



Overview of the TUD Palladian IR Toolkit

Constance, KNIME Open Source Days

05/10/2011

System Architecture – TU Dresden

- Information Retrieval team of about 8 people
 - Focus on IR from the Web
 - Efficient feed reading strategies
 - Index field extraction from PDF files
 - Forum Q/A detection
 - Distributed architecture and relation detection
 - Entity and fact extraction





Palladian?

- shared implementation and experience
- re-use good student's results
- added value to the community

What for?

Capabilities

Classification

- Text
- Numeric
- Nominal
- Language

Extraction

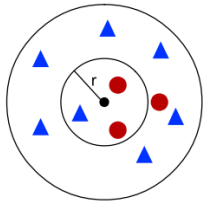
- HTML Text
- Dates
- Tags
- Entities

Pre-Processing

- Tokenization
- N-Grams
- Web Page Segmentation
- Text Similarities

Retrieval

- Search Engines
- Feed Reader
- Wiki Crawler



Where does Palladian excel?

Product Classification



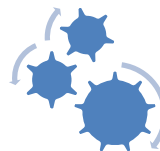
RESEARCH GARDEN
SMART INNOVATION ECOSYSTEMS

Feed Reading



Date Recognition

Combining Algorithms and Sources



8 NERs

3 Keyphrase Extractors



Functionality of Palladian in KNIME

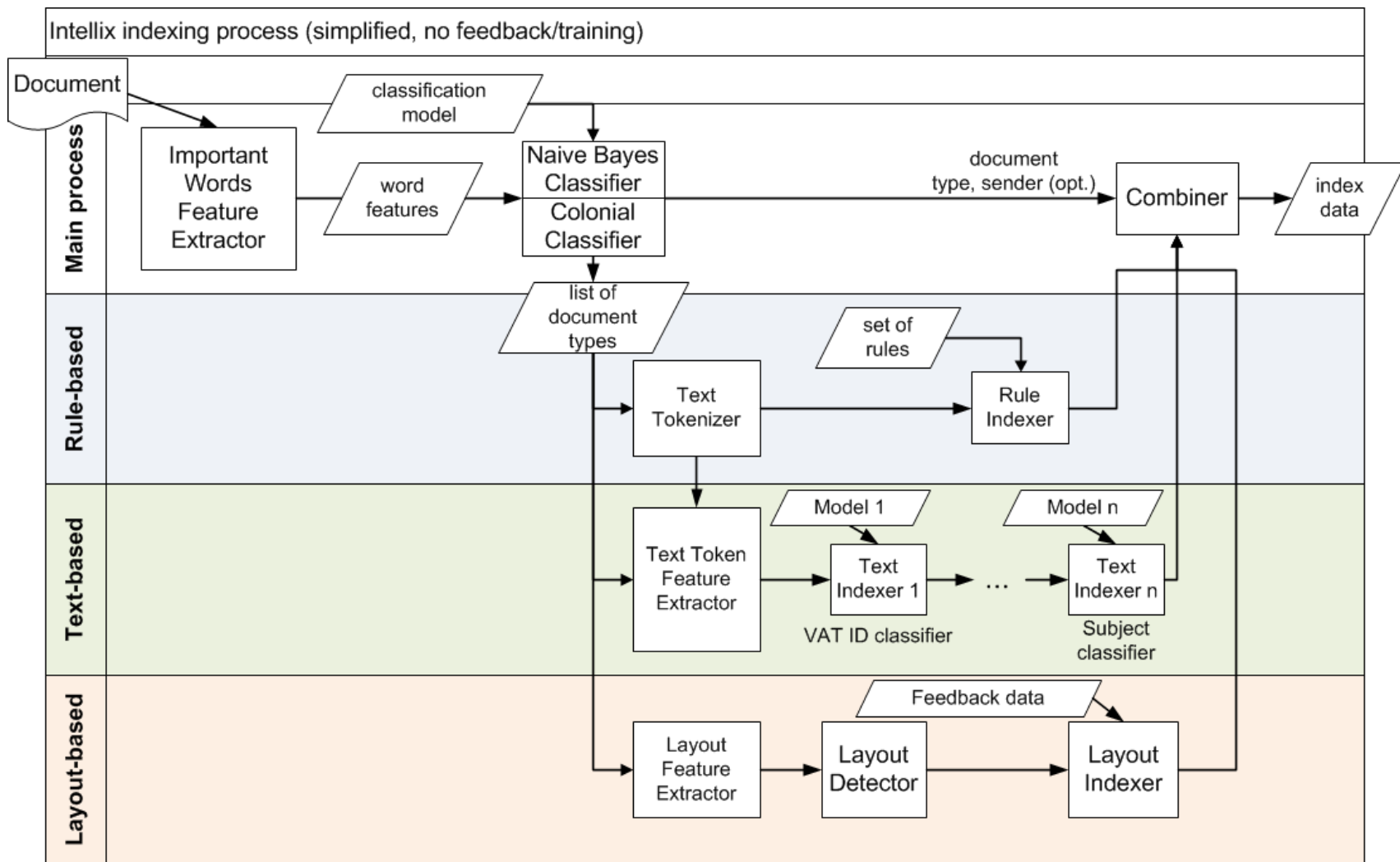
- AP - Average Precision
- RMSE – Root Mean Square Error
- BayesClassifier
- Content Extractor
- Date Recognizer
- Document Retriever
- Feed Discovery
- FeedRetriever
- KNN – K-Nearest Neighbor
- Palladian Named Entity Recognizer
- Ratio Normalizer Node
- TextClassifier
- WebSearcher



Demo



Intellix Workflow





Our Questions

- Any functionality that you would like to have?
- Export of the workflow -> execute (BPMN)?
- Continuous integration? (we use maven)
- Handling big data, streaming
 - Progress
 - Memory consumption



Your Questions



www.palladian.ws