Document Classification
The intelligent way to organize information

Documents exist in many different types and layouts and it is a common task to identify and categorize them based on their diverse contents. The Document Classifier module of OmniPage Capture SDK 20 enables users to separate different documents and sort them based on layout and textual similarities. This intelligent and trainable technology can be used in a wide range of business processes, such as email forwarding, mail-room automation, spam filtering, forms processing and data extraction.

Using the components of the Document Classifier module, users can set up various classes for the different type of documents, train the classifier with typical documents for each class, then test and fine-tune the classifier. The comprehensive data and intelligence collected during this process can be exported and used together with OmniPage SDK in a target application that is designed to execute classification tasks within its multi-step workflow process. This supervised learning model enables great flexibilty and high accuarcy for properly categorizing a vast number of incoming documents and information in a software solution.

Document Classifier Assistant: The intelligence to know your documents
The Document Classifier Assistant is a separate applicaton that provides users with the tools and convenient user interface to create the required document classes. Building on the capabilities of OmniPage SDK and by processing sample documents, it develops the knowledge needed to sort the different documents into those classes and saves the knowledge in the classifier data. Classes in a usual business operation, for example, business letters, invoices, legal contacts and forms, can be defined by the user.

Steps to prepare the classifier project

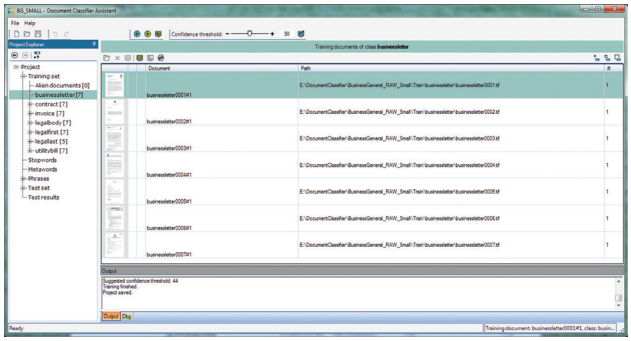
– **Assemble the training document sets:** Collect several documents that represent common characteristics for each class that you want to set up. These documents can be in the format of popular image files, such as JPG, TIFF, PNG, or PDF or TXT files.

– **Set up classes for the collected documents:** Create classes with easily identifiable names (such as Order forms, Messages, Invoices, Reports, etc.) and assign your documents to them as training samples. Documents in the same class should share similar textual and/or layout characteristics.

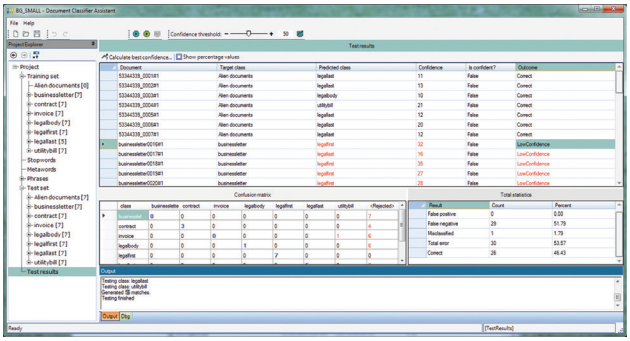
Key benefits

– Automate business processes that require document and information sorting
– Reduce costs accociated with manual pre-processing of documents
– Provide higher value to your customers
– Utilize the world's most accurate and proven OCR SDK for reliable document categorization
– Conveniently create classifier data for your custom application
– Control the classification scheme creation project for best results

NUANCE



Easily create classes and train the classification scheme with a set of sample documents featuring the characteristics of each class.



Fine-tune classification accuracy based on the comprehensive statistics available from the test results.

– **Train your system:** Teach the classifier the unique characteristics of each class from the set of documents you assigned to that class. This is an automated process that takes advantage of the advanced pattern recognition algorithms of the underlying OmniPage SDK.

– **Test your classificaton scheme:** Feed test files to the system to check the accuracy of the classification scheme developed at the training phase. The system returns so called confidence values for each document and based on the highest value it assigns it to a class. You can verify whether or not the test documents are associated with the right class.

– **Fine-tune the classification scheme:** Based on the results of the test, you can re-train your classificaton project by adding more training sample files to each class and/or change the confidence threshold value which is used by the system to decide if a test document is assigned to a class or gets rejected.

– **Export the classifier project data:** Once you are satisfied with the test results, export the data of the classificaton project for use in the target applicaton.

The Document Classifier Assistant provides additional capabilties to optimize the classification accuracy. Users can define phrases that are commonly used in documents that fall into a particular class, e.g. “Yours sincerely” for the “Business letters” class. In addition, they can expand the built-in list of “stop-words” that include generic words, such as “the” or “and” which are not relevant for building the characteristics of any class in the classifier project.

Utilizing the Document Classification API

The OmniPage Capture SDK includes a convenient API that enables users to load the classifier project data exported from the Document Classifier Assistant into the OmnPage SDK. Equipped with this classifier data, it can execute the required classification tasks when integrated into the target solution. The API offers calls to import the documents for analysis and classification. Also, it returns the classification results to the application that include the assigned class together with the corresponding confidence value for each document.

The OmniPage Capture SDK with its Document Classifier Assistant and API provide developers and system integrators all the capabilties to create and integrate flexible and productive classification projects into their application with low effort.

To learn more about Nuance document imaging solutions please call 1-800-327-0183 or visit nuance.com

About Nuance Communications, Inc.

Nuance Communications, Inc. is a leading provider of voice and language solutions for businesses and consumers around the world. Its technologies, applications and services make the user experience more compelling by transforming the way people interact with devices and systems. Every day, millions of users and thousands of businesses experience Nuance’s proven applications. For more information, please visit: www.nuance.com.

NDI_250 JUL 2016