ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Классификация госконтрактов по объектам закупки

Черненко Наталья Алексеевна

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Содержание

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Постановка задачи

Уточнение задачи

Исходные требования

На входе имеются данные карточек госконтрактов с ftp.zakupki.gov.ru. Необходимо на основе данных с ftp.zakupki.gov.ru научиться определять группу, к которой относится контракт с кодом ОКПД-2 41, 42, 43, 71.1.

Группы могут быть следующими:

1. Строительно-монтажные работы (СМР)
2. Проектно-изыскательские работы (ПИР)
3. Строительный надзор
4. Подключение коммуникаций
5. Прочее.

По ОКПД-2 контракты в общем случае должны разделяться так:

1. Строительно-монтажные работы (СМР) - 41, 42, 43(кроме нижеперечисленных)
2. Проектно-изыскательские работы (ПИР) - 41.1, 71.1
3. Подключение коммуникаций - 43.22
4. Строительный надзор – четкой группы нет.

Новая задача

В самом файле данные размечены только по кодам ОКПД2, которые могут быть неверными. Разметка данных в таком случае будет отдельной задачей.

Попробуем создать рекомендательную систему на основе объекта закупки с отнесением объекта к коду ОКПД2.

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Описание используемых методов

Метод кодирования признаков и определения меры сходства

Content-Based подход основан на измерении похожести между объектами на основе их содержания.

Потребуется перевести тексты в числовые данные и определить сходство анализируемого объекта с имеющимися в базе.

Используются:

1. Метод кодирования TF–IDF (Term Frequency–Inverse Document Frequency — частота слова–обратная частота документа)
2. Косинусная мера сходства для выдачи рекомендаций

Разведочный анализ

Анализ данных из файла

Перед работой с текстовыми данными необходимо обработать саму таблицу.

Основные шаги:

1. Чтение файла
2. Работа с пропусками
3. Работа с дублями

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ

Чтение файла

1. С помощью скриптов, написанных на языке python, был подобран диалект для корректной загрузки файла в датафрейм
2. Произведена перезапись всего файла с новым разделителем
3. Отобраны объекты закупки с к кодами ОКПД2, указанными в ТЗ
4. Установлено название практически всех столбцов
5. Созданы переменные для использования в методе read_csv

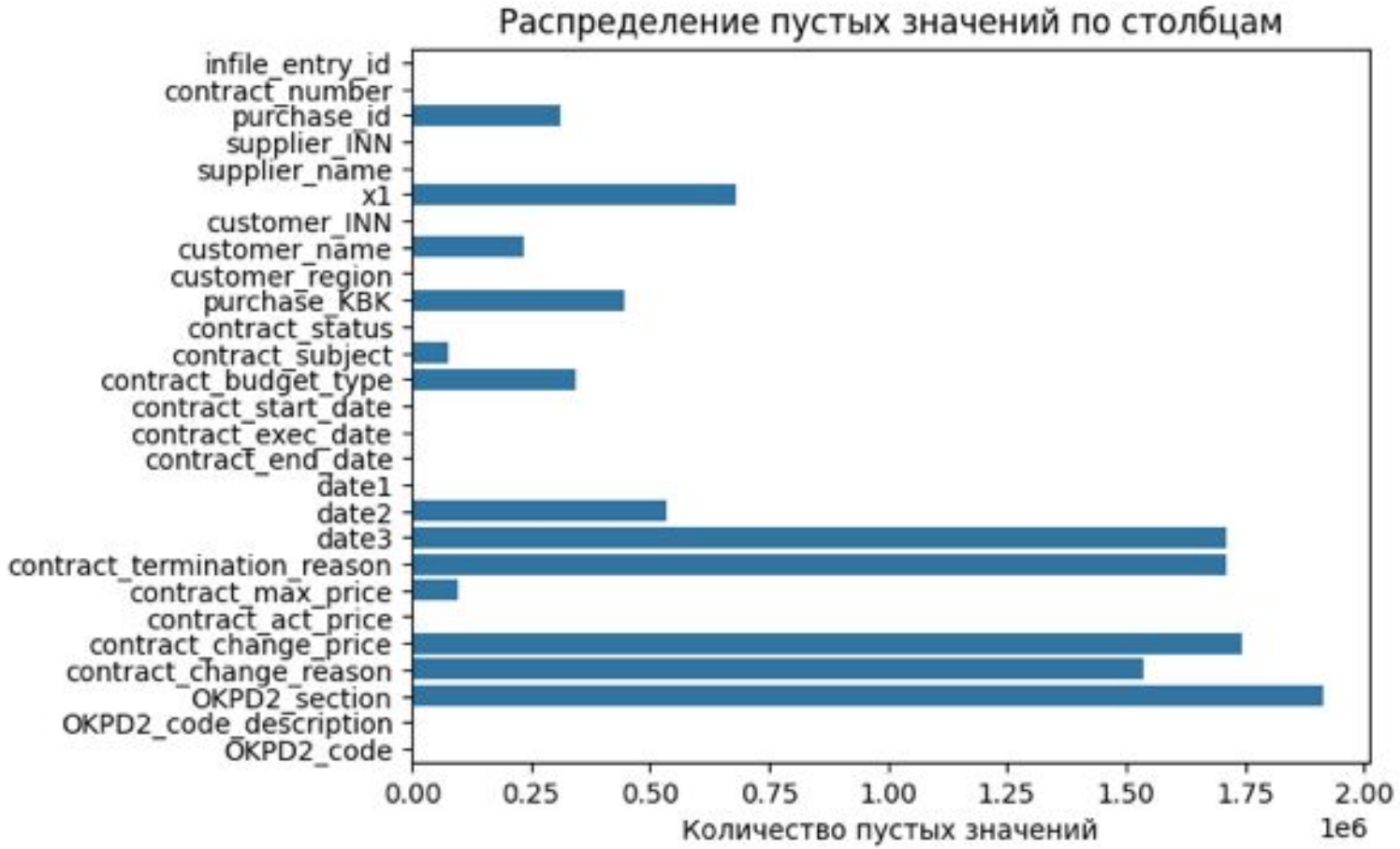
Разведочный анализ

Работа с пропусками

После изучения данных и составления заголовка появилась возможность определить значимые признаки.

Большая часть пропусков находится в признаках, которые не были отобраны для анализа

Строки с пропущенные значения удалены



Разведочный анализ

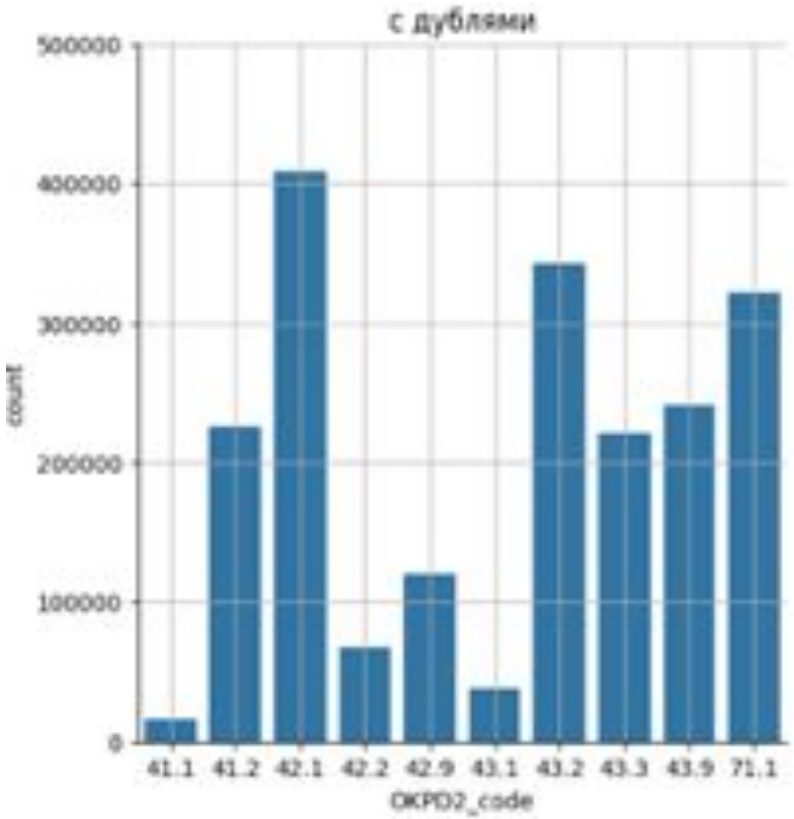
Работа с дублями

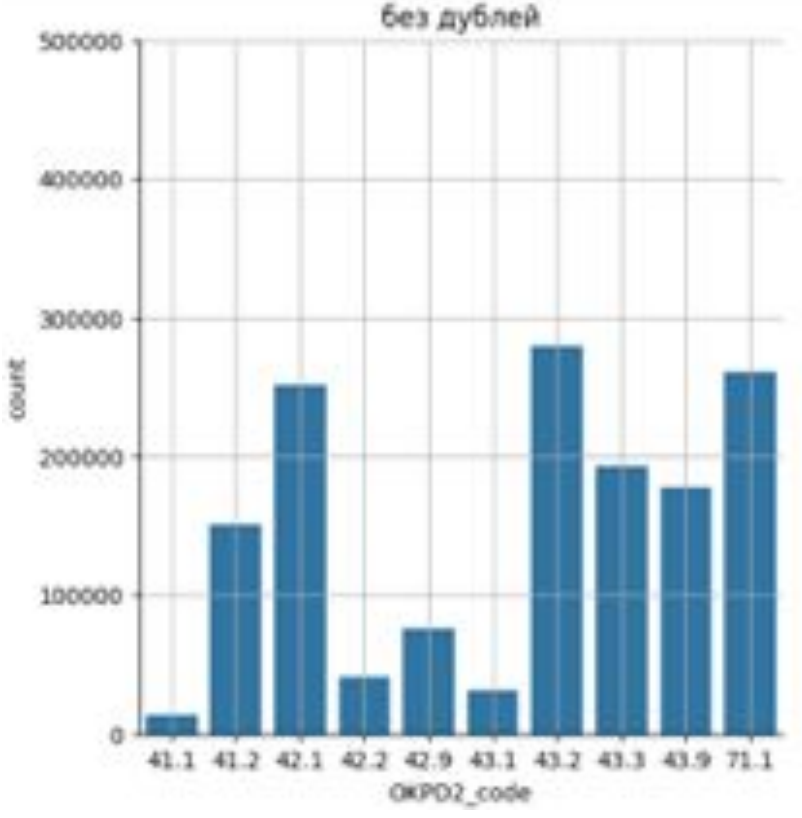
Основным критерием поиска дублей выступал реестровый номер закупки. Итого было определено три основных типа дублей:

1. Совпадающий номер закупки, но разные категории. Записывается в отдельный файл
2. Совпадающий номер закупки и категории, но разные номера КБК. Оставляется одна запись.
3. Совпадающий номер закупки, но есть изменения в признаках, не отобранных для анализа. Оставляется одна запись

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ





Сравнение данных до и после работы с пропусками и дублями

Предобработка текстовых данных

Основные шаги:

1. Очистка текста от пунктуации и специальных символов, удаление лишних слов
2. Стемминг
3. Векторизация

Разведочный анализ

Предобработка текстовых данных

```python
MIN_CHARS = 4
MAX_CHARS = 10
def tokenizer(sent, min_chars=MIN_CHARS, max_chars=MAX_CHARS, lemmatize=True):
    if lemmatize:
        stemmer = nltk.stem.SnowballStemmer("russian")
        tokens = [stemmer.stem(w) for w in word_tokenize(sent)]
    else:
        tokens = [w for w in word_tokenize(sent)]
    token = [w for w in tokens if (len(w) > min_chars and len(w) < max_chars)]
    return token
```

```python
tok_test = tokenizer(clean_test)
tok_test
```

```
['выполнен', 'работ', 'подготовк', 'систем', 'отоплен', 'сезон']
```

Итоги работы функции очистки и токенизации

Построение системы

Основная функция

```
def get_recommendations_tfidf(sent, tfidf_mat):
    clean_sent = clean_text(sent)
    tokens_query = tokenizer(clean_sent)
    embed_query = vectorizer.transform(tokens_query)
    mat = cosine_similarity(embed_query, tfidf_mat)
    best_index = extract_best_indices(mat, topk=3)
    return best_index

vectorizer = TfidfVectorizer(tokenizer=tokenizer)
tfidf_mat = vectorizer.fit_transform(df['cleaned_subject'].values)
```

Функция выдачи рекомендации

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Построение системы

Тестирование

```
test_sentence = 'Выполнение работ по подготовке системы отопления к зимнему отопительному сезону 2022-2023 г.г.'
best_index = get_recommendations_tfidf(test_sentence, tfidf_mat)

display(df[['OKPD2_code', 'contract_subject']].loc[best_index])
```

	OKPD2_code	contract_subject
978022	71.1	Выполнение картосоставительских работ в предел...
612515	41.2	Выборочный капитальный ремонт здания фельдшерс...
841138	42.9	Текущий ремонт здания общежития № 2 Литер К ГБ...

Результат работы рекомендательной системы

ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru