EXTERNAL

SAP AI Business Services
Blueprint Guide for Document Classification

THE BEST RUN

TABLE OF CONTENTS

INTRODUCTION

The purpose of this document is to describe the general value proposition of Document Classification. Therefore, three distinctive business scenarios are described, and a basic solution architecture is depicted. Furthermore, the document guides through the technical setup of the service, including the activation and consumption of the service but also the steps specific to artificial intelligence.

BUSINESS SCENARIO

Three business problems can be identified and solved using Document Classification that all aim to reduce manual efforts and errors in the classification of documents while speeding up the document processing overall:

- **Scenario 1 – From Chaos to Order:** The first scenario addresses enterprises that struggle with the organization and management of large volumes of documents. Such large amounts of unorganized documents can prevent the workforce to work efficiently.
- **Scenario 2 – Enterprise Mail-Inbox:** The second scenario addresses enterprises that deal with large amounts of business documents attached to emails, coming from their customers and partners. The processing of these business documents is often slow, manual and error-prone so that speed and quality of business-critical processes suffer.
- **Scenario 3 – Document Filtering:** The third scenario addresses organizations that need to detect and filter out certain critical documents from a large pool of similar documents. The manual search for critical documents in a pool of unstructured documents is a costly, time consuming and error-prone process.

BUSINESS SERVICE DEFINITION

Document Classification is one out of a set of services for document processing that aim to process unstructured documents into structured information. Document Classification offers automatic and customer-specific classification of documents.

The service analyzes PDF documents and proposes classifications based on previously defined criteria and categories. To do so, the service includes training and inference capabilities to fit a model using a custom dataset.

Thus, Document Classification acts in:

- **Scenario 1 – From Chaos to Order:** Document Classification can help with the organization and classification of large volumes of documents. Therefore, Document Classification applies machine learning models to categorize and classify documents into customer specific document types and classification schemas. By applying automation to the classification of documents, manual work is reduced, the process is accelerated, and the data quality is improved.
- **Scenario 2 – Enterprise Mail-Inbox:** With the help of Document Classification critical documents can be identified within a large volume of documents. Furthermore, the documents are categorized into customer specific document types and classification schemas. Consequently, repetitive tasks are minimized and less time bottlenecks in the processing of documents occur. This frees up time for value-creative tasks and results in faster and more accurate responses to customers.
- **Scenario 3 – Document Filtering:** Document Classification can help with the detection of critical documents in large piles of documents. By applying machine learning models, these documents are filtered out according to customer-specific document types. Thus, the service helps to reduce manual work and classification errors while simultaneously speeding up the classification and identification process.

Data Classification is available as SAP AI Business Service via the Cloud Platform Enterprise Agreement on SAP Business Technology Platform (SAP BTP). Furthermore, the service is already integrated in SAP S/4HANA Document Management and can be integrated with SAP Intelligent Robotic Process Automation.

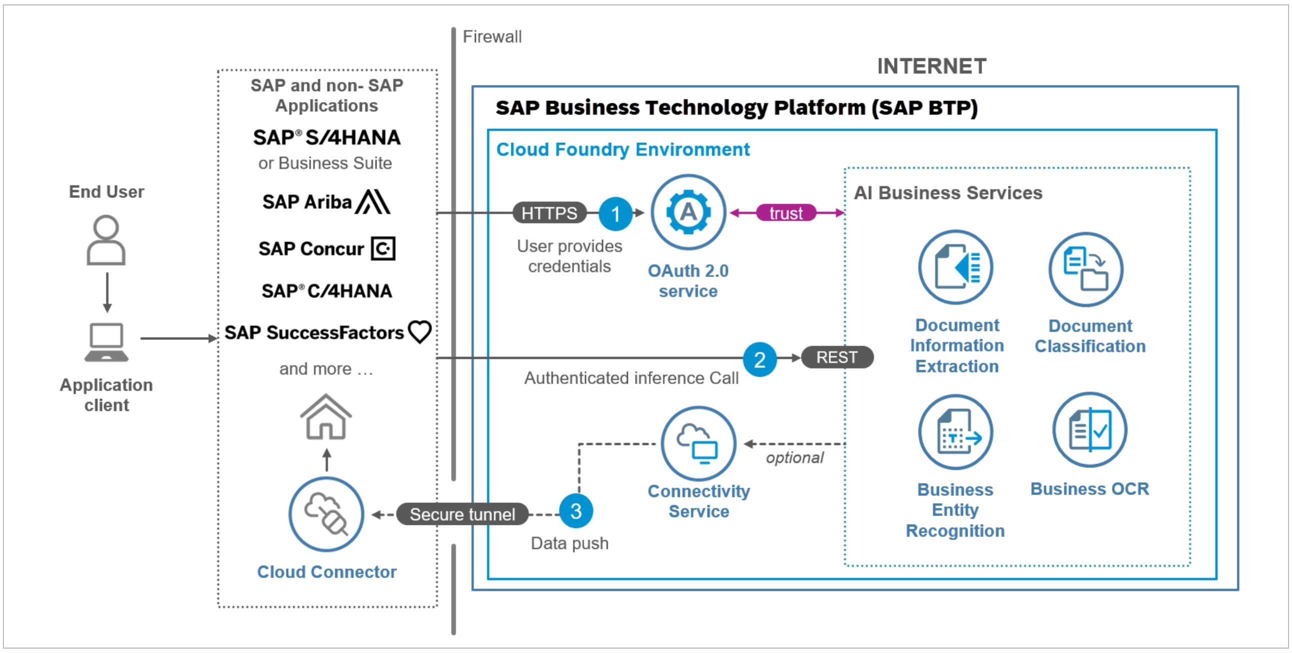
VALUE PROPOSITION

Document Classification enables companies that manage large numbers of business documents to automatically and easily classify documents based on customer-specific classification schemas.

The main benefits of Document Classification include but are not limited to:

- Increased automation of document processing resulting in increased efficiency and reduced errors in the process
- Improved accuracy during the document processing than traditional rule-based methods or manual tasks
- Accelerated document processing and consequently increased customer satisfaction by enabling faster and more reliable responses
- Reduced costs by automating the document processing and thus allowing for higher throughput and quality during the process

SOLUTION ARCHITECTURE



The above schematic diagram provides a general solution architecture of the SAP AI Business Services and, thus, of Document Classification. The service is available on SAP BTP and can be connected and integrated with SAP and non-SAP applications.

To communicate with a service instance of Document Classification, applications can use the **REST API** that the services provides. The **OAuth 2.0** protocol, the standard user authentication and authorization mechanism provided by SAP BTP for Cloud Foundry, is used to secure the communication between the client and the service instance. Thus, the client needs to authenticate itself in order to communicate with the service. Therefore, the client can request an authentication token by providing the credentials of the service instance.

The REST API enables all applications and products to incorporate Document Classification. The whole communication throughout the lifecycle takes place via this interface. This includes the creation and deletion of datasets, the upload of documents, the training and deployment of machine learning models and inference requests (i.e. making predictions).

TECHNICAL SETUP

Please note that the tutorials linked below were created for the service Document Information Extraction. Nevertheless, the same procedure on how to create a service instance applies to Document Classification as well.

Activation of the Business Service

Document Classification is available in the Cloud Foundry environment of SAP BTP.

To activate the service, the correct entitlements must be assigned to your account in the SAP BTP. The entitlements define which service and runtime artifacts the account is able to use. Please follow step 2 in this tutorial that guides through the activation of a service in the Cloud Foundry environment.

In order to use Document Classification, an instance of the service is needed. Please follow steps 4 and 5 in this tutorial that guides through the creation of a service instance.

Business Service API Key/Secret

In order to communicate with the service instance, service keys need to be created and attached to the communication with the service instance.

Please follow step 6 in this tutorial that guides through the creation of service keys for a service instance.

Business Service Consumption

As outlined in the solution architecture, the service provides a REST API for communication. As REST APIs make use of the HTTP protocol the service can be easily integrated into any application and technology stack. Most programming language already have integrated capabilities for HTTP or have several libraries available in its ecosystem.

DATA REQUIREMENTS

Data Protection and Privacy

Data protection is associated with numerous legal requirements and privacy concerns. In addition to compliance with general data privacy acts, it is necessary to consider compliance with industry-specific legislation in different countries.

The data used by Document Classification is controlled and managed by the consuming application/customer which calls the service APIs. Document Classification does not have any means to verify whether the data uploaded to the service contains any personal information.

According to Personal Data Processing Agreement for SAP Cloud Services, SAP acts as data processor. Thus, customers are responsible for obtaining relevant consent to process personal data, including when applicable approval by controllers to use SAP as a processor.

For further details and information about data protection and privacy, please check the Data Protection and Privacy section in the SAP Help Page, and the document Personal Data Processing Agreement for SAP Cloud Services.

Data Preprocessing

Data Format

The general data format for both uploading training data and inference is the PDF format. For the document size limit and other intputs limits, please review the SAP Help Page.

For training, additional data labeling is required to classify the training data. For further details, please see below.

Data Labeling

Data labeling is used to label your training documents with the correct classes. That way, a machine learning model is trained from the correct classified documents.

To label a document, a file in JSON (JavaScript Object Notation)-Format needs to be created with the same name as the PDF document that shall be labeled. For example, to label a document called *“TV_manual_123.pdf”*, a corresponding JSON-File needs to be created called *“TV_manual_123.json”*. Please review the chapter “Classification Model Types” to see what the JSON-File would look like for different classification model types.

TRAINING AND INFERENCE

Classification Model Types

The model schema is defined according to classification scenario that should be used. It can be distinguished between five scenarios:

- **Single characteristic, binary classification:** In the simplest scenario all documents are classified into one out of two classes, e.g. *“Suitable for children”* and *“Not suitable for children”* (review **Figure 1**). Both classes are mutually exclusive so that a document cannot be classified into both classes at the same time.
- **Single characteristic, multi-class classification:** In a multi-class classification documents are classified into one out of multiple classes. Consider the example of language detection so that documents can be classified into *“Englisch”*, *“German”*, *“French”* or *“None of the above”* (review Figure **2**). As for binary classification these classes are mutually exclusive, i.e. a document cannot be classified into *“Englisch”* and *“German”* at the same time.
- **Two characteristics, each binary classification:** In this scenario documents are examined by two characteristics and for each characteristic a binary classification exists. Consider the scenario where a document is first classified by *“Suitable for children”* or *“Not suitable for children”* and second, is classified by *“Document includes images”* and *“Document does not include images”* (review Figure **3**). This result in four cases a document can fall in.
- **Single characteristic, multi-label classification:** In a multi-label classification a document is classified by multiple classes which are not mutually exclusive. For example, a document can be classified by *“Document content contains animals”*, *“Document content contains humans”* and *“Document content contains plants”* (review Figure **4**). Thus, a document can contain neither animals, humans nor plants, i.e. none of the above classes applies, but the document can also contain all three of them, i.e. all three classes apply.
- **Two characteristics, one multi-label and one binary classification:** In this more complex scenario multi-label classification and binary classification are combined. For the multi-label classification, recall the example of animals, humans and plants from the above scenario. For the binary classification consider the example that documents are classified by *“Document includes images”* and *“Document does not include images”* (review Figure **5**). This results in 4 x 2 = 8 cases a document can fall in.

Training Process

The training process starts with the creation of a dataset and the subsequent upload of data into this dataset. The uploaded data will then be validated by the service. Once done, the training of a machine learning model can be initiated based on the uploaded data. To use the trained model productively, it should be deployed subsequently. The model is now able to serve inference calls and, thus, can classify documents.

Inference Process

To classify documents and, thus, to make an inference call, upload a new PDF document that shall be classified. As inference process is asynchronous, the upload will not return the classification result. Instead, it is necessary to poll for the result using a different endpoint. Once the classification is done, this endpoint will return the result.

Deployment and Un-deployment

To use a trained model for inference, it is necessary to deploy the model. Once a model is deployed and, thus, made productively available, you will be charged for it. One model is included for free; every additional model will be charged accordingly.

Consequently, unused models should be undeployed to avoid unnecessary cost. Undeployed models that are not used in a period of time are also cleaned up.

Sample Coding

A Python client library for Document Classification can be found on GitHub. The client library provides convenient methods to access the service that issue calls to the REST API of the service. To use and understand the client library better, please review its API documentation.

Additionally, the client library contains examples in form of Jupyter Notebooks and sample datasets. Please find them here.

MONITORING AND TROUBLESHOOTING

In case of an incident or an error, use the SAP Support Portal to get help. Please visit this help page that explains in detail how to use the SAP Support Portal.

When creating an incident in the SAP Support Portal related to the Document Classification service, make sure to use the following component:

Component Name	Component Description
CA-ML-BDP	Services related to Business Document Processing

Additionally, please provide the following information in the incident description:

- Region information (Canary, EU10, US10)
- Subaccount technical name
- URL of the page where the incident or error occurs
- Steps or clicks used to replicate the error
- Screenshots, videos, or the code entered

ATTACHMENTS

Data Labeling and Classification Model Types

Single Characteristic, Binary Classification

```
{
  "classification": [
    {
      "characteristic": "SuitableForChildren",
      "value": "yes"
    }]
}
```

```
{
  "classification": [
    {
      "characteristic": "SuitableForChildren",
      "value": "no"
    }]
}
```

Figure 1: Data labeling for single characteristic, binary classification.

Single Characterisitc, Multi-Class Classification

```
{
  "classification": [
    {
      "characteristic":
        "Language",
      "value": "English“
    }]
}
```

```
{
  "classification": [
    {
      "characteristic":
        "Language",
      "value": "German"
    }]
}
```

```
{
  "classification": [
    {
      "characteristic":
        "Language",
      "value": "None"
    }]
}
```

Figure 2: Data labeling for single characteristic, multi-class classification.

Two Characteristics, Each Binary

```
{
  "classification": [
    {
      "characteristic": "SuitableForChildren",
      "value": "yes"
    },
    {
      "characteristic": "IncludesImages",
      "value": "no"
    }]
}
```

```
{
  "classification": [
    {
      "characteristic": "SuitableForChildren",
      "value": "no"
    },
    {
      "characteristic": "IncludesImages",
      "value": "no"
    }]
}
```

Figure 3: Data labeling for two characteristics, each binary.

Single Characteristic, Multi-Label Classification

```
{
  "classification": [
    {
      "characteristic": "Language",
      "values": ["animals", "plants"]
    }]
}
```

```
{
  "classification": [
    {
      "characteristic": "Language",
      "values": ["animals", "humans", "plants"]
    }]
}
```

Figure 4: Data labeling for single characteristic, multi-label classification.

Two Characteristics, One Multi-Label and One Binary Classification

```
{
  "classification": [
    {
      "characteristic": "Contains",
      "values": []
    },
    {
      "characteristic": "IncludesImages",
      "value": "no"
    }]
}
```

```
{
  "classification": [
    {
      "characteristic": "Contains",
      "values": ["animals", "plants"]
    },
    {
      "characteristic": "IncludesImages",
      "value": "yes"
    }]
}
```

Figure 5: Data labeling for two characteristics, one multi-label and one binary classification.

www.sap.com/contactsap

© 2018 SAP SE or an SAP affiliate company. All rights reserved.
No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies. See www.sap.com/copyright for additional trademark information and notices.

THE BEST RUN SAP®