Метод нечеткой классификации платежей по полнотекстовому описанию на русском языке

Студент: Лаврова Анастасия Андреевна

Группа: ИУ7-85Б

Научный руководитель: Волкова Лилия Леонидовна

Актуальность

Сфера применения – организация определения кода операции платежных документов в банках

Задача поставлена одним из крупных российских банков

Проблемы:

- Задействовано большое количество персонала
- Значительные временны́е затраты

Работа посвящена автоматизации классификации платежных документов

Цель и задачи

Цель — разработать метод классификации платежа по полнотекстовому описанию.

Задачи:

- Рассмотреть существующие типы алгоритмов машинного обучения
- Проанализировать существующие алгоритмы классификации
- Разработать метод нечеткой классификации платежей по полнотекстовому описанию на русском языке
- Программно реализовать метод
- Исследовать разработанный метод на применимость

Примеры данных: платежи и их описания

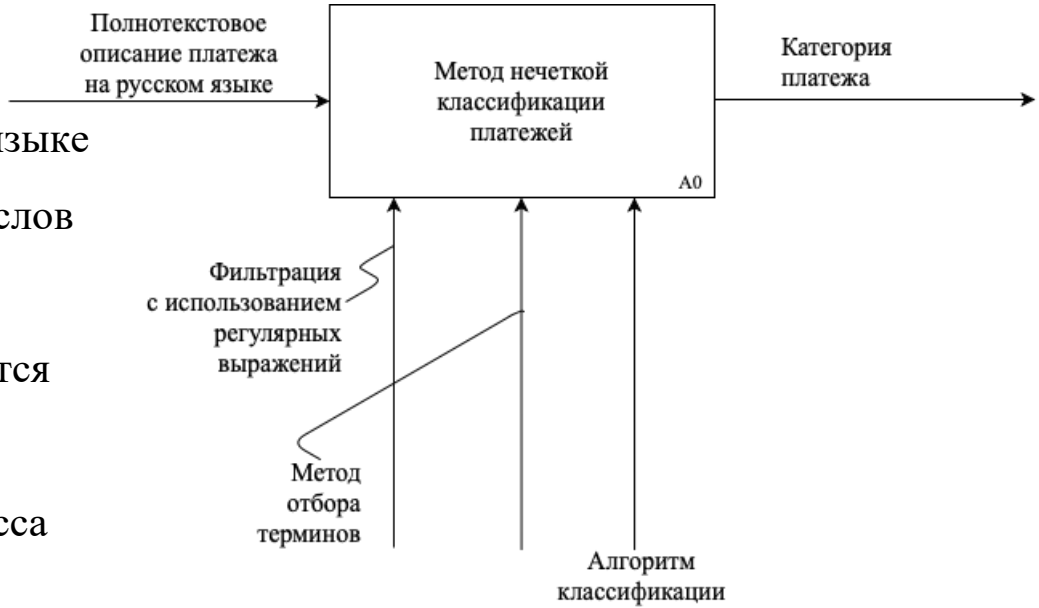
Код операции	Наименование	Назначение платежа
1800	Перевод средств на р/с	Оплата страхового взноса по Договору страхования № 0122130464484 от 05.09.2018, ФИО страхователя Ермошкина Екатерина Ивановна, сумма цифрами 550,00.
14	Перевод средств пенсионных накоплений в 153	Доход от инвестирования (срочная пенсионная выплата) по договору № 22-03У008 от 08.10.2003 г. НДС не облагается.

Постановка задачи

Ограничения, накладываемые на метод:

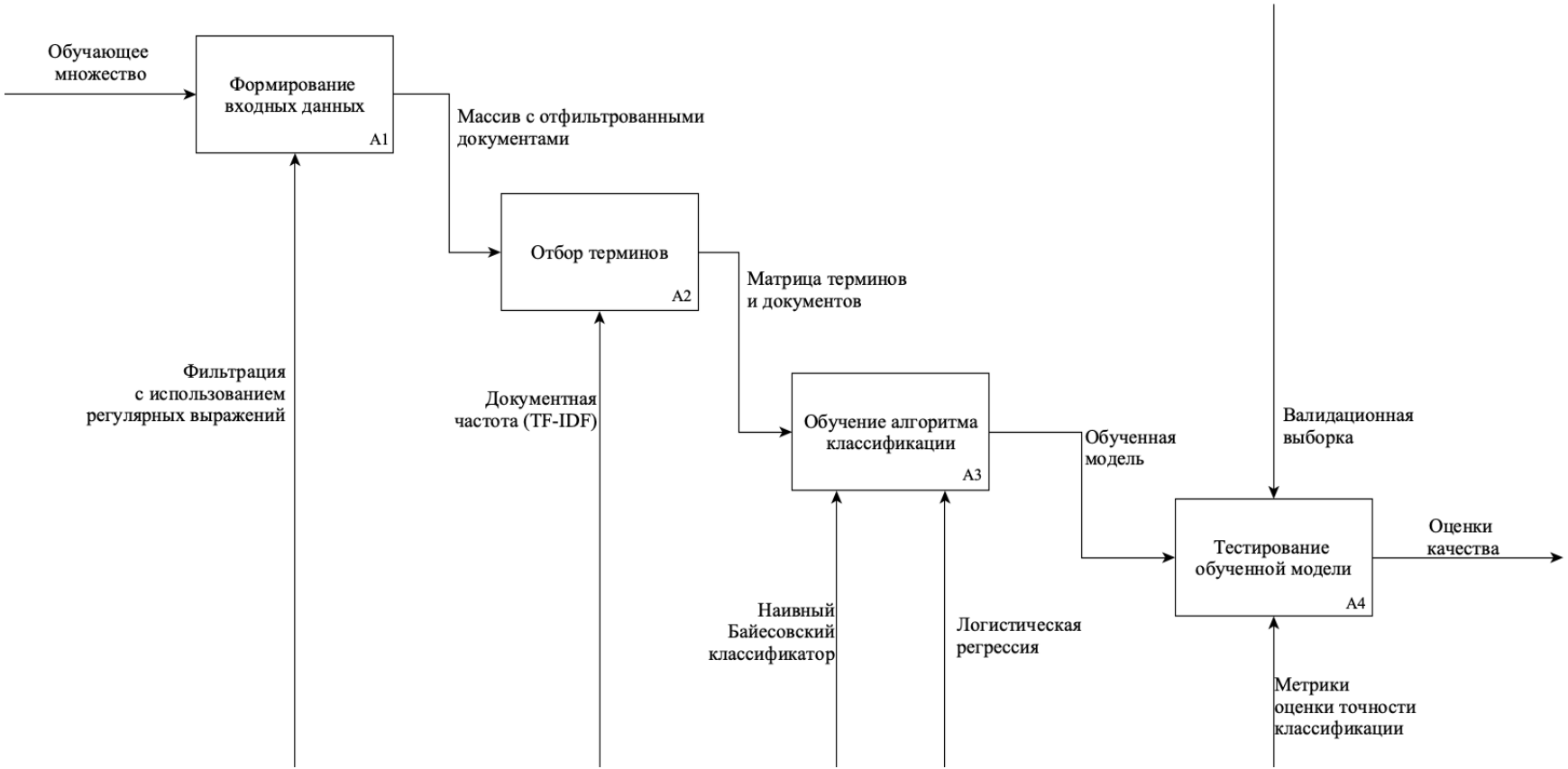
- На вход подается строка на русском языке
- Длина строки составляет от 10 до 30 слов
- Количество классов = 144+1

Вводится дополнительный класс «требуется ручной ввод» для случаев недостаточной уверенности классификатора в метке класса



Будет разработан метод нечеткой классификации для оценки уверенности в присвоении класса платежу

Функциональная модель обучения метода классификации



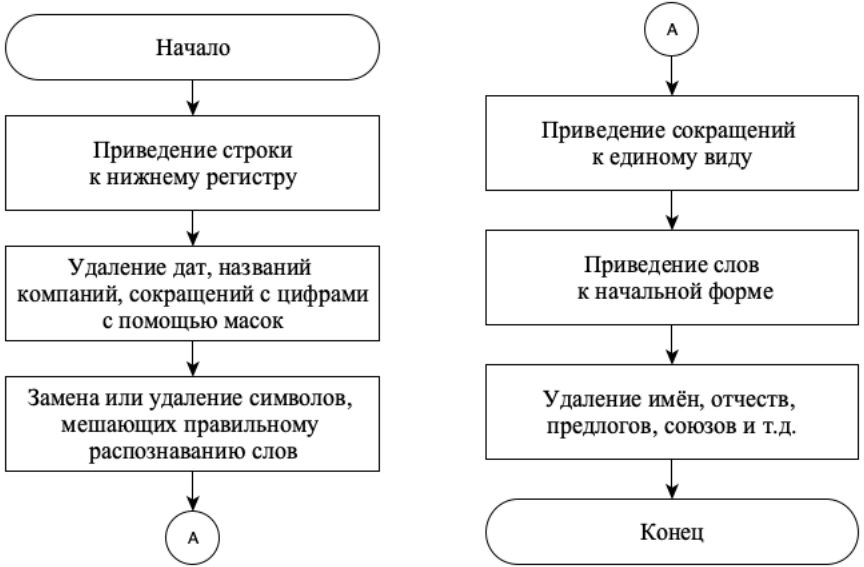
Предобработка данных

Исходное текстовое описание:

“Оплата страхового взноса по Договору страхования № 0122130464484 от 05.09.2018, ФИО страхователя Ермошкина Екатерина Ивановна, сумма цифрами 550,00.”

После токенизации и фильтрации шумовых слов и символов:

“оплата страховой взнос договор страхование фио страхователь сумма цифра”



Отбор терминов

Термин — это слово в начальной форме. Его вес рассчитывается как TF-IDF.

t – термин
D – коллекция документов
d – документ из коллекции D

$$TF - IDF(t, D) = TF(t, d) \times IDF(t_i, D)$$

$$TF(t, d) = \frac{n_t}{\sum_k n_k}$$

n_t – количество вхождений термина t в документ
$\sum_k n_k$ – общее количество слов в документе

$$IDF(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right)$$

$|D|$ – количество документов
$|D_i \in D|$ – число документов, где t_i встретилось хотя бы один раз

Выбор метода классификации текста

	Наивный Байесовский классификатор	Логистическая регрессия
Вычислительная сложность обучения	O(|Ω|)	O((f+1)ΩEc)
Вычислительная сложность тестирования	O(|c|)	O((f+1)c)

Ω – множество документов
c – количество классов
f – число терминов
E – количество эпох градиентного спуска

Наивный Байесовский классификатор

Наивный Байесовский классификатор - простой вероятностный классификатор, основанный на формуле Байеса.

$$c^* = \arg_{c_j \in C} \max P(c_i|d_j)$$

$$P(c_i|d_j) = \frac{P(c_i)P(d_i|c_j)}{P(d_i)} \approx P(c_j)P(d_i|c_j)$$

$P(c_j)$ – априорная вероятность, что документ принадлежит классу c_j

$P(d_i|c_j)$ – вероятность встретить документ типа d_i среди документов, класса c_j.

Логистическая регрессия

Логистическая регрессия является методом построения линейного классификатора, которая позволяет оценить апостериорные вероятности принадлежности объектов классам.

$$h_\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

h_θ – гипотеза
β_0 и β_1 – коэффициенты линейного уравнения, определяющие положение прямой в пространстве

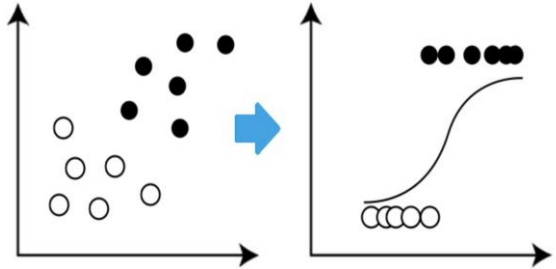
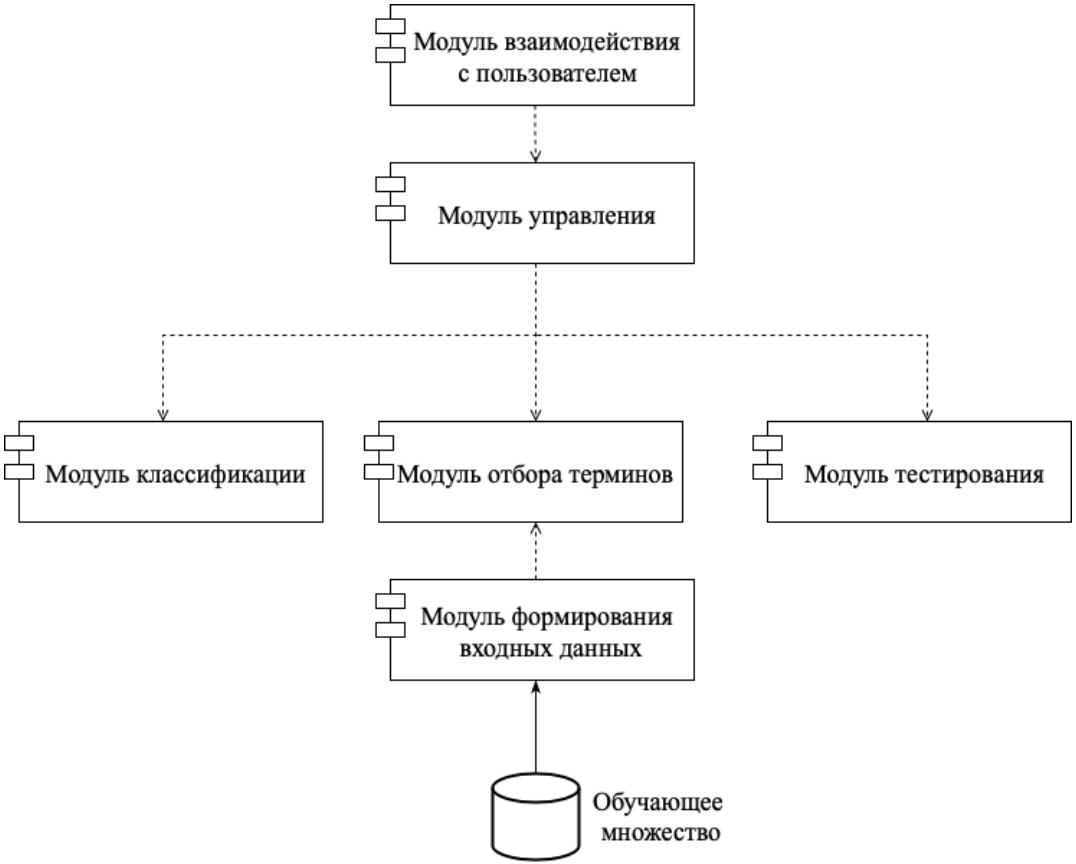


Схема ПО


Модуль взаимодействия с пользователем
Модуль управления
Модуль классификации
Модуль отбора терминов
Модуль тестирования
Модуль формирования входных данных
Обучающее множество

Оценка точности классификации

F-мера — это гармоническое среднее между точностью и полнотой

$$f1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$precision = \frac{TP}{TP + FP}$$

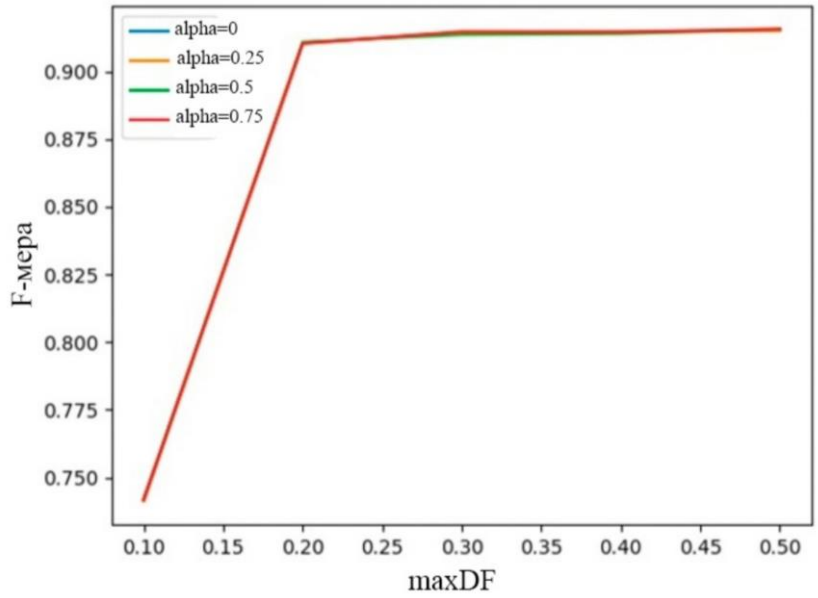
$$recall = \frac{TP}{TP + FN}$$

		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	True Positive (TP)	False Positive (FP)
	Отрицательная	False Negative (FN)	True Negative (TN)

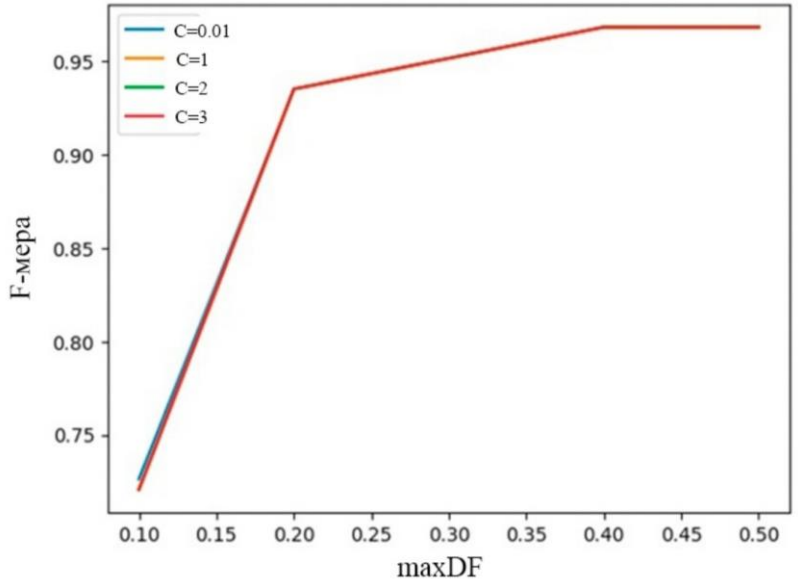
Параметризация методов классификации

- Для анализируемых данных явно выделена чувствительность методов к maxDF
- Влияние внутренних параметров слишком мало (графики для разных внутренних параметров накладываются)

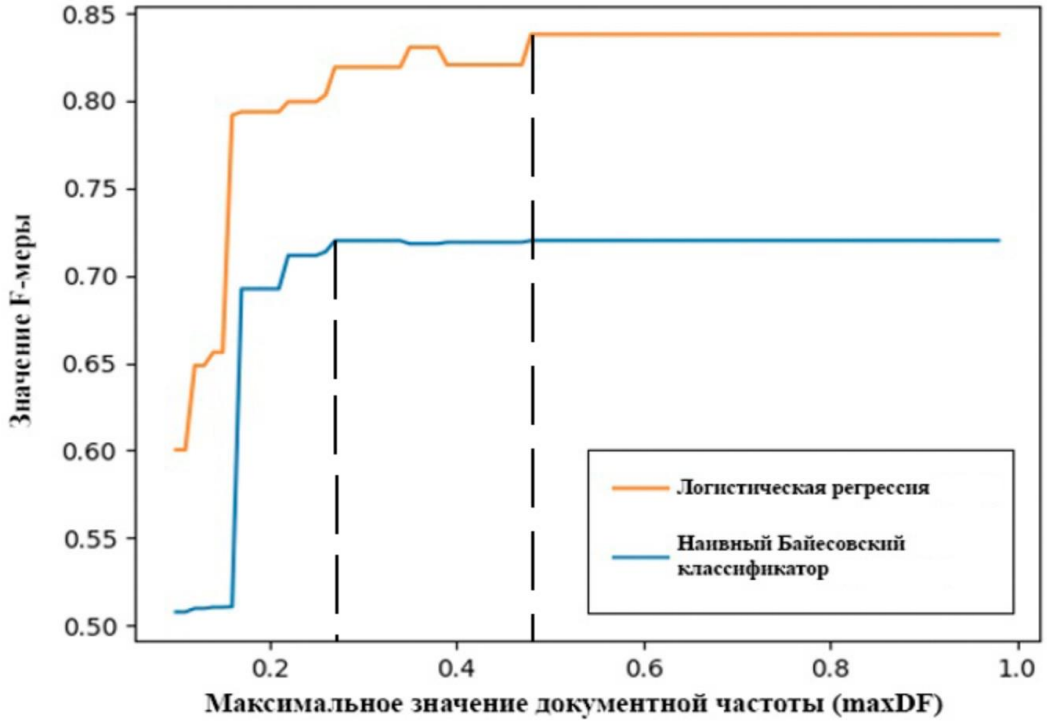
Наивный Байесовский классификатор



Логистическая регрессия



Зависимость точности от порогового значения документной частоты



Рекомендуется пороговое значение maxDF = 0.3 для наивного Байесовского классификатора и maxDF = 0.5 для логистической регрессии

Точность метода логистической регрессии выше

Заключение

Достигнута цель: разработан метод классификации платежей по полнотекстовому описанию.

Решены поставленные задачи:

- Рассмотрены существующие типы алгоритмов машинного обучения
- Проанализированы существующие алгоритмы классификации
- Разработан метод нечеткой классификации платежей по полнотекстовому описанию на русском языке
- Сконструировано и разработано программное обеспечение, демонстрирующее работу метода
- Проведено исследование точности классификации

Предложенный метод рекомендуется к применению

Дальнейшее развитие

- Применение других стандартных методов классификации с модификацией
- Учёт не только отдельных терминов, но и словосочетаний
- Написание правил для использования дополнительных замен при помощи регулярных выражений для покрытия примеров вида <Дата + ФИО + Номер счета>