Palladian

KNIME User Group Meeting, Zürich
02.02.2012

Philipp Katz, TU Dresden

About us

- Information retrieval team at Lehrstuhl Rechnernetze, TU Dresden
- Current research focus
  - Efficient feed polling strategies
  - Index field extraction from OCR data
  - Forum Q/A detection
  - Distributed retrieval architectures
  - Relation detection
  - Entity and fact extraction

Palladian?

- Java-based toolkit for information retrieval
- Developed by David Urbansky, Klemens Muthmann, Philipp Katz, and Sandro Reichert
- Provide users with a basic set of tools
    - Used in our dissertations
    - Used by our students
- Growing set of tools, implementations which might be of general interest, are contributed

Palladian's Capabilities

- **Preprocessing:** feature extraction, tokenization, n-grams, page segmentation, content extraction
- **Classification:** text classification, product classification, sentiment analysis, language detection, …
- **Extraction:** named, entities, dates, keyphrases
- **Retrieval:** web search, feed reading, ranking retrieval, wiki crawling

Palladian’s Strengths

- Where does Palladian excel?
  - Product classification
  - Feed reading
  - Date recognition
  - Combining algorithms and techniques:
    9 NERs, 6 keyphrase extractors,
    3 content extractors

Palladian KNIME Nodes

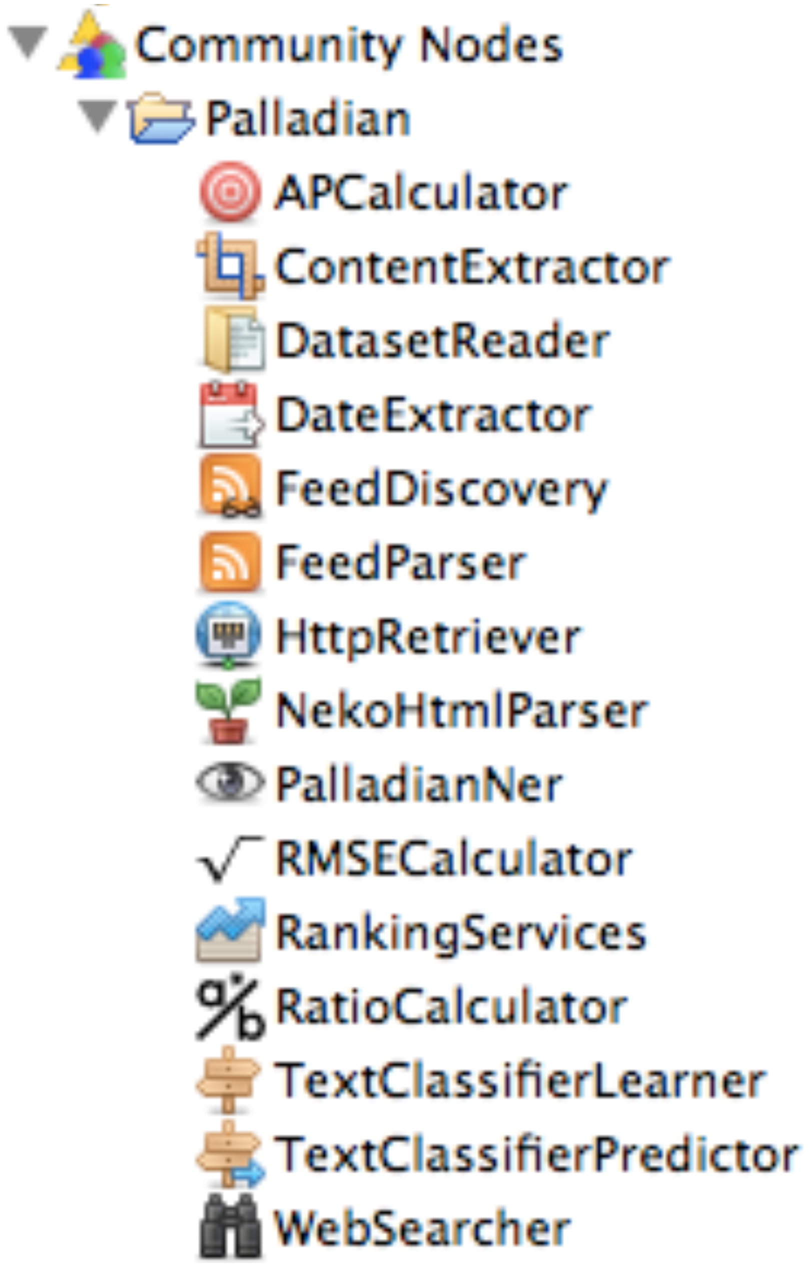
- We have been using KNIME since for a long time, mainly the ML nodes
- Idea
  - Provide Palladian's functionalities as KNIME nodes
  - Allow for quick prototyping of Palladian functionality without having to write code

Palladian KNIME Nodes

- Recent activities
  - Refactoring session during holidays
  - Proxy support :)
  - Implementation of *“Noding Guidelines”*
  - Integration with existing nodes (Text Processing, XML)

Palladian KNIME Nodes

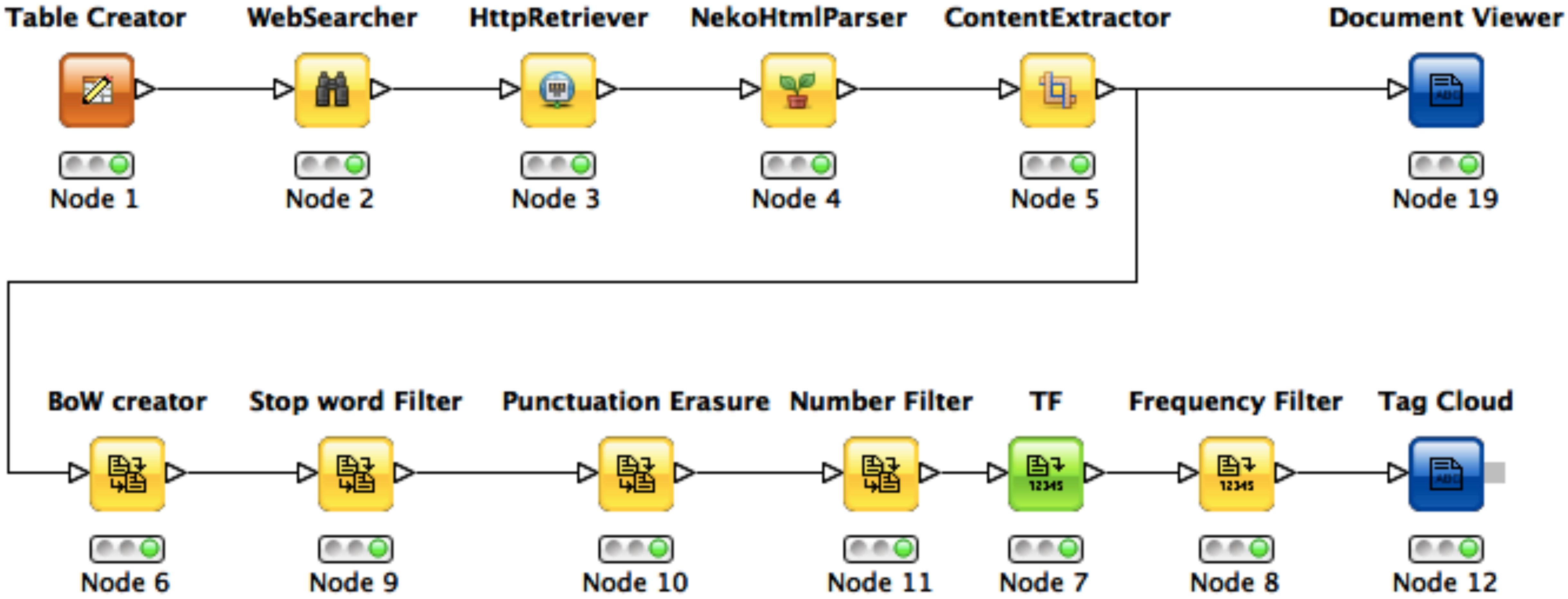
- Text classification
- Content extraction
- Date extraction
- Named entity recognition
- Web search
- HTML, RSS, Atom parsing
- Web 2.0 ranking retrieval
- Evaluation metrics



- Community Nodes
  - Palladian
    - APCalculator
    - ContentExtractor
    - DatasetReader
    - DateExtractor
    - FeedDiscovery
    - FeedParser
    - HttpRetriever
    - NekoHtmlParser
    - PalladianNer
    - RMSECalculator
    - RankingServices
    - RatioCalculator
    - TextClassifierLearner
    - TextClassifierPredictor
    - WebSearcher

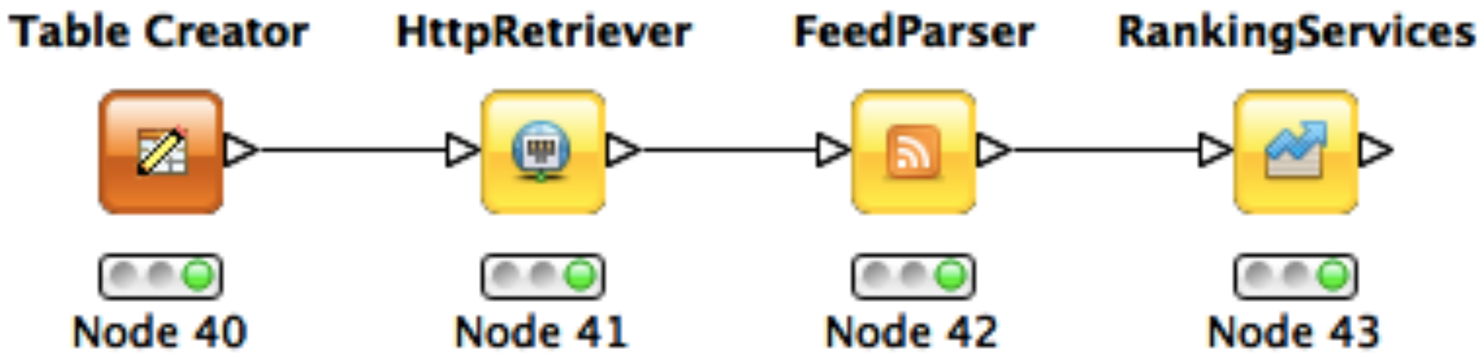
Examples

- Palladian + KNIME Text Processing



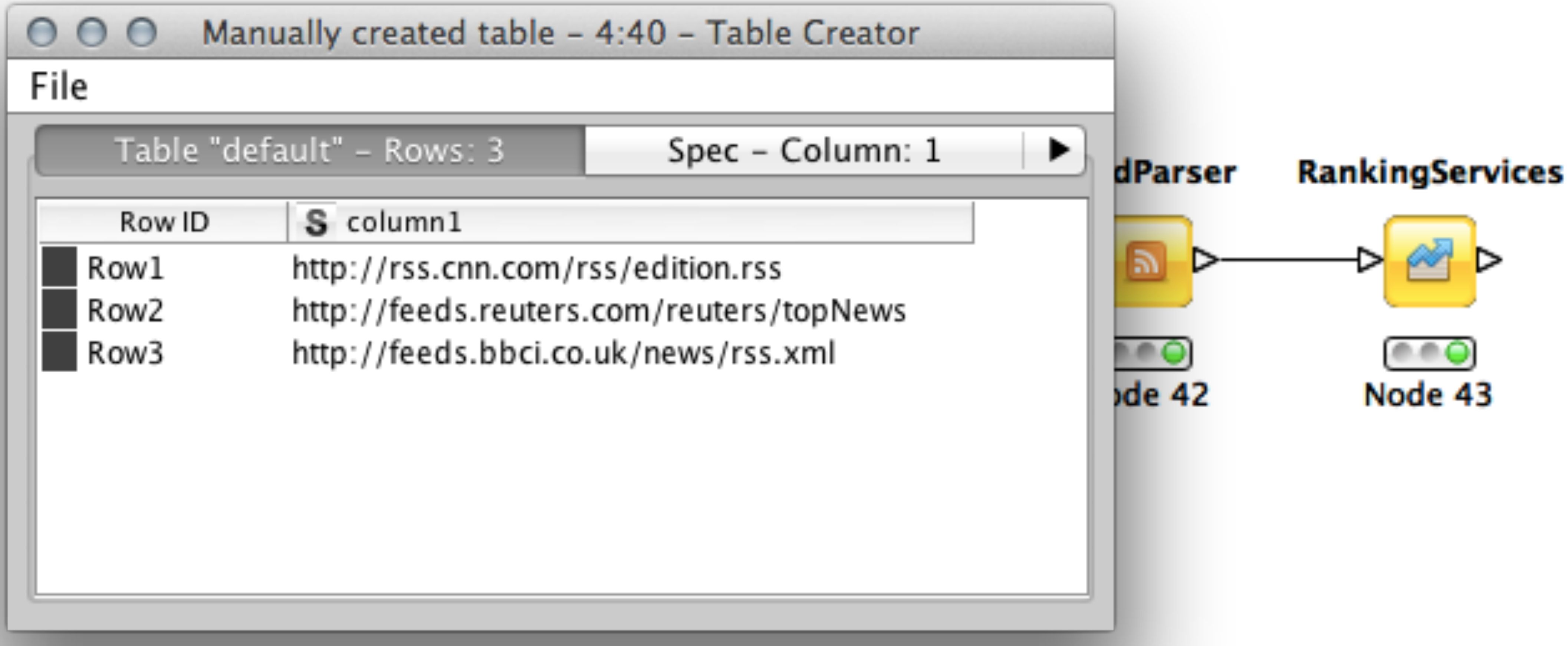
Examples

- Check news feeds, retrieve ranking values from Facebook, Bit.ly, Google, …



Examples

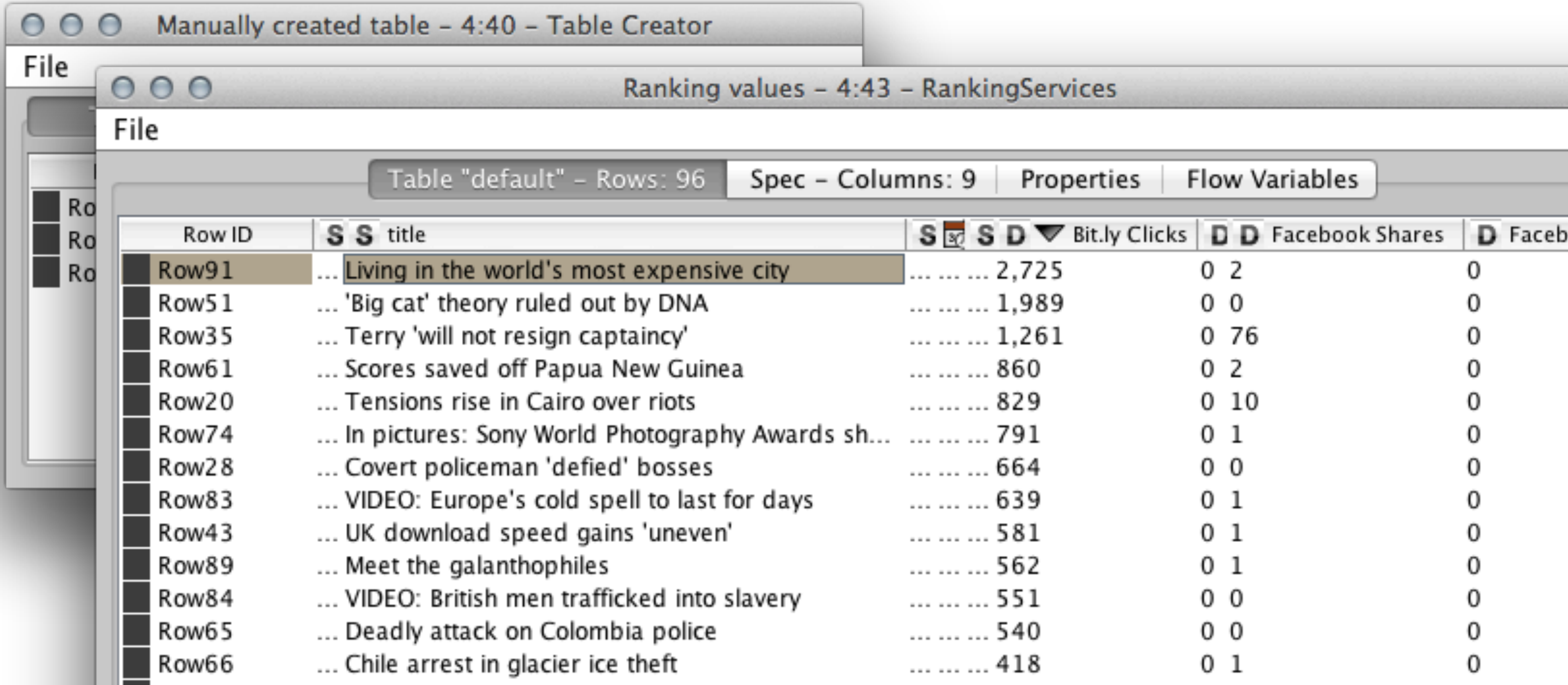
- Check news feeds, retrieve ranking values from Facebook, Bit.ly, Google, …


Manually created table - 4:40 - Table Creator
File
Table "default" - Rows: 3
Spec - Column: 1
dParser
RankingServices
ode 42
Node 43

Row ID	S column1
Row1	http://rss.cnn.com/rss/edition.rss
Row2	http://feeds.reuters.com/reuters/topNews
Row3	http://feeds.bbci.co.uk/news/rss.xml

Examples

- Check news feeds, retrieve ranking values from Facebook, Bit.ly, Google, …



Manually created table – 4:40 – Table Creator

File

Ro

Ro

Ro

Ranking values – 4:43 – RankingServices

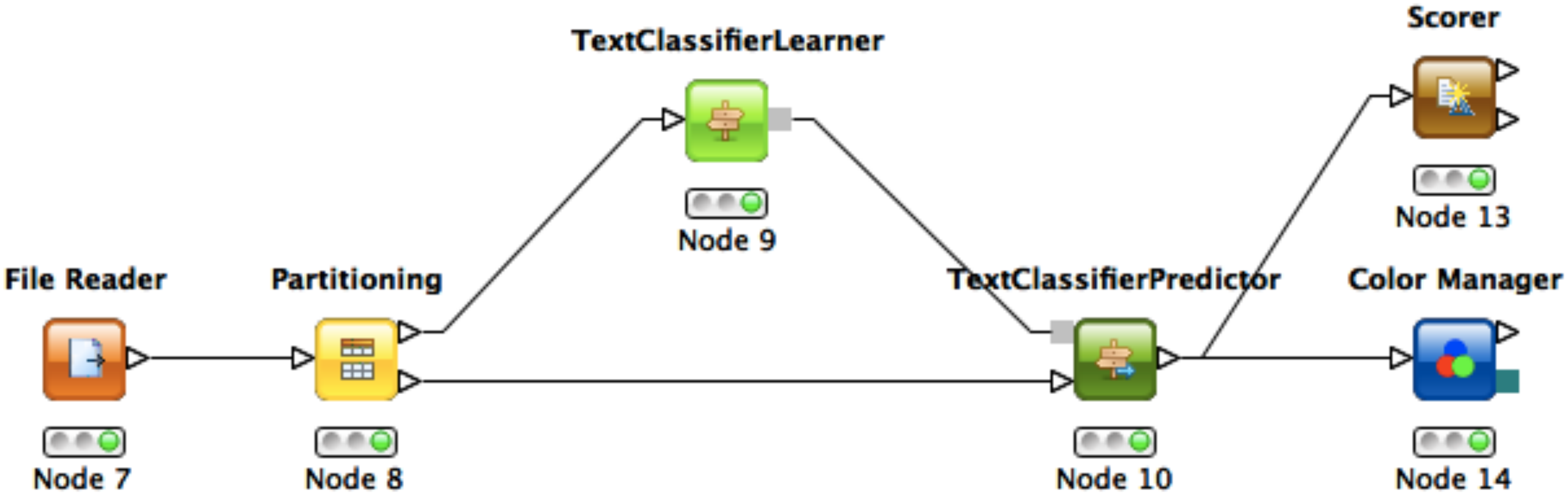
File

Table "default" – Rows: 96 | Spec – Columns: 9 | Properties | Flow Variables

Row ID	S	S title	S		S	D ▼ Bit.ly Clicks	D	D Facebook Shares	D Faceb
Row91	...	Living in the world's most expensive city	...	...	...	2,725	0	2	0
Row51	...	'Big cat' theory ruled out by DNA	...	...	...	1,989	0	0	0
Row35	...	Terry 'will not resign captaincy'	...	...	...	1,261	0	76	0
Row61	...	Scores saved off Papua New Guinea	...	...	...	860	0	2	0
Row20	...	Tensions rise in Cairo over riots	...	...	...	829	0	10	0
Row74	...	In pictures: Sony World Photography Awards sh...	...	...	...	791	0	1	0
Row28	...	Covert policeman 'defied' bosses	...	...	...	664	0	0	0
Row83	...	VIDEO: Europe's cold spell to last for days	...	...	...	639	0	1	0
Row43	...	UK download speed gains 'uneven'	...	...	...	581	0	1	0
Row89	...	Meet the galanthophiles	...	...	...	562	0	1	0
Row84	...	VIDEO: British men trafficked into slavery	...	...	...	551	0	0	0
Row65	...	Deadly attack on Colombia police	...	...	...	540	0	0	0
Row66	...	Chile arrest in glacier ice theft	...	...	...	418	0	1	0

Examples

- Palladian text classifier for sentiment analysis



Examples

- Palladian text classifier for sentiment analysis

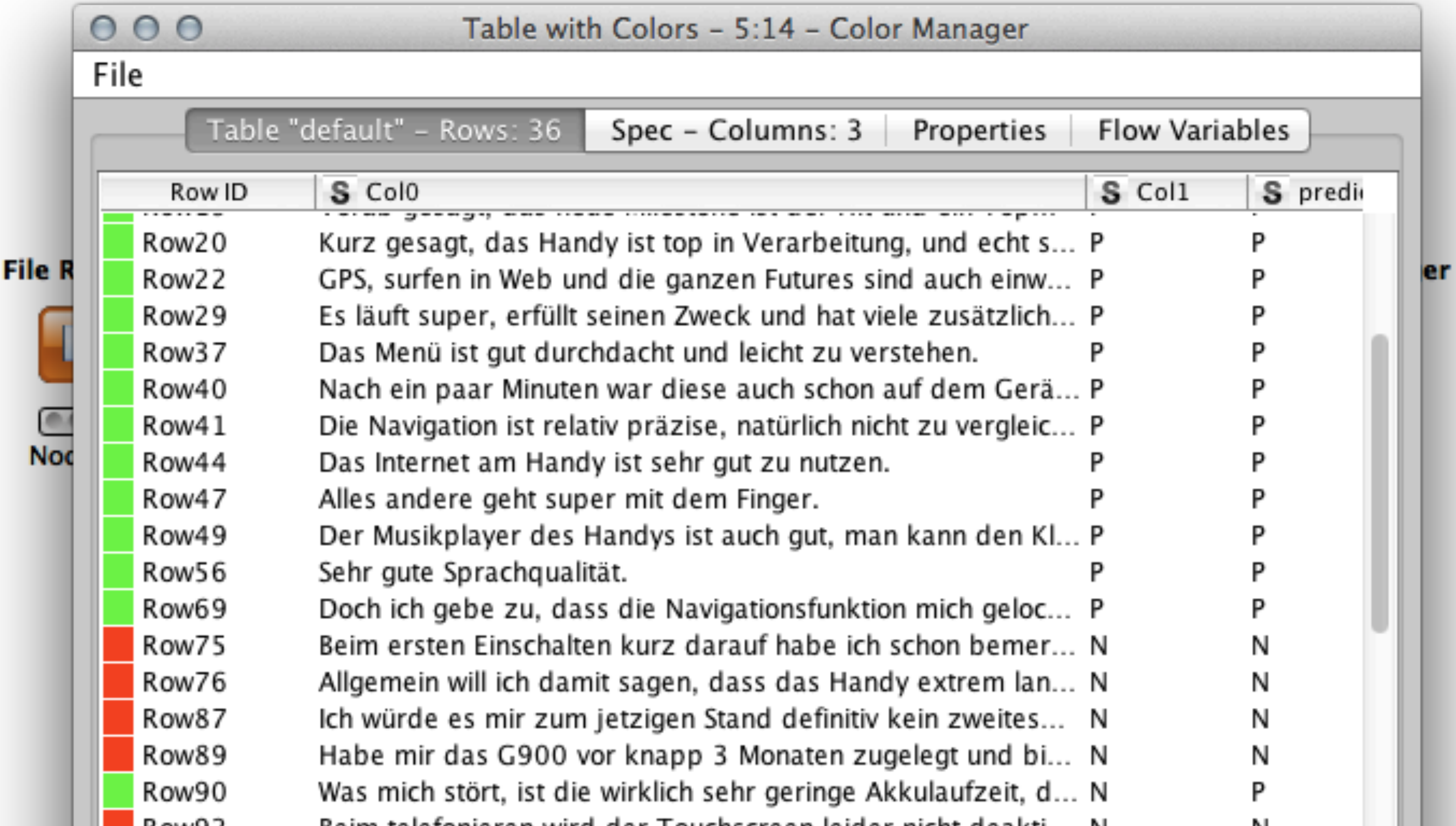


Table with Colors – 5:14 – Color Manager
File
Table "default" – Rows: 36
Spec – Columns: 3
Properties
Flow Variables
File R
Noc
er

Row ID	S Col0	S Col1	S predi
🟩 Row20	Kurz gesagt, das Handy ist top in Verarbeitung, und echt s...	P	P
🟩 Row22	GPS, surfen in Web und die ganzen Futures sind auch einw...	P	P
🟩 Row29	Es läuft super, erfüllt seinen Zweck und hat viele zusätzlich...	P	P
🟩 Row37	Das Menü ist gut durchdacht und leicht zu verstehen.	P	P
🟩 Row40	Nach ein paar Minuten war diese auch schon auf dem Gerä...	P	P
🟩 Row41	Die Navigation ist relativ präzise, natürlich nicht zu vergleic...	P	P
🟩 Row44	Das Internet am Handy ist sehr gut zu nutzen.	P	P
🟩 Row47	Alles andere geht super mit dem Finger.	P	P
🟩 Row49	Der Musikplayer des Handys ist auch gut, man kann den Kl...	P	P
🟩 Row56	Sehr gute Sprachqualität.	P	P
🟩 Row69	Doch ich gebe zu, dass die Navigationsfunktion mich geloc...	P	P
🟥 Row75	Beim ersten Einschalten kurz darauf habe ich schon bemer...	N	N
🟥 Row76	Allgemein will ich damit sagen, dass das Handy extrem lan...	N	N
🟥 Row87	Ich würde es mir zum jetzigen Stand definitiv kein zweites...	N	N
🟥 Row89	Habe mir das G900 vor knapp 3 Monaten zugelegt und bi...	N	N
🟩 Row90	Was mich stört, ist die wirklich sehr geringe Akkulaufzeit, d...	N	P
🟥 Row92	Beim telefonieren wird der Touchscreen leider nicht deakti...	N	N

Thank you!

Any questions?

http://palladian.ws
philipp.katz@tu-dresden.de

The Palladian nodes are available via
KNIME community contributions.