МИНОБРНАУКИ РОССИИ
федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»
Институт дополнительного образования
Высшая инженерная школа

СОЗДАНИЕ ЧАТ-БОТА, ОПРЕДЕЛЯЮЩЕГО СРОК ХРАНЕНИЯ ДОКУМЕНТА

по программе профессиональной переподготовки:
«Анализ данных на языке Python»

Выполнил(а):
Власова Татьяна Григорьевна

Руководитель:
Кандидат экономических наук,
доцент по научной специальности
«Математические и
инструментальные
методы экономики»,
Заграновская Анна Васильевна

Санкт-Петербург
2022

Обзор ситуации

- «Перечень типовых управленческих архивных документов, образующихся в процессе деятельности государственных органов, органов местного самоуправления и организаций, с указанием сроков хранения» (утв. Приказом Росархива от 20.12.2019 № 236),
- ч.4 ст. 7 № 115-ФЗ «О противодействии легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма» от 07.08.2001,
- ст.22 № 125-ФЗ «Об архивном деле в Российской Федерации» от 22.10.2014(ред. от 11.06.21)
- и т.д.

Формулировка проблемы

Неверное определение сроков хранения документов может привести к нежелательным последствиям как непосредственно для сотрудника, так и для организации в целом:

- отсутствие первичных учетных документов и документов кадрового учета часто влечет за собой налоговые последствия и административную ответственность,
- утрата документов, содержащих государственную тайну, согласно статье 284 УК РФ, наказывается ограничением свободы на срок до трех лет.

Потенциальные пути решения проблемы

- Номенклатура
- Запрет
- Обучение
- Архив с инструментарием

Цель

Целью работы является разработка вспомогательного инструмента, подсказывающего рекомендуемый срок хранения документа для сотрудников, не обладающими специальными знаниями и навыками в сфере документооборота.

Задачи

- Собрать данные о фонде организации.
- Провести разведочный анализ собранных данных, чтобы понять, документы с какими сроками хранения образуются в ходе деятельности организации.
- Выбрать лучшую модель для решения задачи классификации документов по срокам хранения, с долей правильных ответов не менее 80%.
- Создать инструмент, которым сможет воспользоваться любой сотрудник организации.

Требования к разрабатываемому инструменту

Тематика	**Созданный справочник должен давать ответ на вопрос, сколько рекомендуется хранить тот или иной документ**
Целевая аудитория	Сотрудники организации, не обладающими специальными знаниями в области архивного законодательства
Основные пользовательские функции	Решение задачи по определению сроков хранения, без необходимости просмотра пользователем всех возможных нормативных документов.
Дизайн	Удобный, интуитивно-понятный пользователю интерфейс справочника

Универсальные признаки

- Место формирования документа.
- Наличие не разрешенных споров/разногласий по вопросам, затронутым в документе.
- Наличие в организации вредных/опасных условий труда или условий приравненных к ним.
- Периодичность отчетности.

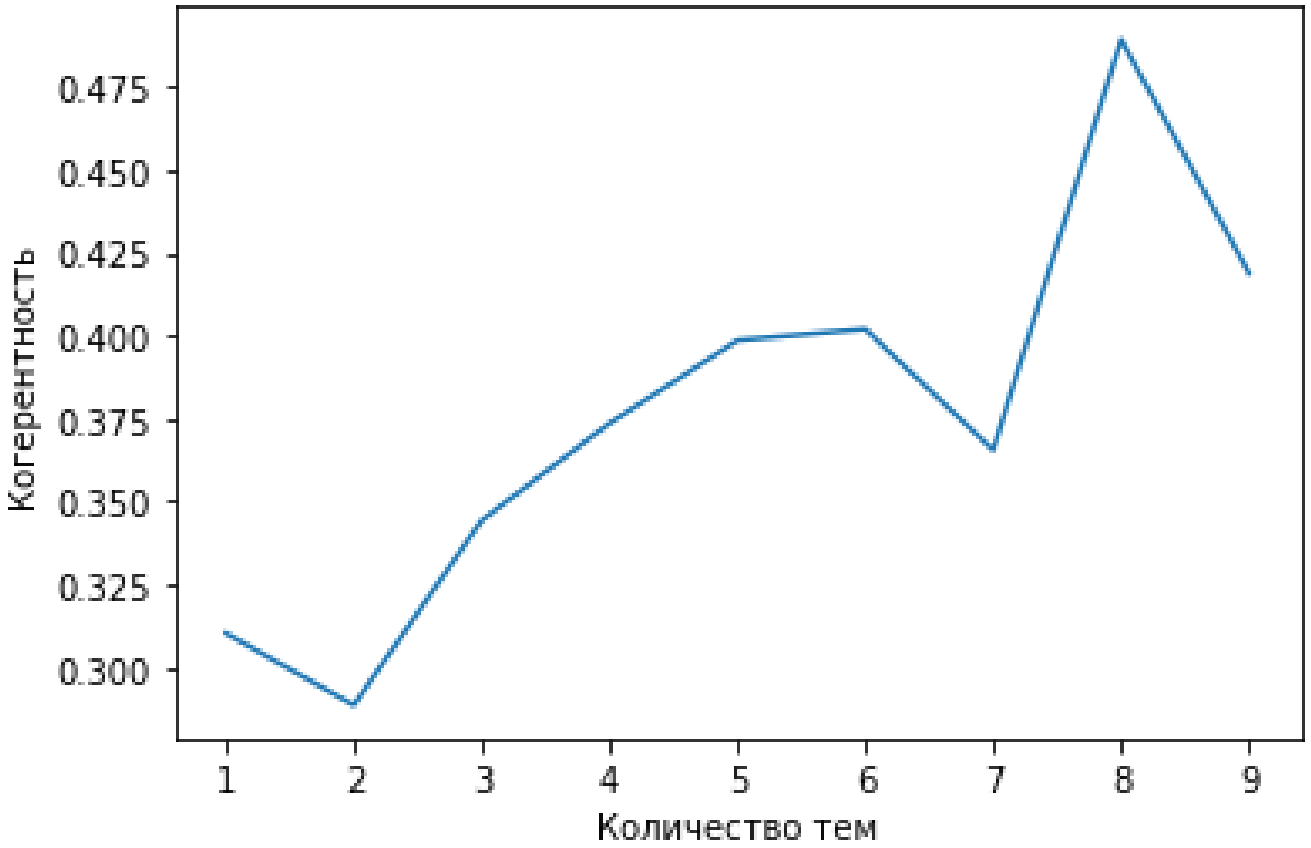
Признаки, уникальные для фонда

- Отдел, в ходе деятельности которого был создан документ.
- Тематика документа.

Тематическая модель для автоматического выявления тематики документа

1. Создание корпуса.
    1.1. Загрузка документов с сайта.
    1.2. Преобразование их из формата .pdf в формат .txt.
    1.3. Создание датасета с текстами.
2. Предобработка текста.
    2.1 Удаление стоп слов и пунктуации.
    2.2. Выделение устойчивых оборотов.
    2.3 Лемматизация или приведение слов к нормальной форме.
3. Подбор оптимального количества тем.
4. Создание модели.

Когерентность тем



Темы и их названия

1. **Организационно-правовые документы**
('0.024*"вуз" + 0.019*"образование" + 0.019*"университет" + ‘ '0.012*"россиискои_федерации" + 0.011*"деятельность" + 0.009*"год" + ' '0.008*"программа" + 0.007*"совет" + 0.007*"филиал" + 0.007*"подготовка"')

2. **Документы о приеме**
('0.041*"документ" + 0.019*"обучение" + 0.016*"лицо" + 0.013*"квалификация" + ' '0.012*"приложение" + 0.011*"образование" + 0.011*"программа" + ' '0.010*"приём" + 0.007*"поступать" + 0.007*"бланк")

3. **Документы по экзаменам**
('0.021*"обучаться" + 0.014*"гэк" + 0.013*"экзамен" + 0.011*"технология" + ' '0.010*"членов_гэк" + 0.010*"образовательный" + 0.009*"университет" + ' '0.009*"применением_электронного" + 0.009*"вкр" + 0.008*"работа")

4. **Документы по научной деятельности**
('0.032*"программа" + 0.012*"организация" + 0.010*"деятельность" + ' '0.010*"образование" + 0.009*"обучение" + 0.009*"университет" + ' '0.009*"развитие" + 0.008*"работа" + 0.008*"реализация" + 0.007*"научно"')

5. **Документы по учебным программам**
('0.025*"ооп" + 0.020*"программа" + 0.011*"обучаться" + 0.010*"обучение" + ' '0.010*"университет" + 0.010*"организация" + 0.009*"работа" + 0.009*"вид" + ' '0.009*"подготовка" + 0.008*"компетенция"')

6. **Развитие университета**
('0.001*"программа" + 0.001*"университет" + 0.000*"деятельность" + ' '0.000*"образование" + 0.000*"развитие" + 0.000*"вуз" + 0.000*"обучение" + ' '0.000*"образовательный" + 0.000*"реализация" + 0.000*"обучаться"),

7. **Деятельность филиалов**
('0.028*"образование" + 0.020*"образовательнои" + 0.016*"деятельность" + ' '0.012*"программа" + 0.010*"россиискои_федерации" + 0.010*"филиал" + ' '0.008*"лицензиат" + 0.007*"образовательный" + 0.007*"направление" + ' '0.007*"копия"')

8. **О коррупциии**
('0.033*"срок" + 0.019*"год" + 0.019*"размер" + 0.018*"лицо" + ' '0.016*"деятельность" + 0.016*"заниматься_определеннои" + 0.013*"должность" ' '+ 0.011*"взятка" + 0.010*"такового_лишением" + 0.010*"период") - **о коррупции**

Описание выборки

Документарный фонд лизинговой компании с использованием информации из книг учета поступлений документов в архив:

- 185 объектов
- 6 номинальных признаков + 1 результативный признак – срок хранения

Разведочный анализ

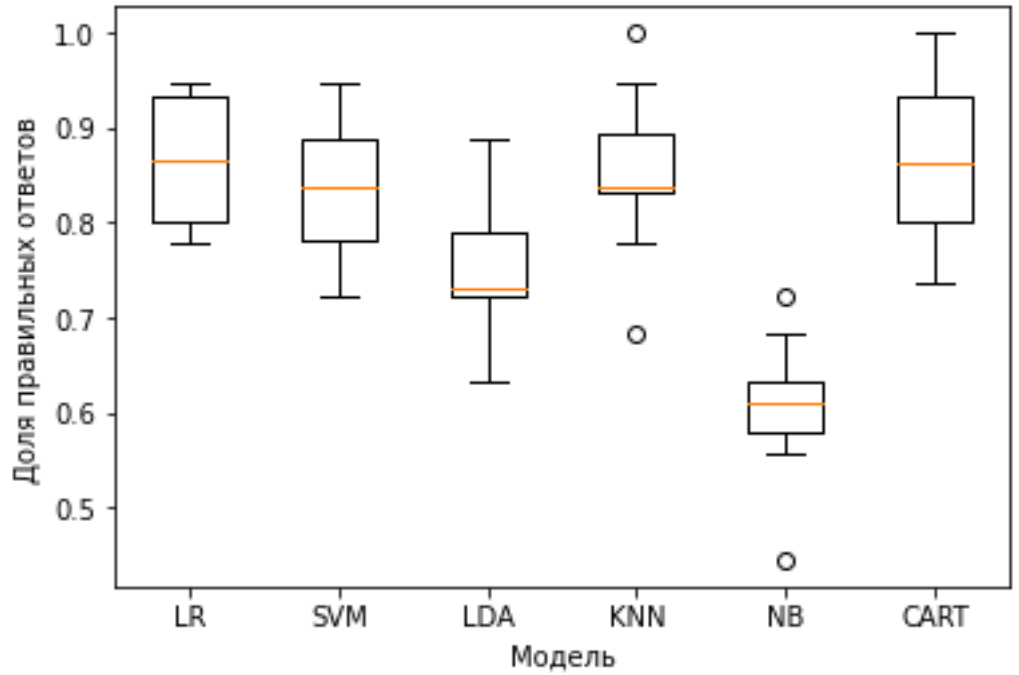
	count	unique	top	freq
Название	185	106	Приказы о приеме, переводе, увольнении	6
происхождение	185	2	в ходе деятельности организации	151
сфера деятельности	185	9	ЛЗ	46
Наличие споров/разногласий	185	2	нет	137
отдел	185	5	бухгалтерия	66
Относится к годовой отчетности	185	2	нет	173
Относится к кадровым документам при вредных условиях труда	185	2	нет	168
срок хранения	185	4	5 лет	106

Многоклассовая классификация

```
df['срок хранения'].value_counts()
```

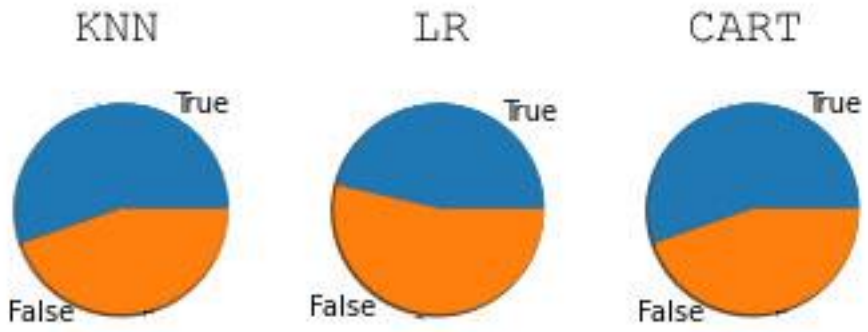
```
5 лет                   106
до разрешения спора      38
Постоянно                21
75 лет                   20
Name: срок хранения, dtype: int64
```

Простые методы: результаты работы на кросс-валидации (kfold = 10)



Методы многоклассовой классификации со сбалансированной обучающей выборкой

- Метод k-го ближайшего соседа - дает на тестовой выборке 55 % правильных ответов.
- Логистическая регрессия - 46 % правильных ответов.
- Деревья решений для классификации - дает 55 % долю правильных ответов.



Ансамблевые методы: результаты работы на кросс-валидации (kfold = 10)

Ансамбль	Доля правильных ответов
Ансамбль классификаторов (k-го ближайшего соседа, логистическая регрессия и деревья решений)	0.83
Классификатор дополнительных деревьев	0.81
XGBoost	0.82
Градиентный бустинг	0.84

Градиентный бустинг

«Бустинговые алгоритмы являются итеративными алгоритмами, которые размещают различные веса в тренировочном распределении на каждой итерации». «An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics» [Текст]/ Victoria López, Alberto Fernández, Salvador García, Vasile Palade, Francisco Herrera

Градиентный бустинг 100 деревьев решений -
дает на тестовой выборке **94% правильных ответов**

Тестирование модели градиентного бустинга

№ п/п	Заголовок дела (групповой заголовок документов)	Годы	Коли-чество ед. хр.	Сроки хранения и номера статей по перечню	Примечание
1	Счета-фактуры выданные	2011	28	5 лет Ст. 368	
2	Авансовые отчеты (авансовые отчеты, товарные чеки, служебные записки)	2009-2010	13	5 лет Ст. 362, 665	
3	Касса (оборотно-сальдовые ведомости, кассовые книги)	2009-2010	2	5 лет Ст. 361, 362	
4	Акты сверки (акты сверки взаиморасчетов)	2005-2010	8	5 лет Ст. 366	
5	Налоговая (требования об уплате налогов, переписка с налоговой, акты проверок, требования об истребовании документов, информации)	2008-2010	7	5 лет Ст. 382, 398, 402	
6	НДС (налоговые декларации по налогу на добавленную стоимость, книги покупок, книги продаж)	2009-2010	8	5 лет Ст. 392, 459 у)	
7	Транспортный налог (налоговые декларации по транспортному налогу)	2010	1	5 лет Ст. 392	
8	Налоговые регистры (бухгалтерские справки, оборотно-сальдовые ведомости, налоговые регистры)	2007-2010	1	5 лет Ст. 361	
9	ОСВ (оборотно-сальдовые ведомости)	2007-2009	2	5 лет Ст. 362	
10	Приказы о краткосрочных командировках, ежегодных оплачиваемых отпусках (без деления на филиалы)	2005-2009	12	5 лет Ст. 19 в)	

	Происхождение	Тема	Наличие споров/разногласий	Отдел	Относится к годовой отчетности	Вредность	Срок хранения
0	0	6	1	3	1	1	0
1	0	0	1	0	1	1	0
2	0	6	1	3	1	1	0
3	0	6	1	3	1	1	0
4	0	3	1	3	1	1	0
5	0	3	1	3	1	1	0
6	0	3	1	3	1	1	0
7	0	3	1	3	1	1	0
8	0	6	1	3	1	1	0
9	0	1	1	4	1	1	0

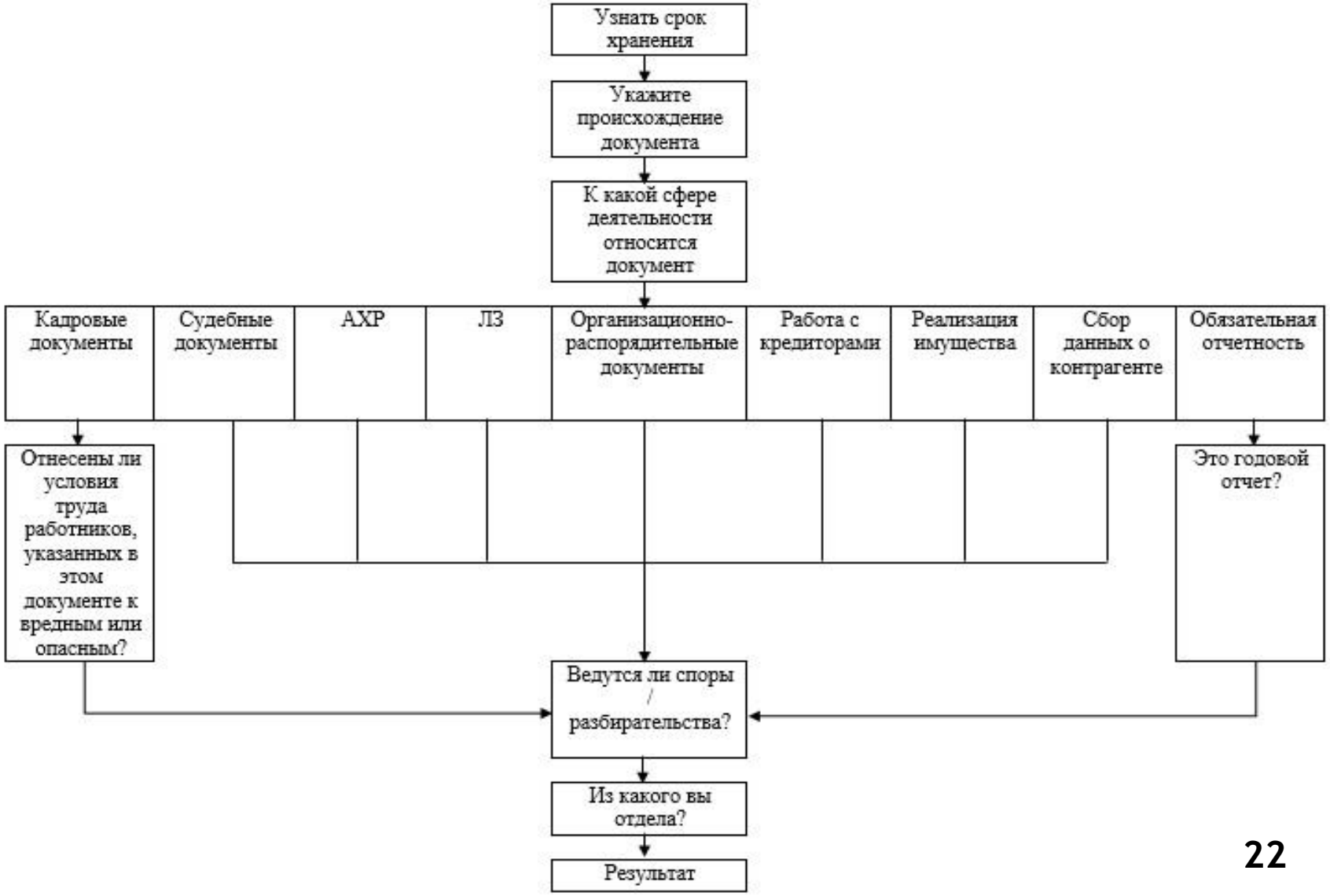
```
[ ]  accuracy_score(Y, predictions)
```

```
0.9
```

Чат-бот: вопросы пользователю

- Укажите происхождение документа
- К какой сфере деятельности относится документ?
- Ведутся ли споры/разбирательства?
- Отнесены ли условия труда работников, указанных в этом документе, к вредным или опасным?
- Это годовой отчет?
- Из какого вы отдела?

Чат-бот: сценарии диалогов



Реализация чат-бота

Итоги

- Собраны данные о документарном фонде организации.
- Построена тематическая модель для определения тематики документа.
- Проведен разведочный анализ собранных данных и выделено 4 класса документа по срокам хранения.
- Построены следующие модели классификации: логистическая регрессия, k ближайших соседей, решающие деревья, Ансамбль классификаторов (k-го ближайшего соседа, логистическая регрессия и деревья решений), классификатор дополнительных деревьев, XGBoost, градиентный бустинг.
- Выбрана лучшая модель классификации на несбалансированных данных - градиентный бустинг.
- Разработан и реализован чат-бот на языке Python.