averbis
text analytics

INFORMATION discovery

Whitepaper

October 2021

Summary

Content

1 ABOUT AVERBIS

Averbis GmbH offers life sciences text mining and machine learning solutions for the analysis of unstructured data. We help analyzing company internal and external documents and other textual resources. Our ultimate goal is to turn text into actionable information, automate intellectual processes and make useful predictions.

Our cross functional team is constantly working on new innovations to always give our customers from the pharmaceutical and health services a decisive competitive edge.

Averbis was founded in 2007 in Freiburg im Breisgau, Germany.

The Averbis Team is made up of motivated and dedicated staff who has committed itself to the goal of providing our customers with the best possible service and highest quality. Our team (computer science, medicine, biology,) possesses long years of experience and excellent skills in the designing and implementation of software applications for processing natural language. Together with our customers we determine your individual needs and compile tailor-made solutions in project teams to incorporate low-cost, optimal software solutions in the daily routine within a very short time.

Averbis stands for cutting edge technology in text analysis, natural language processing, semantic search and labeling, information extraction and machine learning. Our involvement in international research projects guarantees always being up to date in IT research. Together with leading research institutions, we develop trend-setting solutions for the most varied branches.

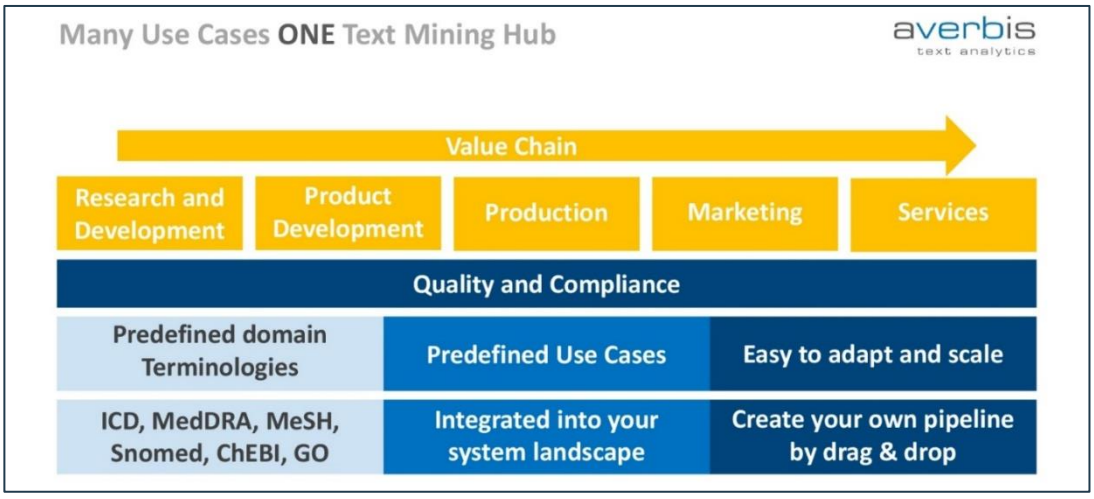
We have great competence and knowledge in agile software development and the underlying and supporting infrastructures (scrum, test driven development & test automation, continuous integration with Jenkins, clean code principles ...).

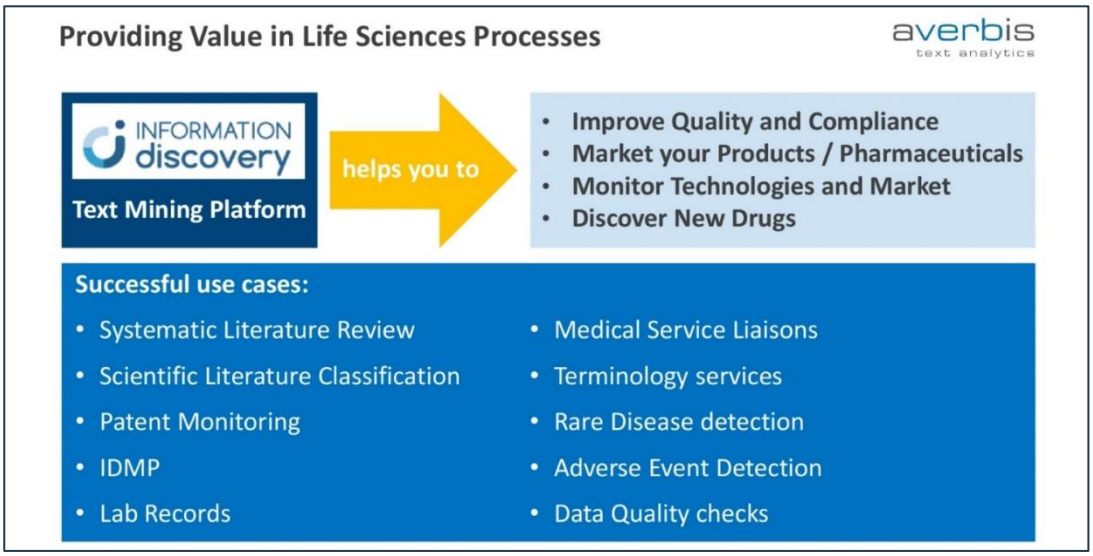
2 WHAT MAKES AVERBIS UNIQUE?

- We offer a perfect ensemble of technologies necessary for successful AI-based projects - bundled in a single application framework. Our solution includes statistical and non-statistical text mining, machine learning, terminology management and a faceted semantic search.
- We combine technical expertise with pharma and health care knowledge.
- We have an extensive proven experience in many use cases and a broad customer base. We solve all your text mining problems, so you don't have to choose a different provider for each use case.
- If you still would like to use another provider for a special task, then it is to your advantage that our software framework employs the text mining framework UIMA. This framework allows you to easily integrate third-party text mining components. This avoids vendor lock-in and gives you maximal flexibility.
- The framework provides many possibilities to integrate your own text mining pipelines or configurations. This allows you to decide whether you want us to implement it, by a third party or yourself, and to always choose the most cost-effective way to implement your project.
- We are the creators and committers of Apache UIMA Ruta, the most powerful Text Mining Rule Engine in the world. Ruta enables a simple and ultra-fast non-statistical implementation of text mining problems.

3 AVERBIS LIFE SCIENCE PLATFORM

Life Sciences text mining and machine learning platform fitting right into your company infrastructure. We help Life Sciences companies using full potential of their distributed and unstructured data to improve processes and compliance. Support your business through quick and integrated development cycles and fast implementation of business-critical use cases.





4 AVERBIS INFORMATION DISCOVERY

Text Mining & Machine Learning with Information Discovery

With the platform "Information Discovery" we have a leading text mining and machine learning platform that allows you to get insights in your unstructured data and explore important information in the most flexible way. Information Discovery collects and analyzes all kind of documents, such as office documents, reports, patents, research literature, databases, websites, and other enterprise repositories.

By parsing and analyzing content, and creating a searchable index, Information Discovery helps to perform text analytics across all relevant data in your enterprise and make that data available for analysis and search. It allows you to explore facts and relationships across many sources that would otherwise be hidden in unstructured data. The full manual for Information Discovery is accessible online[1].

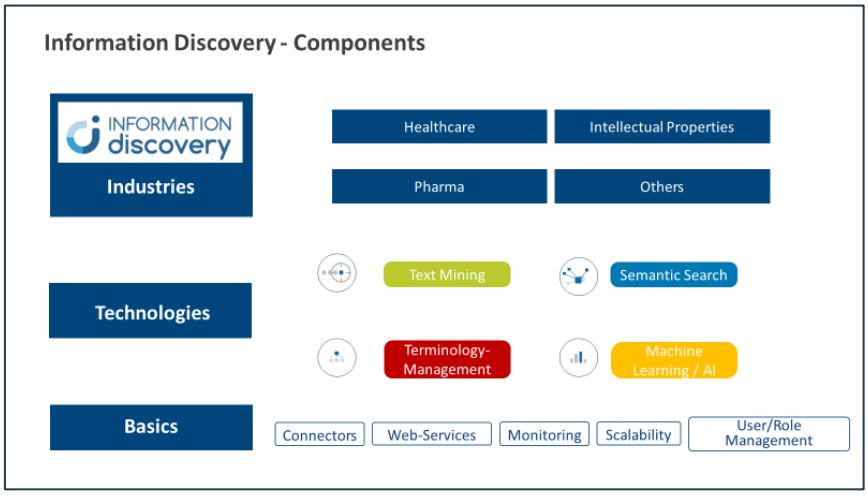


Figure 1: Information Discovery's components and framework

[1] https://help.averbis.com/

Information Discovery allows you to leverage your data for business-critical decisions:

- ✓ Integrate heterogeneous data from various sources in a single application
- ✓ Analyze your Big Data in a short time
- ✓ Structure your unstructured content
- ✓ Easily drill down your results using advanced filters
- ✓ Discover hidden facts and relations in your data
- ✓ Develop your own data-driven applications with Information Discovery

5 FUNCTIONAL COMPONENTS OF INFORMATION DISCOVERY:

The platform comes with four functional components: Text Mining, Machine Learning, Terminology Management and Semantic Search. All components are accessible via RESTful APIs to support existing workflows. Furthermore, the system covers other non-functional requirements such as user and role management (also via LDAP), scalability, monitoring functionalities and connectors to different data sources.

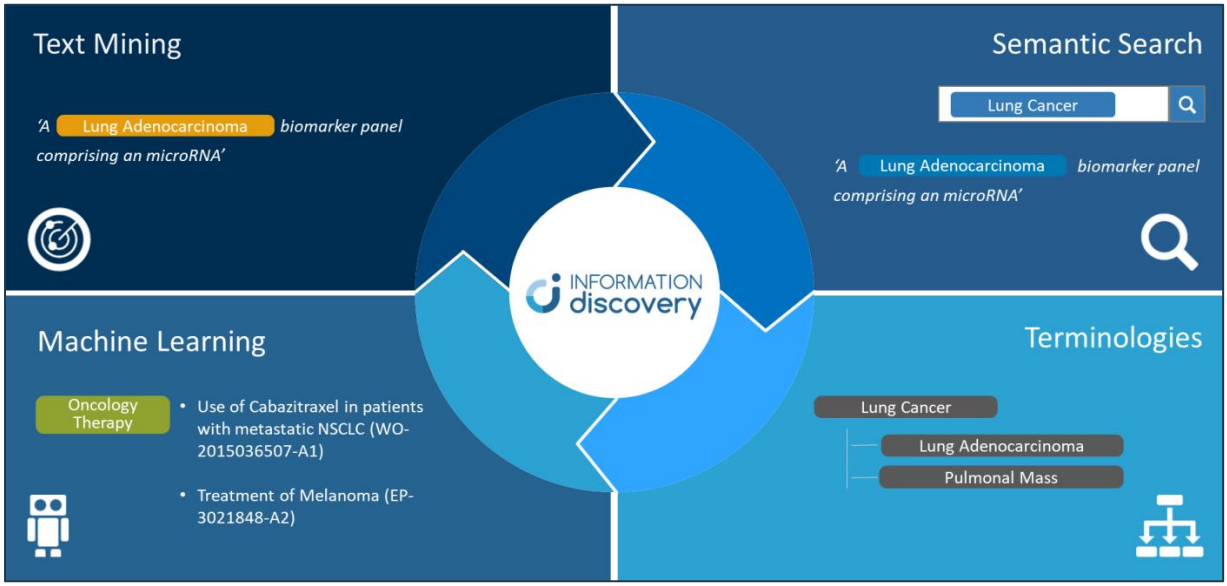


Figure 2: The four building blocks of Information Discovery

5.1 Text Mining

Information Discovery's Text Mining is a leading text-mining solution for the extraction of information from large document collections. The solution identifies single information units, as well as relevant facts and contexts at both high coverage and precision.

The Text Mining Solution is 100% embedded in the free and open Apache UIMA Framework (http://uima.apache.org/). This framework provides interfaces to connect individual analysis components to each other and to determine their order. Thus, a transparent and standardized connectivity between the individual analysis components is guaranteed. The members of the Averbis team are official UIMA committers and constantly working on UIMA advancements.

The platform comes with either the option to configure processing pipelines programmatically or using a graphical configuration editor, which is easy to use for end-users.

On top of UIMA, many modular software components are integrated in a comprehensive text mining solution:

- **Paragraph and Section Detection:** Our software is able to detect paragraphs and sections both using heuristic and statistic approaches.
- **Syntax Analysis**: We provide both rule-based and statistical annotators for syntactic analysis of text. The annotators are able to identify sentences, tokens, POS tags and chunks (=shallow parsing) on documents. The statistical approaches are based on Apache OpenNLP, the leading machine learning based toolkit for the processing of natural language text.
- **Term Detection**: here, the terminology structure is flexible and enables the organization of synonyms and various attributes which may play a role in the annotation process. The terminology matching procedure can be carried out on linguistic variants (e.g. "innovation ability" vs. "the ability to innovate"). Multiple Terminologies can be managed in several languages via a web-based editor. Existing, specialized terminologies and other term lists can be imported, edited and made available for the extraction of information and term annotation.
- **Rule Engine**: Ruta is a rule-based system designed for information extraction tasks, but it is also applicable for many natural language processing use cases. This Apache UIMA component consists of two major parts: An Analysis Engine, which interprets and executes the rule-based scripting language, and the Eclipse-based tooling (Workbench), which provides various support for developing rules. As it allows complex and nested patterns, it is a powerful and flexible way for all kind of rule-based information extraction tasks.
- **Entity Recognition**: entities are recognized by mere statistical calculation of scores of different pieces of information and attributes of context words, thus precisely identifying person or product names, organizations or geographic information.
- **Third Party integration**: It is easy to integrate third party text mining solutions using the UIMA PEAR mechanism. That allows you to not only rely on one text mining provider, but to combine the solutions of different vendors in a best-of-breed approach.

We deliver a huge set of further annotators ranging from domain independent to domain specific solutions, e.g. for the healthcare and the pharmaceutical domain. Each of these annotators ships with unique outstanding features resulting in high performance annotation pipelines. The full list of annotators and terminologies integrated in Information Discovery, as well as some performance metrics, is available in the annex of this document.

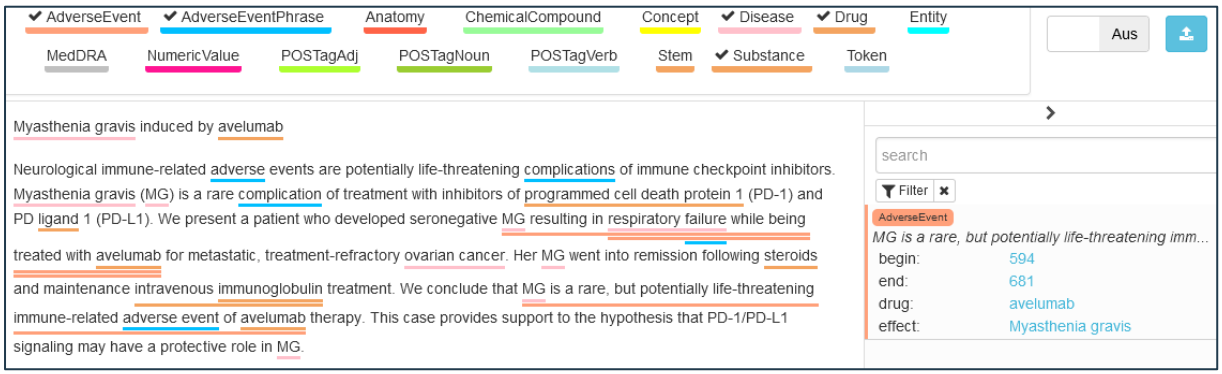


Figure 3: Text Mining functionalities are accessible via the graphical user interface or an API

5.2 Machine Learning for Text Classification

The text classification module of the platform allows customers and integrators to create Artificial Intelligence applications based on text data. We provide classification techniques based on advanced Text Mining and Machine learning algorithms. They can be accessed both via a powerful graphical user interface and with a simple RESTful web interface.

Document classification can be integrated with any project through the web API. By this, our customers can implement capabilities to facilitate sentiment analysis, content monitoring, technology categorization, predictive coding, clustering, alerting and concept searching.

Machine learning techniques massively support human experts in complex annotation and labeling tasks. The expert, instead of programming a rule for every possible outcome, provides a set of training data that shows examples of how the decision should be made. Computers learn from the experience of information professionals and produce useful predictions on new, unseen examples after being trained on a learning data set.

Automatically categorizing large data sets of documents with a high number of (hierarchical) categories while still opting for excellent prediction quality requires a sufficient number of learning data. The concept of active learning minimizes the effort of manual creation of such data by intelligent data sampling and iterative supervised learning.

The software can be used to interactively train the classification model (active learning) and to integrate the resulting classification via API into other (existing) applications. The platform provides feedback on the current prediction quality (via 10fold cross-evaluation), thus, the user always knows about the consistency of the training data.

Besides support vector machines we also have great experience with current deep learning models such as convolutional neural networks[2] or BERTs.

[2] We published the best results ever achieved for several standardized patent classification settings, see: Optimizing Neural Networks for Patent Classification. Louay Abdelgawad (Averbis GmbH, Freiburg), Peter Kluegl (Averbis GmbH, Freiburg), Erdan Genc (Averbis GmbH, Freiburg), Stefan Falkner (Albert-Ludwigs University of Freiburg), Frank Hutter (Albert-Ludwigs University of Freiburg). ECML 2019.

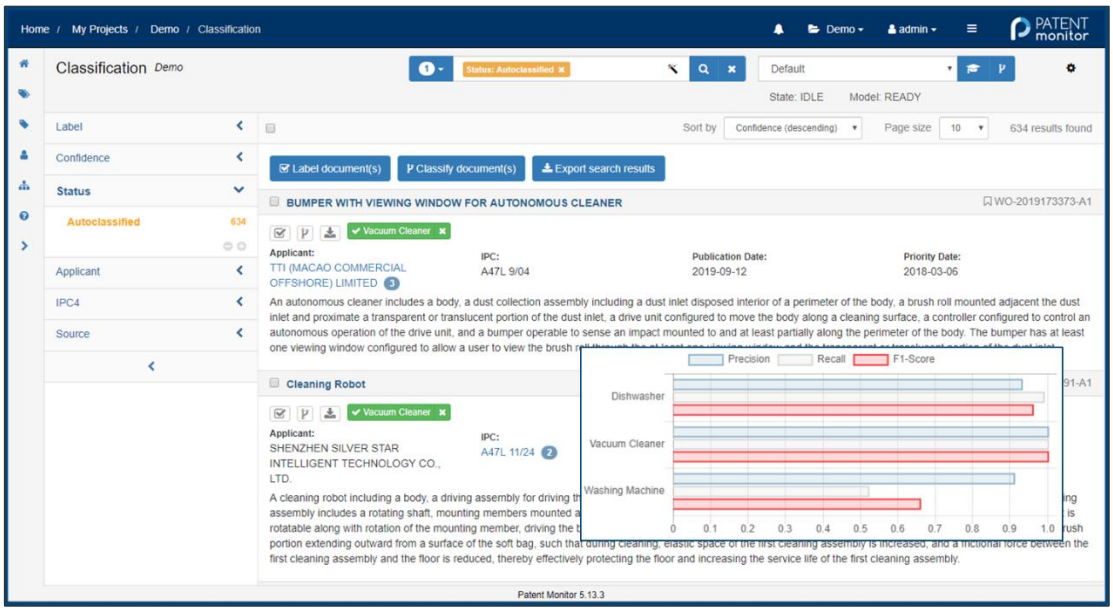


Figure 4: Machine learning for text classification is accessible via the graphical user interface or an API

5.3 Terminology Management

The Averbis Terminology Management System offers innovative tools for handling terminologies and ontologies. These term hierarchies aid in the formal sorting of the individual fields of expertise and establish rules on the contexts of the corresponding expressions, thus enabling conclusions to be drawn from the existing data, contradictions to be detected and missing knowledge to be added.

Via a web-based editor, existing terminologies and other term catalogues can be imported, edited and made available for the extraction of information and term indexing.

Multi-linguality is supported, as is the enhancement of word synonyms and cross references to other terminologies. The editor supports the entry of new terms by means of automatic validation and consistency checks and helps with the enhancement of information from various external sources.

The terminologies are used to easily integrate domain knowledge into the Text Mining component of the platform via configuration. Therefore, different use cases can be realized without further implementation.

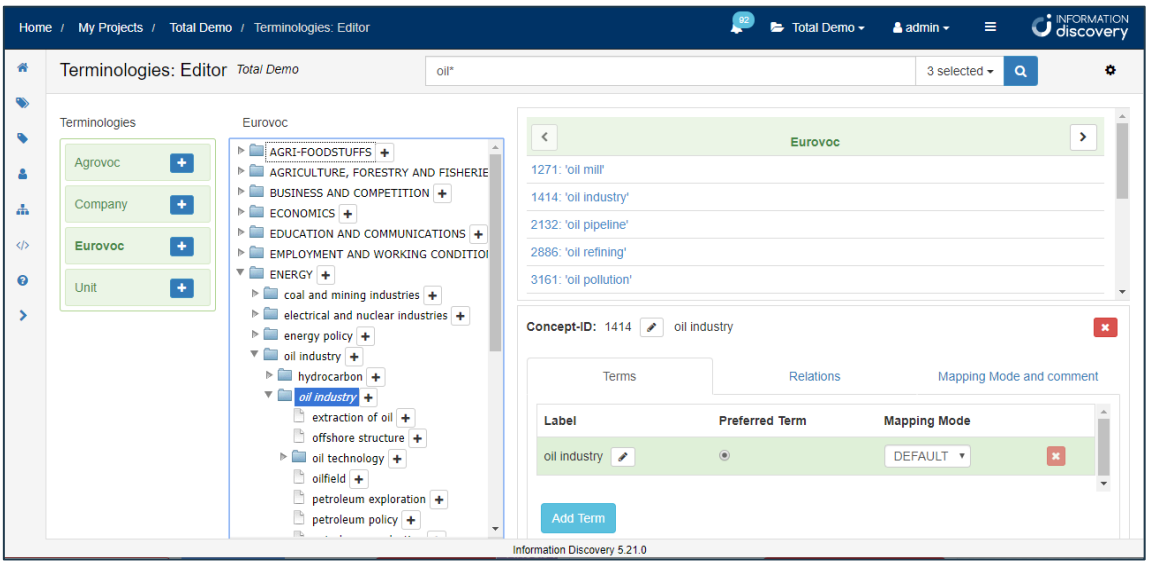


Figure 5: Terminology editor for Text Mining is also accessible via the graphical user interface or an API

5.4 Semantic Search

To be able to find specific relevant documents or evaluate document collections in great amounts of data, we offer a high-performance, scalable semantic search engine that can process user queries within milliseconds. The search engine distinguishes itself by a high degree of parameterization and can make the most varying types of documents searchable.

Due to the integration of Text Mining and Machine Learning components, the search engine offers comprehensive treatment of linguistic phenomena. Even phrases, synonyms or single components of compound words are recognized, and laymen and expert language are normalized.

The incorporated search engine of the platform is based on Solr and can be accessed through a web service. However, any other search engine can be fed with the semantically annotated metadata as result of Text Mining and Machine Learning via the appropriate APIs.

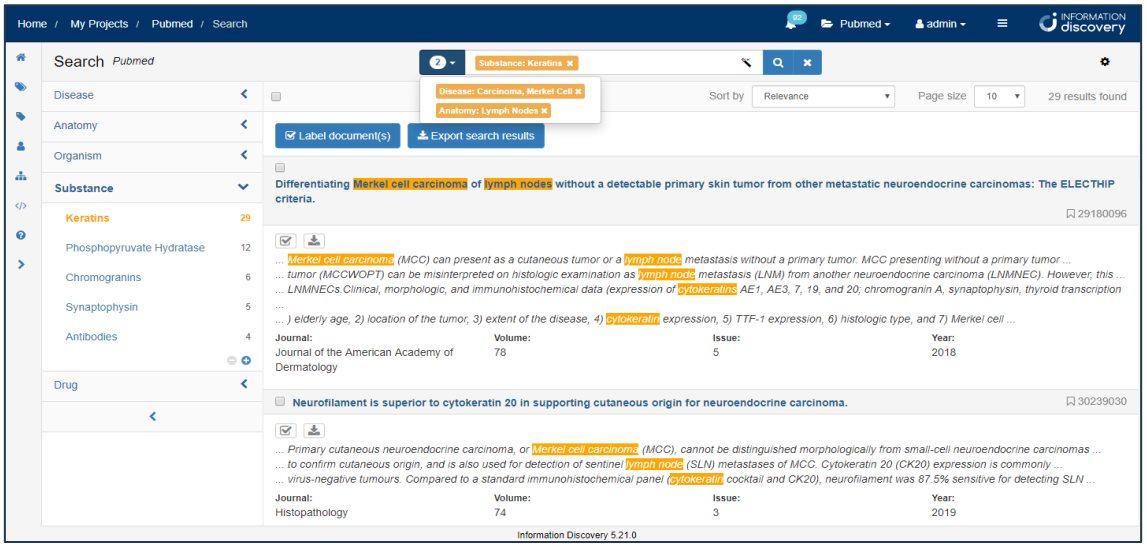


Figure 6: Semantic search is also accessible via the graphical user interface or an API

6 NATURAL LANGUAGE PROCESSING (NLP) AND MACHINE LEARNING FUNCTIONALITY

Standard NLP Components

Information Discovery contains a 100% UIMA[3] compatible text mining platform. It offers numerous annotators for the semantic analysis of text. Standard NLP components are already included in the software, such as: language detection, sentence detection (statistical and rule-based), tokenizer (statistical and rule-based), stemmer, segmenter, abbreviation detector, annotators for numeric values, measurements, times and dates, part-of-speech tagger, chunker, statistical entity recognition, concept mapper, wordlist annotator, keyword extractor, etc.

For a complete overview, please refer to the "Text Analysis Component Reference"[4].

Individual components for specific use-cases can easily be defined by using the Ruta rule language[5] and/or the PEAR mechanism[6]. Furthermore, third party components using the UIMA standard can be integrated and used.

Domain-specific NLP Components

We offer a wide range of domain specific annotators for the healthcare and the life science industry. See the appendix for the list of available components which are constantly being extended.

Adding of new NLP components

New NLP components are integrated into Information Discovery in terms of UIMA PEAR packages. The components themselves can be implemented in Java or Apache UIMA Ruta. A connection to Python libraries is possible.

Language Support

The NLP standard components (see above) support major European languages. However, some of the components are language-independent and can therefore be used for many languages. By using language-specific wordlists and terminologies that can be imported or defined at any time, language-specific use-cases can be served.

Access to Cloud Services

Existing NLP cloud services can basically be integrated into pipelines via REST interface. For this purpose, a client for the cloud service can be implemented, packaged as a PEAR component and then integrated into Information Discovery pipelines.

Configuration of NLP tasks

NLP tasks can be configured using a simple-to-use graphical user interface[7] and via REST API.

Machine-Learning based Document Classification

With Information Discovery you can easily train statistical classifier for document categorization. You are free to define your target categories, based on relevancy criteria, technological areas or other topics that you are interested in. By using the GUI or by uploading training data you provide classification examples for each of the categories. Only 10 to 20 examples for each category are sufficient to enable

[3] **https://uima.apache.org/**
[4] Information Discovery Manual: **Text Analysis Component Reference**
[5] Averbis Knowledge Base: **Apache UIMA Ruta Tutorial**
[6] Averbis Knowledge Base: **Building and Integrating your own Text Analysis Component**
[7] Information Discovery Manual: **Pipeline Configuration**

first predictions. Afterwards, the system automatically categorizes new documents. It also provides a level of confidence for each of its predictions. Take a look at the system's predictions and correct the results if needed. This is how the system learns from you and steadily improves.

Connectors

Information Discovery provides connectors for importing documents from popular file formats such as .docx, .pdf, .xlsx and relational databases that can be accessed via JDBC.

Information Discovery does currently offer connectors for automatic delivery to external data sinks via individual provision of corresponding PEAR components.

Annotation Editor

The front-end of the annotation viewer (see above) can also be used to manually label new data or to make modifications to manually or automatically created annotations, e.g. in order to create gold standards. These gold annotations can be used to automatically carrying out evaluations of specific NLP components (in terms of precision, recall, f1-score etc.).

Manually created data are stored in the database and are available for export.

Visualization & GUI

The application comes with an annotation viewer (and editor) that graphically displays the results of applying a NLP pipeline to texts. Using this tool, all documents from different document sources can be easily viewed, section by section, whilst all annotations are graphically highlighted.[8]

For the quality assessment and improvement of text analysis pipelines, an aggregated overview of the assigned annotations is often helpful, for which a graphical user interface is also provided.[9]

A graphical user interface is also available for the configuration of NLP pipelines.[10]

Monitoring

The graphical user interface of Information Discovery displays key figures for the configured text analysis pipelines. These include, for example, throughput at pipeline and component level.

6.1 Frontend & Webservices via REST API

Information Discovery provides a browser-based user interface for managing projects, administering users, configuring text analysis pipelines and components, monitoring performance metrics, managing terminology, etc. Furthermore, Information Discovery has an integrated annotation editor and evaluation tools for text analysis results.

The REST API provides access to Averbis Information Discovery functionality for third-party applications. The API is HTTP-based, so it can be used with any language that has an HTTP library, such as curl (see online documentation[11] for more information).

The webservice uses API tokens to protect resources against unauthorized use. Administrators can create personalized API tokens and use them for authentication on API calls.

Averbis Information Discovery comes with a built-in browser interface for the REST API based on Swagger UI. It allows you to get an overview of the API and to submit sample requests directly from the browser.

[8] Information Discovery Manual: **Viewing and Editing Annotations**
[9] Information Discovery Manual: **Annotation Overview**
[10] Information Discovery Manual: **Pipeline Configuration**
[11] Information Discovery Manual: **REST API**

6.2 Users, Role Management and Access Control

Information Discovery provides an integrated user management with predefined roles that can be assigned to users on a project-specific basis. This means, for example, that access for certain users can be restricted to individual projects. Furthermore, Information Discovery offers an integration of LDAP / Active Directory. Therefore, user groups are mapped to roles within Information Discovery.

6.3 Architecture & Hosting

Information Discovery supports container-based cloud deployments. The software is available as multiple Docker containers, which enable easy and fast installation and upgrades both locally and from cloud vendors such as AWS or Azure.

6.4 Technologies

With respect to natural language processing and Machine Learning, we are experts for Apache UIMA (Unstructured Information Management Architecture) and UIMA RUTA (both active committers), as well as uimaFIT and UIMA-AS (asynchronous scaleout). Further competences are OpenNLP, Stanford NLP, Mellet, Factorie, DKPro, Tensorflow, DL4J, fastText, BOHB, pyTorch, Solr, to name a few.

From the perspective of the development of software services, the following technologies are in the foreground: Java EE, Javascript, Python, Angular, CSS/Bootstrap, Tomcat, Maven, Spring, SQL/JPA, Eclipse, Junit, NPM, Swagger (RESTful APIs).

ANNEX A: Information Discovery – Text Mining Using Domain Specific Vocabularies

Averbis has extensive experience with managing own or standard terminologies. The following list gives a non-exhausted overview about terminologies that we already integrated in our terminology management software, including the terminology domain and the publisher of the terminology. Note that license constraints of source vocabularies may apply if used for text mining. Other terminologies can easily be integrated.

HEALTHCARE / PHARMA Terminologies	Categories	Publisher
ATC	Drugs	WHO
ChEBI	Chemical Compounds	EMBL-EBI
ChEMBL	Chemical Compounds	EMBL-EBI
DrugBank	Drugs	University of Alberta
Human Phenotype Ontology	Phenotypes	Human Phenotype Ontology Consortium
FMA	Human Anatomy	University of Washington
ICD-10	Indications	WHO, DIMDI
ICD-O	Oncology	WHO, DIMDI
LOINC	Laboratory	Regenstrief Institute
MedDRA	Misc	MedDRA MSSO
MeSH	Misc	National Library of Medicine
NCI-Thesaurus	Misc	National Cancer Institute
OPS	Procedures, German	DIMDI
RadLex	Radiology	RSNA
RxNorm	Drugs	National Library of Medicine
SNOMEDCT	Misc	IHTSDO
Uberon	Mammalian Anatomy	Mungall et. al
UCUM	Units	Regenstrief Institute
UMLS	Misc	National Library of Medicine
Uniprot	Genes	EMBL-EBI

OTHER Terminologies	Categories	Publisher
Agrovoc	Agriculture	Food and Agriculture Organization
eClass	Products	eCl@ss e.V.
GeoNames	Geo	GeoNames
GND	Misc	German National Library
Quantities	Units, German	Averbis
TEMA (EN, DE)	Scientific Technical Domain, English and German	WTI Frankfurt

ANNEX B: Information Discovery - Pharma Specific Components & Usecases

Components	Comment
Adverse Events	Detection of adverse event mentionings
Anatomy	Anatomical entities, structures, body parts, cells, tissues aligned to NCI, MeSH, (Uberon on request)
Chemicals & Drugs	Substances, Brand names aligned to Drugbank, RxNorm, MeSH
Chemicals	Entity recognition based on FLAIR (tob e released)
Cellines	Entity recognition based on FLAIR (tob e released)
Devices	Medical devices aligned to CHV, NCI, MeSH
Disease	Entity recognition baised on FLAIR (tob e released)
Diseases, Disorders, Indications	ML-based, highly generic w.r.t. features used, classification algorithms (supported: SVM, NB, MaxEnt, Trees). A customized version for patents is Diseases, Syndromes, Abnormalities, Injuries, Signs & Symptoms aligned to CHV, MedDRA, NCI, MeSH
Gene/Protein	Entity recognition baised on FLAIR (tob e released)
Genes and Molecular Sequences	Genes and gene products aligned to NCI
IDMP	Identification of medicinal products: Product characteristics (SmPC), composition (module 3), manufacturers (module 3), 50+ data elements most relevant for IDMP iteration 1, aligned to EMA, MedDRA (en, de, fr)
Laboratory values	Extraction of lab values, measurements (including SI normalization) according to LOINC
Organisms	Living beings, bacteria, fungi, plants, vertebrata & mammals , viruses aligned to NCBI, NCI
Phenomena	Human-caused phenomena and processes, environmental effects, natural phenomena aligned to MeSH, NCI
Physiology	Biological functions: cells, genes, molecules, organisms aligned to CHV, MeSH
Procedures	Diagnostic, laboratory, therapeutic and preventive procedures aligned to CHV, MedDRA, MeSH, NCI
Species	Entity recognition baised on FLAIR (tob e released)

ANNEX B: Information Discovery - Healthcare Specific Components & Usecases

Components	Comment
Diagnoses	Disease classification according to ICD10 (en, de), including contextual resolution
Laboratory values	Lab value detection including parameters (LOINC), values (quant./qual.), units, normalization, interpretation (normal/high/low)
Medication	Medication detection including ingredients, brand names, strengths, dose forms and dosing schemes
Morphology	Morphological classification according to ICD-O (en, de)
Temporal aspects	Hospital admission, length of stay, creation date
Topography	Topographic classification according to ICD-O (en,de)
TNM	Tumor, lymph nodes, metastasis according to TNM nomenclature
Tumor classification	Grading, gleason, R & V classification, side detection

ANNEX C: Information Discovery - Standard Components

Framework	Comment
UIMA Java Framework	
UIMA C++ Framework	
UIMA Default Viewers & Tooling	
PEAR Packaging Facilities	
UIMA-AS Scaleout Framework	
UIMA-AS in the Cloud	
Infrastructure	**Comment**
Simple Server (UIMA REST Service)	
Generic Typesystem	
Web-based Annotation Client	
Scripting Language for Pipeline Configuration	
Core Components	**Comment**
Collection Readers (CR)	
Simple File Reader	
XMI Reader	
Generic XML Reader	
Generic Database Reader	
Annotators	
Tika Annotator	Apache Tika is a toolkit for detecting and extracting metadata and structured text content from various documents using existing parser libraries.
Generic Document Zoning	ML-based, must be trained on specific document-collection (no pre-trained models)
Language Detection	based on document fingerprint statistics; pretrained on 40 languages; can be extended to other languages or adapted to other domains/genres
Document Classification	ML-based, highly generic w.r.t. features used, classification algorithms (supported: SVM, NB, MaxEnt, Trees). A customized version for patents is available
Sentence Splitting, Rule Based	
Sentence Splitting, Trainable	based on ML; trained on news- and bio-nlp genre
Tokenization, Rule Based	
Tokenization, Trainable	based on ML; trained on news- and bio-nlp genre

Part-Of-Speech Recognition	based on ML; trained on news- and bio-nlp genre
Shallow Parsing / Chunking	based on ML; trained on news- and bio-nlp genre
Stemming	based on well-known and approved Porter-Stemmer algorithm
Morphological Analysis	Dictionary-based approach, optimized for biomedical domain
Decompounding	Dictionary-based approach, optimized for biomedical domain
Decompounding, Rule Based	based on JWordSplitter
Stopword Recognition	based on language dependent stopword lists
Invariant Annotator	Used to identify parts of texts (typically tokens) which should not be subject to linguistic analysis (such as e.g., stemming). This holds for proper names, brand names, chemicals etc. Based on definition of patterns (regular expressions). Might be extended according to application-specific needs.
Acronym and Abbreviation Resolution	based on Hearst-Schwarz algorithm to detect and resolve acronyms in text; optimized on abstract of bio-medical research literature
Regular Expression Annotator	
Lemmatizer, Lexicon Based	
Terminology Mapper	allows to find terms of a terminology in the text; allows for different mapping modes on different linguistic levels (e.g., original mode, stems, lemma etc.); allows for some forms of fuzzy matching but also allows for more stricter matching of typical terms by requiring certain Part-Of-Speech tags
Numeric Value Annotator	This component can recognize a wide variety of numeric expressions and their numerical value. These include simple numbers such as 2.3, but also more difficult expressions such as ½ million or fuenfundzwanzig. Furthermore, the component is able to recognize roman numerals and assign an equivalent numeric value. Written-out numbers are currently only supported in English, German and French.
Measurement Annotator	This component detects units, measurements and quantities. It can trace the given unit back to SI base units and at the same time normalize the numerical value. For example, the text passage 10cm is recognized as 0.1 m (dimension L).
Temporal Expression Annotator	This component can recognize different temporal expressions and normalize their values. This includes simple date formats such as "10.2.2015" or "12:30". The component supports the English and German language.
Named Entity Recognition, Trainable	ML-based (Conditional Random Fields), pre-trained models available for standard newspaper entities (PER, LOC, ORG), but re-trainable for any other entity type. Flexible and configurable w.r.t. to features used.
Concept Disambiguation	Several stages of rules and statistics for disambiguation; which stages to use is configurable; stages include, e.g., making use of additional information given by the concept store (i.e., the terminology) such as attributes or relations. Disambiguation is performed globally for complete document.
Keyword-Extraction, Controlled	focusses on free keywording; several approaches supported: a) based on statistics (TF-IDF), b) based on text coherence (graph algorithm) and c) based on keyword pattern

Keyword-Extraction, Uncontrolled	focusses on controlled keywording, i.e., suggests most informative concepts from a terminology matching the given text at hand; several approaches supported: a) based on statistics (TF-IDF), b) based on text coherence (graph algorithm) and c) based on keyword pattern
Table Format Recognition	rule based, tries to derive table format from file automatically
RUTA (Rule Engine)	UIMA Ruta is a rule-based system designed for information extraction tasks, but it is also applicable for many natural language processing use cases
Drools Annotator	
UIMA Default Annotators (HMM Tagger, BSF Annotator, Alchemy, OpenCalais)	
Evaluation Modules	Allows for automatized performance testing of components: given gold standard annotations (e.g., for entity recognition or keywording), this module compared made annotations with gold standard, marks true positives/false positives and false negatives and calculated standard performance statistics (Precision, Recall, F-Score) on it
CAS Consumer (CC)	
XML Writer	
Lucene CAS Indexer (Lucas)	
Solr CAS Consumer (Solrcas)	
Flow Controller (FC)	
Document Language Flow Controller	flow can be controlled based explicitly set or automatically identified language; allows for language specific pipeline configurations
Document Category Flow Controller	flow can be controlled based on category being explicitly set or automatically identified for given document; allows for topic-specific pipeline configurations

ANNEX D: Training Corpora and Performance Figures of Standard Components

We set up extensive quality and performance for all our components. Usually, we have gold standards available for the components, against which we evaluate our improvements. As an example, we have generated our own or acquired third party corpora for each of our NLP annotators relying on statistical models, which we use for training and evaluation:

Corpora	Description	License
TIGER	German newspaper text (Frankfurter Rundschau).	bought
GENIA	Biomedical text information extraction corpus	free
ELRA Crater	Morpho-syntactically tagged telecommunication manuals. The corpora consists of 1,500,000 tokens for English and French and of 1,000,000 tokens for Spanish, with morpho-syntactical annotations (human-edited).	bought
FRAMED	German clinical documents.	owned
ECI_MCI	A 98 million word corpus, covering most of the major European languages, as well as Turkish, Japanese, Russian, Chinese, and Malay.	bought
MASC	MASC is a balanced subset of 500K words of written texts and transcribed speech drawn primarily from the Open American National Corpus (OANC)	free
REUTERS	The TRC2 corpus comprises 1,800,370 news stories	free for research and development
MULTITEXT-JOC	The corpus contains approx. 5 million words from Written Questions and Answers of the Official Journal of the European Community in English, French, German, Italian and Spanish	bought
PTPAR	PTPARL Corpus contains approximately 975,806 running words of European Portuguese. It includes 1076 texts consisting of adapted transcriptions of the Portuguese parliament sessions, which were made available in 2004	bought

The following table gives a brief overview over the annotation quality of each annotator with respect to these corpora.

SentenceDetector

Corpus	Language	F1-Score
FRAMED	DE	0.960
TIGER	DE	0.970
GENIA	EN	0.986
OANC-State	EN	0.939
MASC-3	EN	0.915
ELRA Crater	FR	0.420
ELRA Crater	ES	0.406
ELRA Crater	PT	0.873
ELRA Crater	IT	0.701

Tokenizer

Corpus	Language	F1-Score
ELRA Crater	FR	0.990
ELRA Crater	ES	0.990
ELRA Crater	PT	0.990
ELRA Crater	IT	0.990
FRAMED	DE	0.997
TIGER	DE	0.999
GENIA	EN	0.998
OANC-State	EN	0.998
MASC-3	EN	0.990

POS-Tagger

Corpus	Language	F1-Score
ELRA Crater	FR	0.971

ELRA Crater	ES	0.992
ELRA Crater	PT	0.970
ELRA Crater	IT	0.949
FRAMED	DE	0.971
TIGER	DE	0.966
GENIA	EN	0.985
OANC-State	EN	0.979
MASC-3	EN	0.952

Chunker

Corpus	Language	F1-Score
FRAMED	DE	0.942
TIGER	DE	0.959
GENIA	EN	0.939
OANC-State	EN	0.960
MASC-3	EN	0.887