# Document Conversion
# 1.0版

2016/09/01

SoftBank | IBM

# 変更履歴

| 版数 | 適用日 | 変更内容 |
|---|---|---|
| 1.0 | 2016年9月1日 | 初版 |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

# はじめに

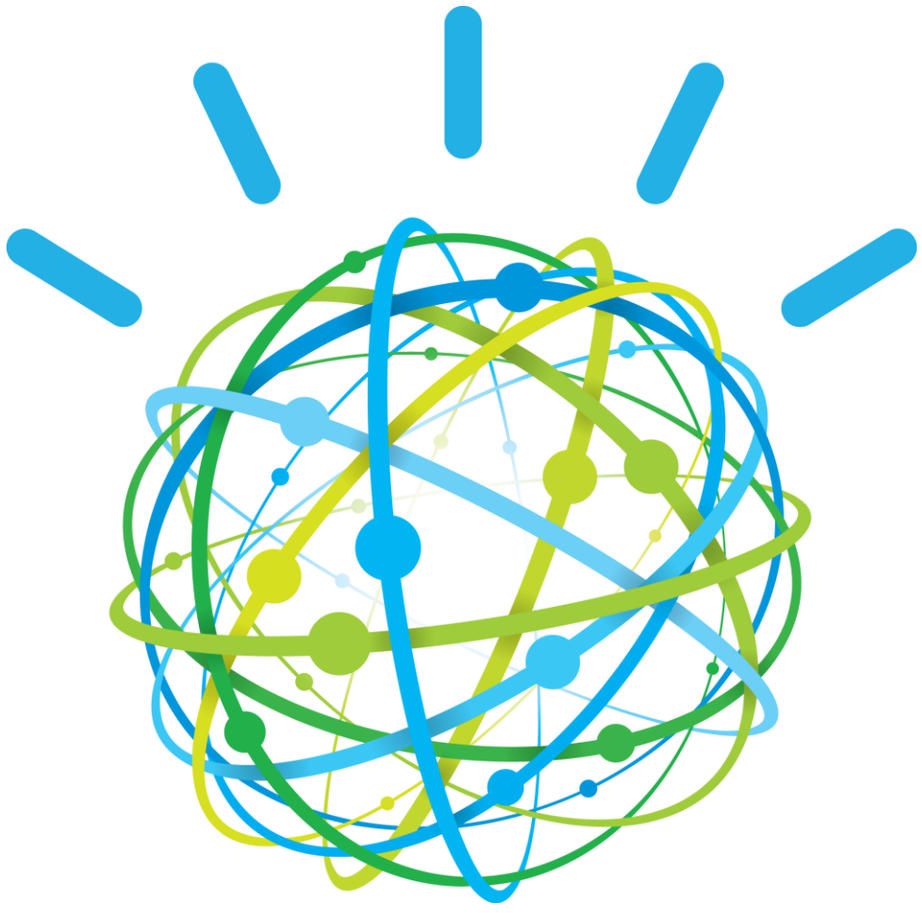
本資料は、ソフトバンク株式会社が、米国IBMの日本法人である日本アイ・ビー・エム株式会社(以下「IBM」という。)との提携により日本語化を行った、IBM Watson(以下「Watson」という。)のサービスのひとつである、Document Converterについて説明しています。

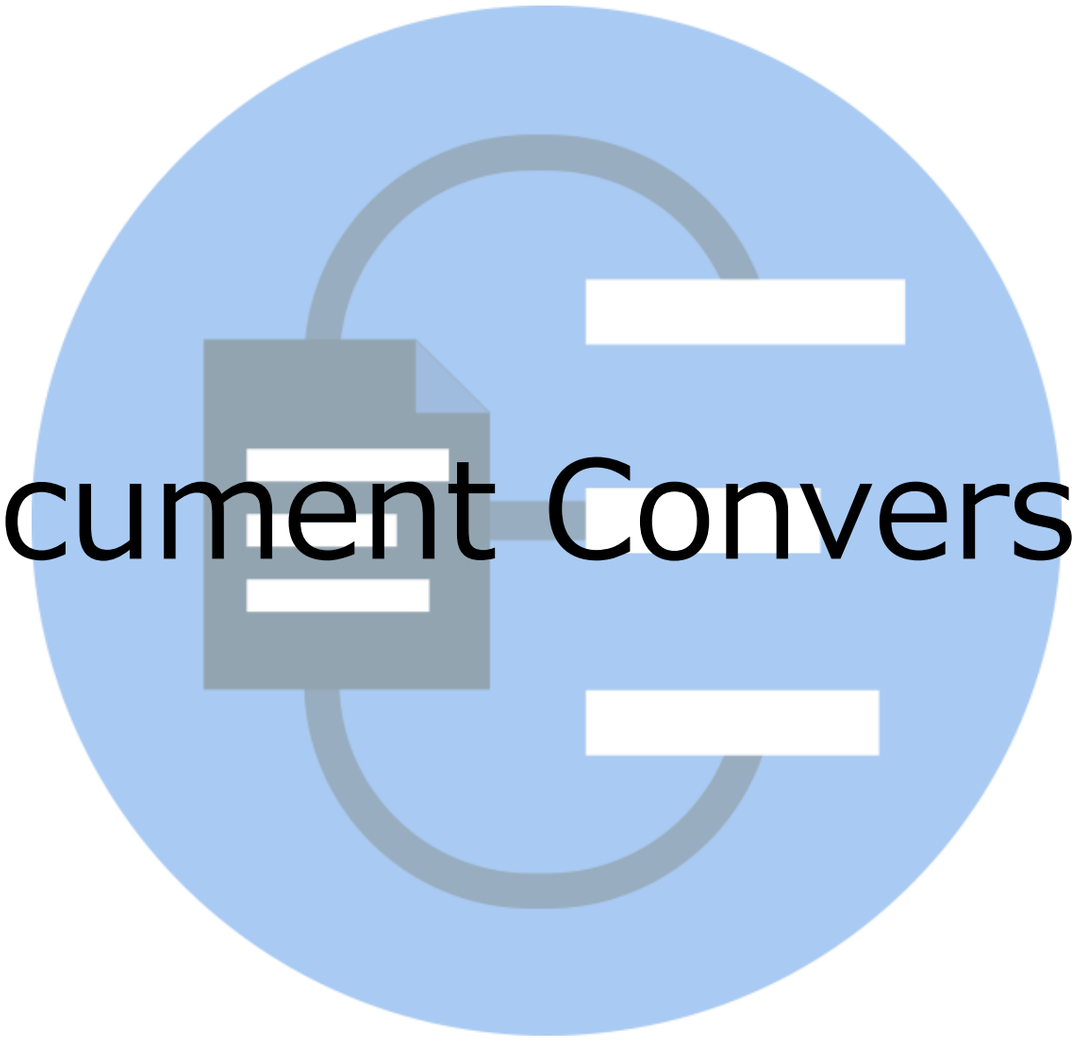
本資料とWatson Developer Cloudに記載の内容と齟齬がある場合は、特段の記載がない限りWatson Developer Cloudの内容を優先とします。

本資料に記載の内容は予告なく変更、また追記・削除されることがありますのでご了承ください。

# もくじ







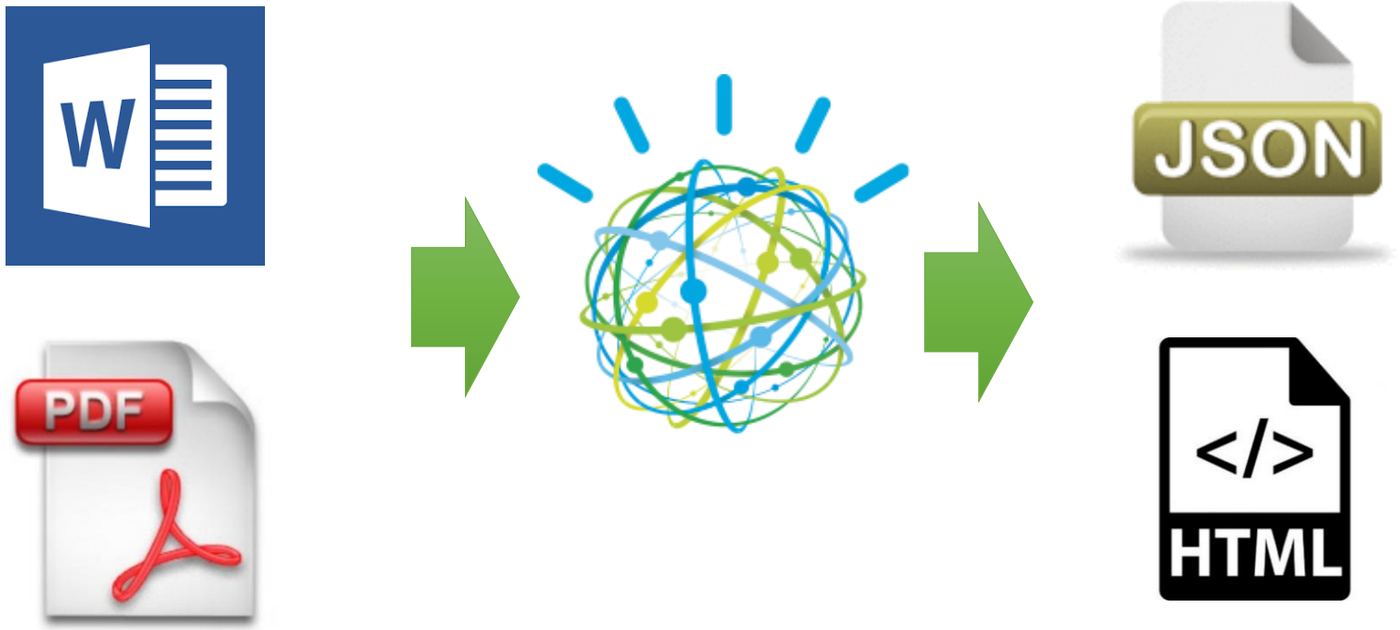
# 1. Document Conversionとは

# Document Conversion

IBM Watson™

## テキスト文書を指定したフォーマットに変換します。

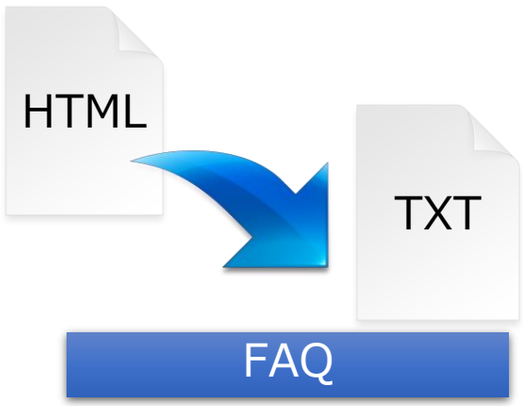
HTML, PDF, Microsoft Word™ 文書を、正規化されたHTML, プレーン・テキスト, JSON形式設定されたAnswerユニット・セットに変換し、他の Watsonサービスで使用できるようにします。



# 想定されるユースケース

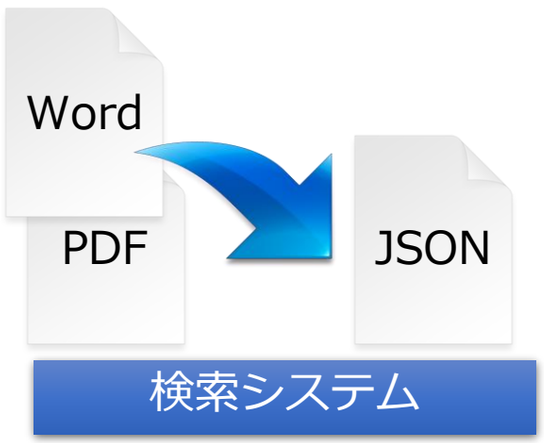
IBM Watson™

運用性の向上だけでなく、散在していたナレッジを統合的に
活用することが可能です。




## HTMLからテキストに変換

HTMLとして存在するFAQ情報
→NLCの学習データに利用




## WordやPDFからJSONに変換

WordやPDFとして存在する技術文書
→R&Rの学習データに利用

# DoCのAPI一覧

IBM Watson™

DoCのAPIでは言語(ロケール)に依存する設定はありません。

| 分類 | メソッド | URL | 詳細 |
|---|---|---|---|
| 変換 | POST | /v1/convert_document | ドキュメントを変換する |

## リクエストパラメータ

| 項目 | 説明 | 種別 | 形式 |
|---|---|---|---|
| config | 変換設定 | multipart/form-data | JSON |
| file | 入力ファイル | multipart/form-data | 50Mbyte |
| type | 入力ファイルの種別<br>(オプション) | multipart/form-data | text/html', 'text/xhtml+xml',<br>'application/pdf',<br>'application/msword'. |
| version | APIバージョン | query | YYYY-MM-DD |

# APIの利用イメージ1

IBM Watson™

## HTMLからscript要素を除去する

入力HTMLファイル

```
<html>
<body>
  <h1>1. 通信利用動向調査</h1>
  <p>今年度の情報通信 …</p>
  <script>
    function x() { none; }
  </script>
</body>
</html>
```

出力HTMLファイル

```
<html>
<body>
  <h1>1. 通信利用動向調査</h1>
  <p>今年度の情報通信 …</p>


</body>
</html>
```

リクエストパラメータconfig(JSON)

```
{ “conversion_target":"NORMALIZED_HTML",
  “normalize_html”: {
    “exclude_tags_completely”: [ “script” ]
  }
}
```

# APIの利用イメージ2

IBM Watson™

## PDFをテキストに変換する

### 入力PDFファイル

1．通信利用動向調査
今年度の情報通信サービスの利用状況等について調査した通信利用動向調査の結果を取りまとめました。

1.1 インターネット等の普及状況
インターネットの普及状況について、利用状況や端末について調査しました。

***1.1.1*** インターネットの利用状況 (個人)
インターネット利用者の割合を男女別にみると、男性は 86.3%、女性は 79.4%となり、とも に前年から…

1.1.2　端末別インターネットの利用状況 (個人)
平成 26 年の 1 年間のインターネットの利用状況を端末別にみると、「自宅のパソコン」 が 53.5%と最も多く…

### 出力テキストファイル

1. 通信利用動向調査結果

今年度の情報通信サービスの利用状況等について調査した通信利用動向調査の結果を取りまとめました。

1.1. インターネット等の普及状況

インターネットの普及状況について、利用状況や端末について調査しました。

1.1.1. インターネットの利用状況 (個人)

インターネット利用者の割合を男女別にみると、男性は 86.3%、女性は 79.4%となり、とも に前年から…

1.1.2　端末別インターネットの利用状況 (個人)

平成 26 年の 1 年間のインターネットの利用状況を端末別にみると、「自宅のパソコン」 が 53.5%と最も多く…

### リクエストパラメータconfig(JSON)

```
{ “conversion_target":"NORMALIZED_TEXT",
}
```

# APIの利用イメージ3

IBM Watson™

## Wordファイルの見出しレベル３をイタリックフォントのみ対象にして、JSONに見出しレベル2で分割して出力する

### 入力Wordファイル

1．通信利用動向調査
今年度の情報通信サービスの利用状況等について調査
利用動向調査の結果を取りまとめました。

1.1 インターネット等の普及状況
インターネットの普及状況について、利用状況や端末
調査しました。

***1.1.1*** インターネットの利用状況 (個人)
インターネット利用者の割合を男女別にみると、男
、女性は 79.4%となり、とも に前年から…

1.1.2 端末別インターネットの利用状況 (個人)
平成 26 年の 1 年間のインターネットの利用状況
ると、「自宅のパソコン」 が 53.5%と最も多く

レベル1

レベル2

レベル3
*italic*

レベル3

出力しない

### JSONファイル

```
{
 “answer_units”: [
  {
  “type”:"body",
  "title":"no-title",
  “content”: {
   "text":"1. 通信利用動向調
  }
 },
 {
  "type":"h2",
  "title":"1.1. インターネット等の普及状況",
  “content”: {
   "text":"インターネットの普及状況について…
1.1.1 インターネットの利用状況(個人) インターネット利用者の割合を…”
  }
 }
 ]
}
```

レベル1

レベル2と３

### リクエストパラメータconfig(JSON)

```
{ “conversion_target”:"ANSWER_UNITS",
 “word” : {
  “heading” : {
   “fonts” : [
    { “level” : 3, “italic” : true }
   ]
  }
 },
 “answer_units”: {
  “selector_tags”: [ “h2” ]
 }
}
```

IBM Watson™

# Appendix

# 変換ルールの設定項目

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| conversion_target | 必須 | conversion_targetの値一覧のいずれか | N/A | 変換後のファイル形式を指定する。 |
| Word | オプション（指定する場合はWord か PDFのいずれか) | Wordファイル入力設定を参照 | なし | Wordファイルから入力する場合のルールを定義する。 |
| pdf | | PDFファイル入力設定を参照 | なし | PDFファイルから入力する場合のルールを定義する。 |
| normalize_html | オプション | HTML正規化設定を参照 | なし | 正規化HTMLに変換するルールを定義する。 |
| answer_units | オプション | Answerユニット出力設定を参照 | なし | Answerユニットデータ形式に出力する場合のルールを定義する。 |

# conversion_target の値一覧

| 値 | 説明 |
| --- | --- |
| ANSWER_UNITS | Answer Units データに変換する。 |
| NORMALIZED_HTML | 正規化されたHTMLに変換する。 |
| NORMALIZED_TEXT | シンプルテキストに変換する。 |

# Wordファイル入力設定

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| heading | 必須 | headingの値 | N/A | |

# Wordファイル入力設定 heading の値

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| fonts | オプション | フォント指定のリスト形式 | N/A | Wordファイル取り込み時の見出しレベルとフォント属性の関連ルールを定義します。 |
| styles | オプション | スタイル指定のリスト形式 | N/A | Wordファイル取り込み時の見出しレベルとスタイル属性の関連ルールを定義します。 |

# Wordファイル入力設定(フォント指定)

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| level | 必須 | 1, 2, 3, 4, 5, 6 | N/A | 指定した見出しレベルに該当する要素の必要条件を後続の項目で定義します。（AND条件） |
| min_size | オプション | 数値 | 0 | 対象となる最小フォントサイズを指定します。<br>０より大きい値を指定してください。 |
| max_size | オプション | 数値 | N/A | 対象となる最大フォントサイズを指定します。<br>min_sizeより大きい値を指定してください。 |
| bold | オプション | true, false | FALSE | trueを指定した場合はボールドフォントの要素が対象となります。<br>（ボールドフォントではない要素は対象外となります）<br>falseを指定した場合はボールドフォントの有無の判定はしません。<br>（全て対象の要素となります） |
| italic | オプション | true, false | false | trueを指定した場合はイタリックフォントの要素が対象となります。<br>（イタリックフォントではない要素は対象外となります）<br>falseを指定した場合はイタリックフォントの有無の判定はしません。<br>（全て対象の要素となります） |
| name | オプション | フォント名のリスト形式 | N/A | 対象となるフォント名を指定します。<br>設定しなかった場合は全てのフォントが対象となります。 |

# Wordファイル入力設定(スタイル指定)

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| Level | 必須 | 1,2,3,4,5,6 | N/A | 指定した見出しレベルに該当する要素の必要条件を後続の項目で定義します。（AND条件） |
| names | オプション | スタイル名のリスト形式 | N/A | 対象となるスタイル名を指定します。<br>設定しなかった場合は全てのスタイルが対象となります。 |

# PDFファイル入力設定

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| heading | 必須 | headingの値 | N/A | |

# PDFファイル入力設定 heading の値

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| fonts | オプション | フォント指定の<br>リスト形式 | N/A | PDFファイル取り込み時の見出しレベル<br>とフォントの関連ルールを定義します。 |

# PDFファイル入力設定(フォント指定)

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| level | 必須 | 1, 2, 3, 4, 5, 6 | N/A | 指定した見出しレベルに該当する要素の必要条件を後続の項目で定義します。（AND条件） |
| min_size | オプション | 数値 | 0 | 対象となる最小フォントサイズを指定します。<br>０より大きい値を指定してください。 |
| max_size | オプション | 数値 | N/A | 対象となる最大フォントサイズを指定します。<br>min_sizeより大きい値を指定してください。 |
| bold | オプション | true, false | FALSE | trueを指定した場合はボールドフォントの要素が対象となります。<br>（ボールドフォントではない要素は対象外となります）<br>falseを指定した場合はボールドフォントの有無の判定はしません。<br>（全て対象の要素となります） |
| italic | オプション | true, false | false | trueを指定した場合はイタリックフォントの要素が対象となります。<br>（イタリックフォントではない要素は対象外となります）<br>falseを指定した場合はイタリックフォントの有無の判定はしません。<br>（全て対象の要素となります） |
| name | オプション | フォント名のリスト形式 | N/A | 対象となるフォント名を指定します。<br>設定しなかった場合は全てのフォントが対象となります。 |

# HTML正規化設定

IBM Watson™

| キー | 条件 | 設定値 | 既定値 | 説明 |
|---|---|---|---|---|
| exclude_tags_completely | オプション | タグのリスト形式 | なし | 変換対象から除外する要素のタグを指定します。 |
| exclude_tags_keep_content | オプション | タグのリスト形式 | なし | 変換対象の要素からタグのみを取り除き、コンテンツは保持するタグを指定します。 |
| keep_content | オプション | xpath式 | | xpath式で指定された要素が保持されます。<br>exclude_contentを設定している場合にはそちらの条件が優先されます。 |
| exclude_content | オプション | xpath式 | | xpath式で指定された要素を除外します。 |
| keep_tag_attributes | オプション（指定する場合はkeep_tag_attributesかexclude_tag_attributesのいずれか） | 属性名, EVENT_ACTIONS, *のリスト形式 | [ “*” ] | 要素内に保持する属性を指定します。<br>EVENT_ACTIONSを指定した場合はJavaScriptの全てのイベント属性が対象となります。<br>*(アスタリスク)を指定した場合は全ての属性が対象となります。 |
| exclude_tag_attributes | オプション | 属性名, EVENT_ACTIONS, *のリスト形式 | なし | 要素から取り除く属性を指定します。<br>EVENT_ACTIONSを指定した場合はJavaScriptの全てのイベント属性が対象となります。<br>*(アスタリスク)を指定した場合は全ての属性が対象となります。 |

両方を指定した場合は
400エラーとなります

# Answer ユニット出力設定

| キー | 条件 | 値 | 既定値 | 説明 |
|---|---|---|---|---|
| selector_tags | 必須 | 見出しレベルのリスト形式 | [ “h1”, “h2”, “h3”, “h4”, “h5”, “h6” ] | 分割する見出しレベル(h1, h2, h3, h4, h5, h6)をリスト形式で指定します。<br>デフォルトでは全ての見出しレベルが分割されるようになっています。 |

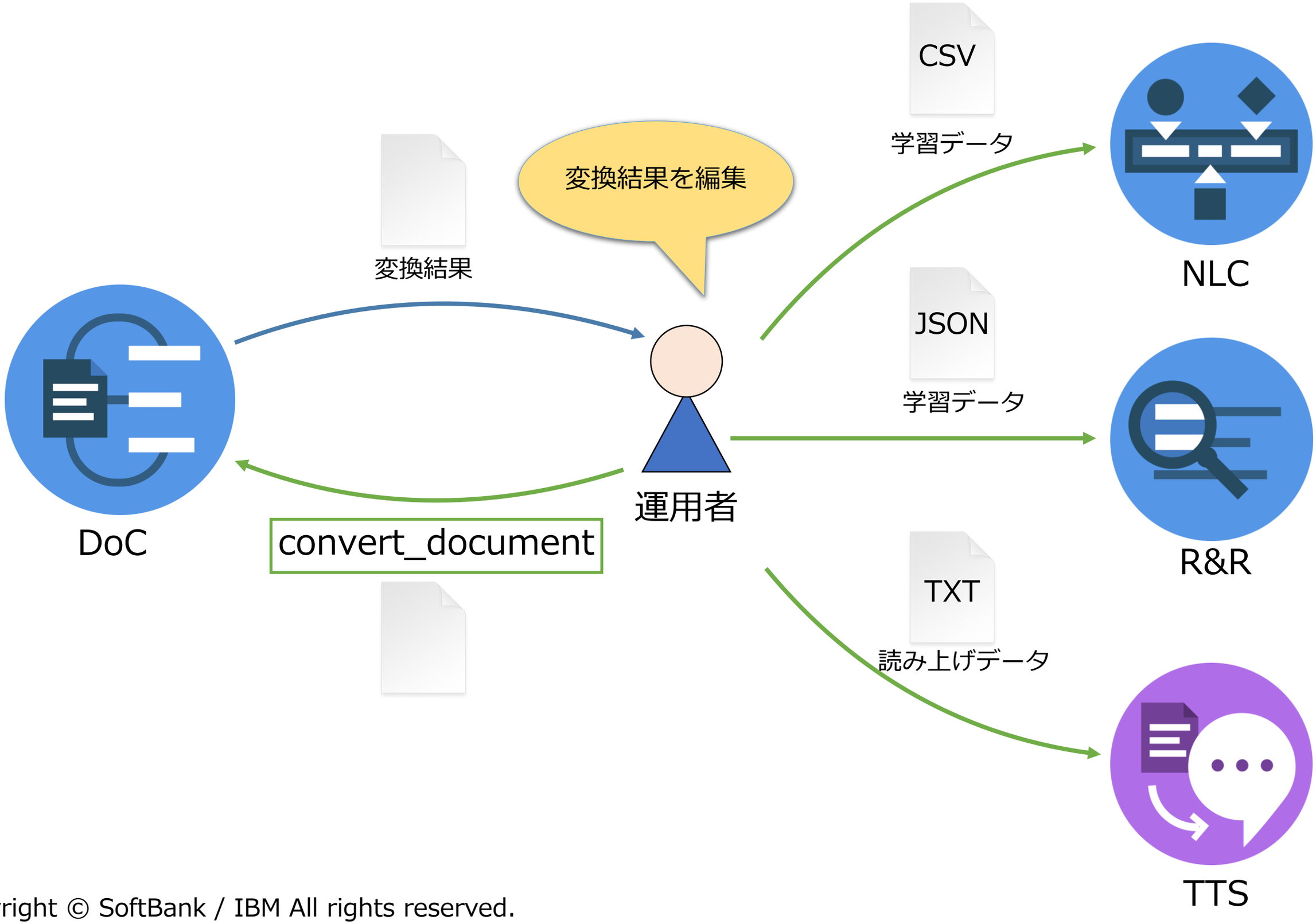
# エラーコード一覧

IBM Watson™

| code<br>(HTTPステータスコード) | error<br>(エラーメッセージ) | 実行結果 |
|---|---|---|
| 200 - OK | N/A | 正常終了 |
| 400 - Bad Request | Missing a required parameter or invalid parameter value. | 要求されたパラメータが無効です。 |
| 401 - Unauthorized | No API key provided or the API key provided was not valid. | APIの要求が認可されませんでした。 |
| 404 - Not Found | The requested item or parameter doesn't exist. | 要求されたリソースが見つかりませんでした。 |
| 413 - Payload Too Large | The uploaded file exceeds the maximum allowed file size | アップロードしたファイルサイズが上限を超えています。(上限:50Mバイト) |
| 500 - Internal Server Error | Internal server error. | 要求先のサーバに問題があるためにリクエストが処理できませんでした。 |

# DoCのAPI

IBM Watson™



IBM Watson™

# EOF