ABBYY Smart Classifier

Technical Product Overview

The Challenge of Classifying Unstructured Information

For storing data, searching through documents and databases by keywords or defined expressions, computers are useful tools. If a piece of text contains the search phrases or unique keywords, computer-based rules can easily be written and applied to help organise and manage content assets in databases. However, rules do not work well for unstructured content written in natural language. Documents such as articles, reports, process descriptions, product documentation, letters, and e-mails, can contain information also in free flowing text and may not necessarily use the exact phrases to make computer-based search or classification possible.

Enterprises and organisations have collected hundreds of thousands of documents which carry relevant information – but automatically identifying subject area, department, or process to which each of them belong, is still an IT challenge.

Structures and categories help the human brain to organise the world. In the same way, categorising the content of documents being stored or arriving at enterprises can greatly improve their process efficiency. In order to enable existing IT systems to classify unstructured content as reliable as structured documents, language-based technology that derives classification categories automatically from text is needed.

What is Smart Classifier?

ABBYY Smart Classifier is a scalable server-based classification application for organising unstructured information based on statistics, morphology and semantics. Integrated via REST API this classification application becomes an intelligent component within existing IT systems, workflows, and solutions.

Smart Classifier’s set-up and classification model optimisation is intuitive. The system automatically selects and optimises the classification features and parameters. Consequently, Smart Classifier delivers high-quality classification results without the need for data experts or scientific classification background. Smart Classifier is production-ready and natively supports 39 languages and a large variety of file formats such as text, office document, PDF, and image formats.

Approaching the Unstructured Data Challenge

ABBYY Smart Classifier can “read” a selection of pre-classified documents. During the training phase the documents will be analysed to determine what document content for each particular category has in common. Once the system is trained and deployed, content assets can be sent for classification. Smart Classifier will analyse the new assets by comparing relevant features and content with the documents it was trained on.

The result is a statistical classification decision which is based on the most applicable classes that the document will belong to. The generated metadata from each document can be used to further optimise information and knowledge management systems or business processes. The scope of use cases how classification helps to automate information processes is vast and include: to determine whether a piece of text is confidential or not, add automatically generated tags to documents stored in repositories, route documents to the relevant downstream business process, workflow or department, or enhance search by restricting it to a particular class of records.

HIGHLIGHTS & BENEFITS

Linguistically enhanced text- and semantic-based classification for managing unstructured information and documents

High quality classification results out-of-the-box

Straightforward set-up and creation of classification models with a webbased Model Editor - no "data science" expertise required

Easy set-up and training of new classification models with a web-based Model Editor

Automatic classification algorithm optimisation

Flexibility to address new and changing process needs. Fast creation of new classification “classes” using small sample training sets

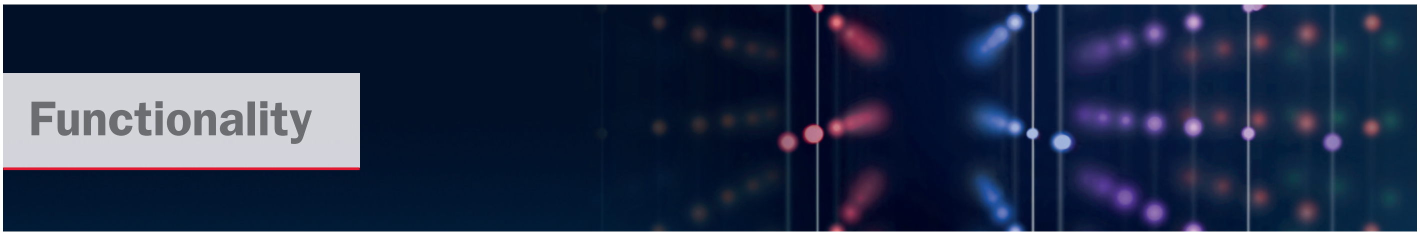
Native support of many file formats such as Office, PDF and image

Embedded optical character recognition (OCR)

Ready for international projects with linguistic classification support for 39 languages

REST APIs for easy integration

Scalable processing backend



Functionality

Getting started

ABBYY Smart Classifier is flexible and capable of processing content from different subject areas, domains and departments.

The system learns automatically based on the training documents. Every organisation has experts with knowledge of the internal documents that are generated or received within that organisation. These people are perfectly suited to train classes in Smart Classifier. No in-depth scientific knowledge is required to generate high classification quality. The system extracts the text from submitted documents and analyses it using the morphology of the language, semantic and textual features, statistics and machine learning.

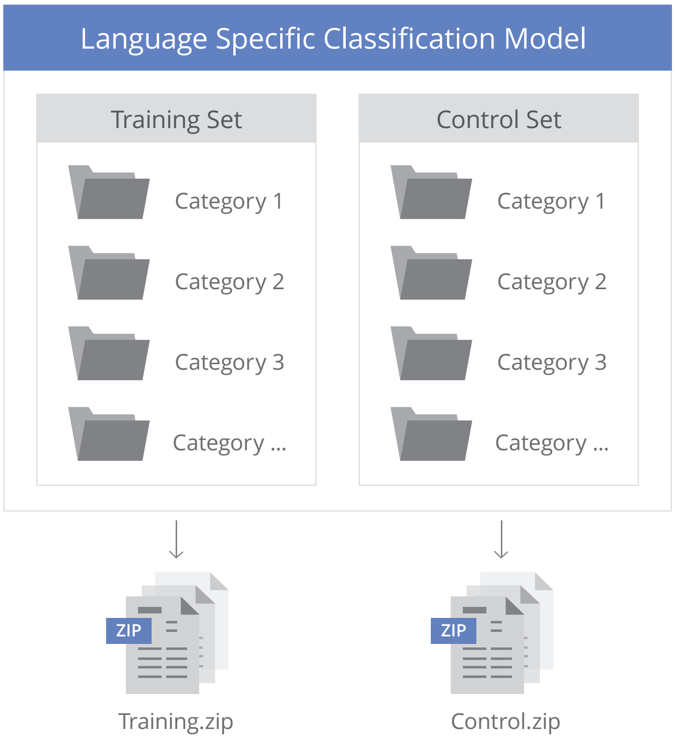
The principle behind machine learning is to identify and use the most relevant features automatically from a set of training documents.

Preparation: Category Definition

Content and process experts within the organisation are perfectly suited to define and train the necessary categories (or classes) in Smart Classifier and select the training documents - prototypical sample documents that represent the categories.

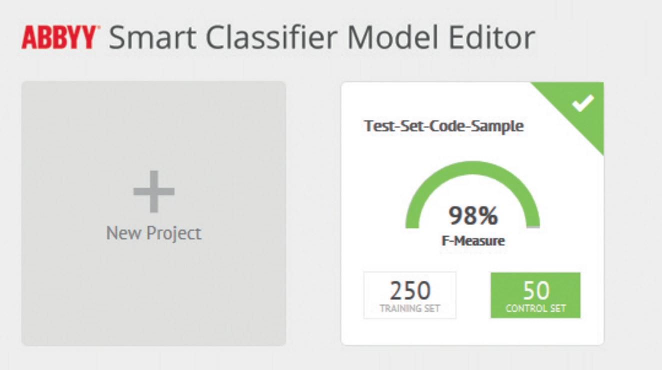
With Smart Classifier, there is no need to use thousands of documents to train a category. The minimum technical number is ten documents for a class, while a set of 100+ documents is recommended to generate reliable statistics. Smart Classifier makes collecting training documents easy since it can natively process a large variety of file formats present in enterprise reality such as plain text, office formats, HTML, etc. Thanks to built-in ABBYY OCR and conversion capabilities, the application can also process PDFs, faxes or scanned documents reliably. The training documents for the different classes should be set up for individual languages and arranged in a folder structure.

Now the training document collection from each category should be divided into a training set and a control set. Once compressed as zip-archives they can be uploaded via the intuitive Model Editor interface to train the system.



System Training: Model Editor

Smart Classifier's Model Editor offers an intuitive and easy-to-use web interface. Classification projects, document training sets and models are created and managed remotely via a browser.



ABBYY Smart Classifier Model Editor

+ New Project

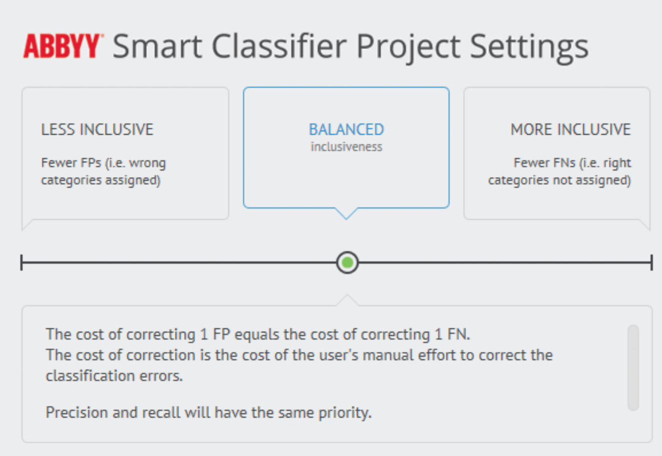
Test-Set-Code-Sample

98%
F-Measure

250 TRAINING SET

50 CONTROL SET

New classification projects can easily be set up by specifying the language for a model language and the inclusiveness of classification. With these settings, the sensitivity of the model is controlled. Depending on the business need, it is possible to adjust the model if the classification should "focus" on a high "true positive" rate or if the model can be more "open" – with a statistical risk of putting too many documents in a particular category (= precision/recall balance).



ABBYY Smart Classifier Project Settings

LESS INCLUSIVE
Fewer FPs (i.e. wrong categories assigned)

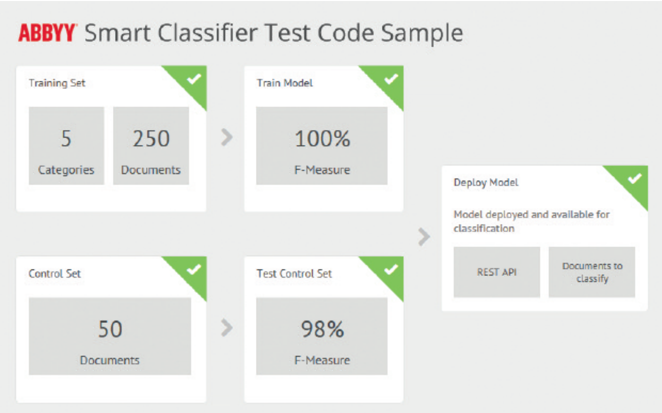
BALANCED
inclusiveness

MORE INCLUSIVE
Fewer FNs (i.e. right categories not assigned)

The cost of correcting 1 FP equals the cost of correcting 1 FN.
The cost of correction is the cost of the user's manual effort to correct the classification errors.

Precision and recall will have the same priority.

The Model Editor provides a status overview for each classification project and gives access to the different components of the workflow including project settings, training sets, control set documents and quality evaluation for each of the classified assets. Unknown or wrongly classified documents can be uploaded to evaluate/debug the classification results and words, even after the model is deployed.



ABBYY Smart Classifier Test Code Sample

Training Set
5 Categories
250 Documents

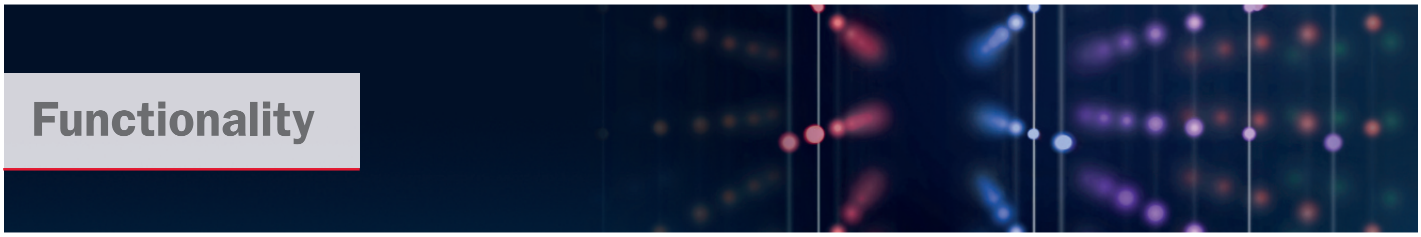
Train Model
100% F-Measure

Control Set
50 Documents

Test Control Set
98% F-Measure

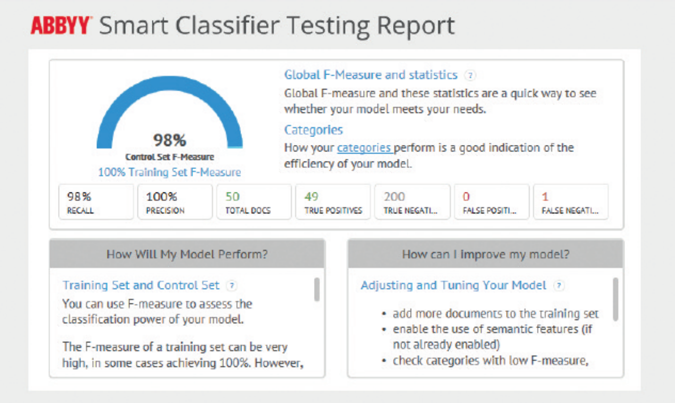
Deploy Model
Model deployed and available for classification
REST API
Documents to classify

Functionality



Quality Evaluation

Smart Classifier makes classification easy because classification model tuning on an algorithmic level is not required. However, since the underlying core technologies are very complex, it is important to review and determine whether the training process was successful and that the classification results in a control set meet expectations before the model is deployed. Performance metrics such as f-measure, precision/recall, true/false positives, and more help to evaluate the classification model quality.



ABBYY Smart Classifier Testing Report

98%
Control Set F-Measure
100% Training Set F-Measure

Global F-Measure and statistics ?
Global F-measure and these statistics are a quick way to see whether your model meets your needs.

Categories
How your categories perform is a good indication of the efficiency of your model.

98%	100%	50	49	200	0	1
RECALL	PRECISION	TOTAL DOCS	TRUE POSITIVES	TRUE NEGATI...	FALSE POSITI...	FALSE NEGATI...

How Will My Model Perform?

Training Set and Control Set ?
You can use F-measure to assess the classification power of your model.

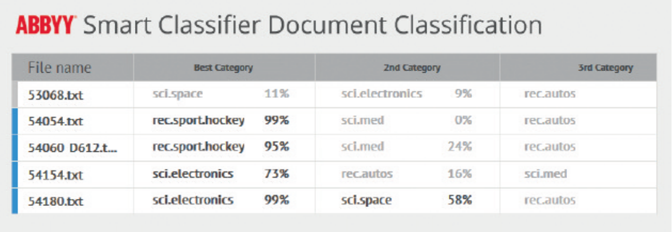
The F-measure of a training set can be very high, in some cases achieving 100%. However,

How can I improve my model?

Adjusting and Tuning Your Model ?

- add more documents to the training set
- enable the use of semantic features (if not already enabled)
- check categories with low F-measure,

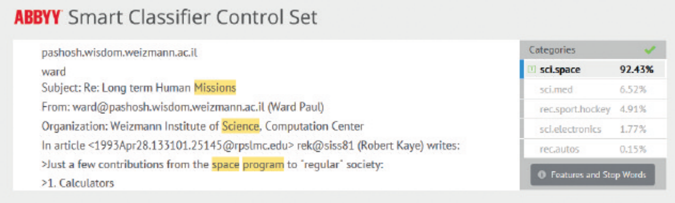
The classification results of the documents used in the training and control sets can be checked and re-assigned if the category does not match.



ABBYY Smart Classifier Document Classification

File name	Best Category		2nd Category		3rd Category
53068.txt	sci.space	11%	sci.electronics	9%	rec.autos
54054.txt	rec.sport.hockey	99%	sci.med	0%	rec.autos
54060 D612.t...	rec.sport.hockey	95%	sci.med	24%	rec.autos
54154.txt	sci.electronics	73%	rec.autos	16%	sci.med
54180.txt	sci.electronics	99%	sci.space	58%	rec.autos

The Model Editor also provides instant visibility of each document within a classification project. Source text and key words picked by the algorithms can be analysed and checked. Terms that should be ignored during classification can be added to a stop word list.



ABBYY Smart Classifier Control Set

pashosh.wisdom.weizmann.ac.il
ward
Subject: Re: Long term Human Missions
From: ward@pashosh.wisdom.weizmann.ac.il (Ward Paul)
Organization: Weizmann Institute of Science, Computation Center
In article <1993Apr28.133101.25145@rpslmc.edu> rek@siss81 (Robert Kaye) writes:
>Just a few contributions from the space program to "regular" society:
>1. Calculators

Categories	✓
sci.space	92.43%
sci.med	6.52%
rec.sport.hockey	4.91%
sci.electronics	1.77%
rec.autos	0.15%

Architecture & Integration

Simple Setup & Configuration

Smart Classifier is built on the ABBYY Compreno processing backend. All necessary components, including Microsoft Internet Information Server (IIS), can be automatically configured during the installation process. IIS hosts the web based Model Editor and allows integration via REST API.

Backend Architecture & Scalability

Smart Classifier is based on a scalable backend, capable of processing large amounts of files. The system will extract plain text out of content assets submitted for classification; for images or PDFs, optical character recognition can be automatically applied. For a high throughput, Smart Classifier can be scaled up with additional processing stations running OCR processes.

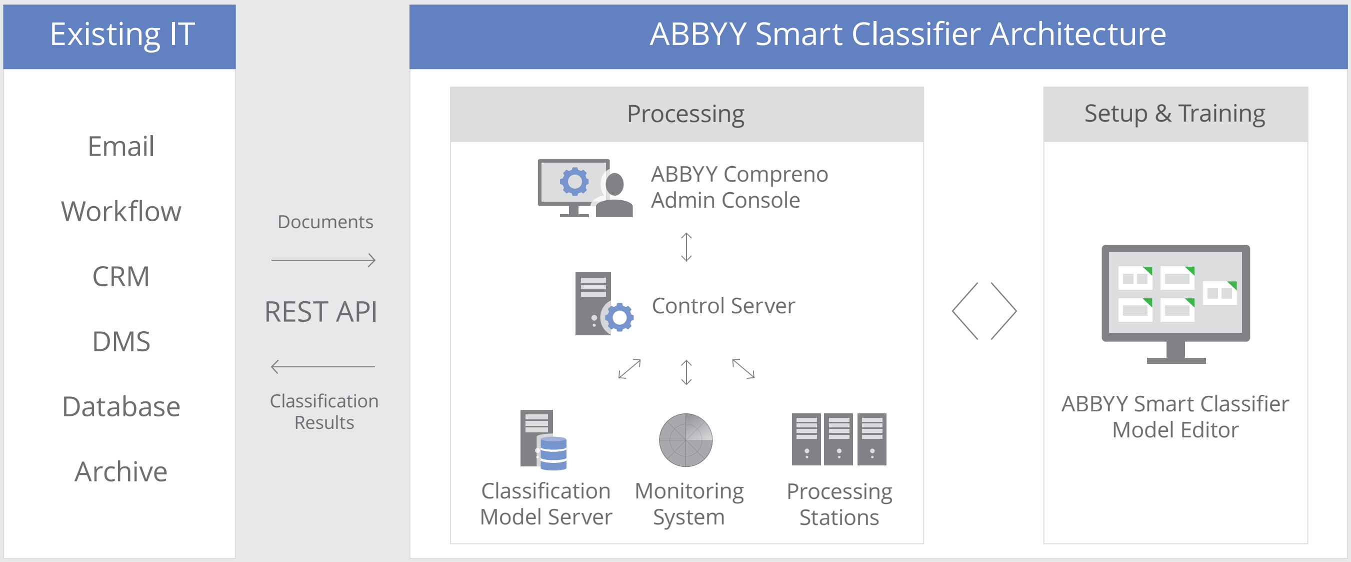
Integration via REST API

Smart Classifier can be easily integrated into document workflows, archiving, records management, e-mail management, or data migration systems via its RESTful API.

Once the system is set up and a classification model is published for operation, incoming classification tasks will be accepted. Depending on the amount and complexity of tasks they can be submitted either in synchronous and asynchronous modes.

Classification Results

Smart Classifier can return the classification results in JSON or RDF/XML. The results include information such as name of the classification model, categories with their probabilities, confidentiality flags, feature/word lists, access to the raw text** or error messages.



Existing IT

Email
Workflow
CRM
DMS
Database
Archive

Documents →
REST API
← Classification Results

ABBYY Smart Classifier Architecture

Processing
ABBYY Compreno Admin Console
Control Server
Classification Model Server
Monitoring System
Processing Stations

Setup & Training
ABBYY Smart Classifier Model Editor

Specifications and Licensing

SYSTEM REQUIREMENTS

System requirements

- PC with 1.5 GHz or faster 32-bit (x86) or 64-bit (x64); 4 CPU cores
- Operating systems (32 & 64-bit): Microsoft® Windows® 7, 8, 8.1, 10 Microsoft® Windows Server® 2008 R2, Windows Server® 2012, Windows Server® 2012 R2,

  The operating system must support the selected interface language.
- Memory: Min. 4 GB RAM, depending on the volume processed.
- Hard-disk space: 2 GB for product installation. The disk space required for product depends on the volume processed.
- Microsoft® .NET Framework 4.5 Full Profile, provided in the installer
- MSXML 6.0, will automatically be installed
- Microsoft® Internet Information Services 7.0 or later

Model Editor (web access)

- Internet Explorer 11.0 or later
- Google Chrome 40.0 or later
- Firefox 36.0 or later
- Microsoft Edge 20 or later

Input formats

- Document Formats: DOC, DOCX, DOCM, HTML, MIF, ODT, ODP, ODS, PDF*, PPT, PPTX, RTF, TXT,XLS, XLSX
- Image Formats*: PDF, JPG, BMP, GIF, TIF, PNG DJVU, DCX, DIB, JB2, JP2, JPF, PCX, WDP

Output Formats

JSON, RDF/XML, Raw-Text of the classified document**

Integration

REST API

Textual/ Linguistic Classification

Supports 39 languages, for a full list see www.abbyy.technology

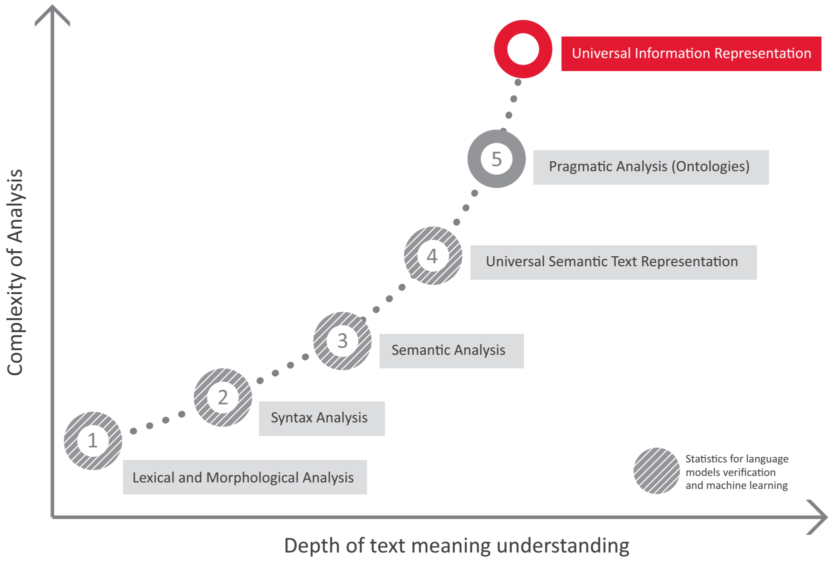
Semantic Classification Languages

English, Russian

* OCR is a license add-on
** Raw-Text-Access is a license add-on

ABBYY Compreno, Smart Classifier & InfoExtractor

ABBYY Compreno natural language processing (NLP) technology enables businesses to understand unstructured information. Its intelligent technology understands the meaning of words and defines relationships between them. Based on these relationships, it creates semantic representations that enables text to be analysed by computers for accurate information extraction, classification and intelligent search. This language-based approach for analysing unstructured information creates new opportunities to action information and to optimise critical business processes where rule-based approaches fail.


Universal Information Representation
5
Pragmatic Analysis (Ontologies)
4
Universal Semantic Text Representation
3
Semantic Analysis
2
Syntax Analysis
1
Lexical and Morphological Analysis
Statistics for language models verification and machine learning
Complexity of Analysis
Depth of text meaning understanding

ABBYY Smart Classifier organises unstructured information in existing or incoming documents. It extracts and analyses text with linguistic and statistical methods and derives the highest probability for the best matching classes.

ABBYY InfoExtractor uses a semantic based approach to extract relevant information. The system can identify not only entities and facts, but also the relationships between them. The technology is currently available for English and Russian language.

Licensing

Smart Classifier is available for testing via time and volume limited trial licences.

ABBYY offers 3 different licensing models for Smart Classifier:

- perpetual licences with software maintenance;
- yearly subscriptions and;
- OEM licensing models for software vendors.

The standard license model is based on a renewable peak volume. The back-end can be scaled up as it is needed; the number of processing stations/CPU cores is not limited.

ABBYY®

ABBYY Europe GmbH
Elsenheimerstr. 49
80687 Munich, Germany
Tel: +49 89 69 33 33 0
sales_eu@abbyy.com
www.ABBYY.com

Regional Contacts

UK & Ireland
sales_uk@abbyy.com

France
sales_france@abbyy.com

Spain & Portugal
sales_es@abbyy.com

Italy
sales_it@abbyy.com

Benelux
sales_benelux@abbyy.com

Nordics
sales_nordic@abbyy.com

©2016 ABBYY. All rights reserved. ABBYY, the ABBYY logo and ABBYY Compreno are either registered trademarks or trademarks of ABBYY Software Ltd. All other trademarks are the property of their respective owners.