training Set

Categories

10% most uncertain

Active Learning Cycle

Category Assign

None

Yes
Hi Specialist!

I was wondering if there would be any "gotchas" in trying to use python multiprocessing within a KNIME Python Script node. I have a stand-alone script that uses multiprocessing and I'd like to adapt for use within KNIME. Simply placing the code within a Python Script Node fails with File "C:\Python27\lib\multiprocessing\pool.py", line 558, in get raise TypeError: expected string or Unicode object, NoneType found. One would normally protect the body of the script that initializes the data and sets up the multiprocessing pool with if name_ == "__main__" then change this tool. __name_ == "__builtin__" which seems to be the correct module name. Any ideas? Steve.

needs a category. Please choose one.

- Integrations
  - ET
  - Data Access
  - Smartening Line

Thank you!
Step 2a - Category Assign

This workflow is part of a number of other workflows that address a data mining scenario at the intersection of active learning, text mining, stream mining and service-oriented knowledge discovery architectures.

This workflow, in particular, provides a graphical interface on the WebPortal for a KNIME specialist to respond to the question with the most uncertain predicted class. It starts by first reading a subset of the training set (10% of the most uncertain predicted classes). Then, it loops over all the questions, and for each one, those, it allows the specialist to choose between one of the predicted classes or the option “Something Else”. The labeling phase takes place in the “Choose Answer” webpage. To complete the execution of the loop the specialist has to complete the labeling for all the no-processed questions of the rest of the training set, or to click “Exit”. If the specialist clicks “Exit” the workflow saves the last step of the loop iteration. Thus, when the specialist starts again the execution of the workflow on the WebPortal, he/she will be able to start labeling the questions from the last loop iteration.

The Variable Condition Loop End data split between questions that have been labeled as “Something Else” and all the other categories. These two datasets are then saved into two different tables.
Adding Active Learning to the Cycle

1. **Ask Question**
   - Teaching bot
     - Translate to Keyword(s)
     - Not Yet
     - "Best Question Answer?"
   - Email
     - Yes
     - None
   - Keywords
     - Categories

2. **KNIME Ontology Model**
   - Model Training
   - Training Set
     - 10% most uncertain

3. **Initial Labelling**
   - Active Learning Cycle
     - Category Assign
     - Category Define
Adding Label
Step 2b - Category Define

TA - System - Manual Labeling of the Questions Classified as "Something Else"

This workflow is part of a number of other workflows that address a data mining scenario at the intersection of active learning, text mining, stream mining and service-oriented knowledge discovery architectures. In particular, this workflow, first reads the questions labeled as "Something Else" in the workflow Relabel_Uncertain_Classes. Then, from the WebPortal, the KNIME specialist can label those questions with one of the possible classes defined in the ontology. Finally, the workflow updates the Training Set with the newly manually labeled questions.

Reading Data

Labeling Questions Manually

Writing the Results back to the Training Set with the Updated Labels

Reading Data

Questions classified as "Something Else" & Training Set

Labeling

Concatenate

Reference Row Filter

Table Writer

Update Training Set
Adding Active Learning to the Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

"Best" Question Answer?

None

Yes

Email

Keywords

KNIME Ontology Model

Model Training

Training Set

Initial Labelling

Active Learning Cycle

Category Assign

Category Define

Extend

10% lowest probability
Step 3 – Extend with k-NN

• Expert has labelled uncertain samples

• k-NN (k=1) extends the expert classes to their neighbor sample
Step 3 - Extend with k-NN

Active Learning 5. Extend Oracle Labels to Training Set with k-NN.
Once we have a subset labelled by the oracle, we need to extend these classes to the whole training set. We do that here with a k Nearest Neighbour.
Adding Active Learning to the Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

“Best” Question Answer?

None

Yes

Email

Keywords

Categories

KNIME Ontology Model

Model Training

Training Set

Initial Labelling

Active Learning Cycle

Category Assign

Category Define

Extend

10% lowest probability
Combining the Teaching Bot and the Active Learning Cycle

1. **Ask Question**
   - Teaching bot
     - Translate to Keyword(s)
     - Not Yet
     - "Best" Question Answer?
     - None
     - Yes
     - Email

2. **Keywords**
   - KNIME Ontology Model
   - Categories

3. **Model Training**
   - Training Set
     - Initial Labelling
     - 10% lowest probability
     - Extend

4. **Active Learning Cycle**
   - Category Assign
   - Category Define
Changes in Training Set

AL Iteration 0

AL Iteration 1

AL Iteration 2
## Answer Evolution

<table>
<thead>
<tr>
<th>AL #</th>
<th>Input Dataset</th>
<th>Output Model</th>
<th>Ver</th>
<th>Accuracy</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Training_Set_v0</td>
<td>RandomForest_v0</td>
<td>0.0</td>
<td>0.59</td>
<td>19/2/2018</td>
</tr>
<tr>
<td>1</td>
<td>Training_Set_v1</td>
<td>RandomForest_v1</td>
<td>1.0</td>
<td>0.56</td>
<td>23/2/2018</td>
</tr>
<tr>
<td>2</td>
<td>Training_Set_v2</td>
<td>RandomForest_v2</td>
<td>2.0</td>
<td>0.52</td>
<td>26/2/2018</td>
</tr>
</tbody>
</table>

![Version 0](image1.png)

![Version 2](image2.png)
Combining the Teaching Bot and the Active Learning Cycle

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

“Best” Question Answer?

None

Yes

Email

Keywords

Categories

KNIME Ontology Model

Model Training

Training Set

Initial Labelling

Active Learning Cycle

10% lowest probability

Category Assign

Category Define

Extend
From building one time to reusing Components: MicroServices

Ask Question

Teaching bot

Translate to Keyword(s)

Not Yet

"Best" Question Answer?

None

Yes

Email

Keywords

KNIME Ontology Model

Categories

Model Training

Training Set

10% lowest probability

Active Learning Cycle

Extend

Initial Labelling

Category Assign

Category Define
Microservices

TA-System
  01_DataAccess
  02_ActiveLearning
  03_Deployment
  Archive
  data

MetanodeTemplates
  Extract Classes
  Get_Probabilities_v0
  Get_Probabilities_v1
  Get_Probabilities_v2

Microservices
  Extract Classes
  Get_Class_Probabilities_v0
  Get_Class_Probabilities_v1
  Get_Class_Probabilities_v2
  Get_Probabilities_v0
  Get_Probabilities_v1
  Get_Probabilities_v2
Microservices - Converting reusable Subflows into Microservices

Metanode Templates

Microservices
Combining the Teaching Bot and the Active Learning Cycle

1. **Teaching Bot**
   - Translate to Keyword(s)
   - Email
   - "Best" Question Answer?
     - Not Yet
     - None
     - Yes

2. **Active Learning Cycle**
   - Category Assign
   - Category Define
   - Extend

3. **Initial Labelling**

4. **KNIME Ontology Model**
   - Model Training
   - Training Set
   - 10% lowest probability
What we have tried to show....

- Creating a basic bot
- Building an Ontology with Active Learning
- Automating the process
- Converting reusable subflows into micro services
What did we learn?

- KNIME forum is used as educational tool
- Support is search
- Keyword extraction is a plus with respect to just keyword search
- Re-adjust your class system (and goals) from time to time
- Accuracy is not all
- New educational page on DataViz
- Optimizing KNIME -> Maybe another blog post?
How could this be extended?

• Improve text processing phase (tagging)
• Use word embedding
  – Problem: Document Vector leads to big and sparse feature spaces
  – Solution: Train a vector representation for each word using the Word2Vec
• Use the Keras integration to replace the Random Forest with a Neural Network which uses LSTM layers.
• Investigate the role of parameters:
  – 10% of uncertain
  – K=1 in k-Nearest Neighbors
  – Forgetting functions?
• Add speech recognition? KNIME YouTube videos as additional resource
Where to find more

- Presentation available immediately
- Series of blog posts in the next weeks
- Workflows on EXAMPLE Server
- Collection of blog posts in a whitepaper
KNIME & Teacher Bots: From Workflows to Micro Services

Kathrin Melcher
Vincenzo Tursi
Rosaria Silipo
Phil Winters
The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.