

Error identification service

Sushkov Ilya, Beregovoy Igor, Goncharov Danil, Mikolae Arion

May 2024

Abstract

The optimization of the system functioning is implemented by checking the correctness of filling in attributes by users associated with a specific Mandatory requirement.

Github: https://github.com/CrystalMath404/Error_identification_service_in_reference_fields.

1 Introduction

The need to search for errors in the text fields of documents arose a long time ago and with the help of ML it became possible to automate these routine tasks. The solution can be useful for business and government agencies because it will reduce the burden on employees to check documents

1.1 Team

Our team:

Sushkov Ilya ML research, preparation of the report

Beregovoy Igor ML research, data structuring

Goncharov Danil ML research

Mikolae Arion data structuring

2 Model Description

Introductory notes:

The Unified Register of Mandatory Requirements system stores Acts to which mandatory requirements are linked reflecting the essence of the act.

This system has the following business process:

1. The administrator analyzes the Act and selects a new mandatory requirements from its content, in addition to those that were already created earlier for this act.

2. The administrator creates a card for a new mandatory requirements, where the card is a set of attributes and characteristics corresponding to this mandatory requirements.

3. The administrator fills in the fields of the card in accordance with the information contained in the mandatory requirements.

4. The administrator saves the card, as a result of which a new mandatory requirements is added to the act, and the number of such requirements becomes $N+1$.

3 Dataset

We are working with data obtained from the National University of Science and Technology "MISIS"

The model should return the probability of an error for each reference field being checked.

We have 9 datasets with the same column names. Let's connect them. Now the set contains 5 columns. 2 of this columns are json types. After parsing this data we got set contains many columns. To begin with, we will delete all columns with a large number of missing values. We have columns named *"title"*, *"npaType"*, *"npaTitle"*, *"fullTitle"*, *"checkQuestion"*, *"npaLinkPravoGovRu"*, *"controlTypes"*, *"assessmentForm"*, *"categoriesOfPersons"*, *"mrEstablishmentObject"*, *"publicRelationSpheres"*, *"regulationLevel"* Column named *"npaLinkPravoGovRu"* got 0.28 missing values, so we filled it with *'omcymcmeyem na www.pravo.gov.ru'*.

Our target variables are *'mrEstablishmentObject'*, *'assessmentForm'*, *'controlTypes'*, *'categoriesOfPersons'*, *'publicRelationSpheres'*

We train the CatBoostClassifier model, transpose the result and calculate Accuracy and f-score

4 Metrics



$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ F1 &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned}$$

Figure 1: Key metrics

5 Results

Accuracy is 0.9329

Now count the F-score for each Model

Model name	Weighted-average F1 score	Micro-average F1 score	Macro-average F1 score
mrEstablishmentObject	0.9632	0.9633	0.9235
assessmentForm	0.9850	0.9855	0.7191
controlTypes	0.9100	0.9180	0.1428
categoriesOfPersons	0.9919	0.9921	0.8821
publicRelationSpheres	0.8037	0.8054	0.1264

Table 1: F-Score Table