Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина» (УрФУ)
Институт радиоэлектроники и информационных технологий - РТФ

ОТЧЕТ
о проектной работе
по теме: Разработка вопросно-ответной системы для нормативно-правовых актов России
по дисциплине: Проектный практикум

Команда: Клюква

Екатеринбург
2022

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современном мире самым простым способом найти актуальную информацию о чём-либо является сеть Интернет. Однако не всегда поисковик может корректно ответить на вопрос пользователя. Бывает, что ответ скрыт под тоннами текста и найти его поиском по странице не всегда возможно. В частности, иногда люди обращаются в сеть Интернет в надежде получить ответы на юридические вопросы. И зачастую найти такие ответы крайне проблематично.

Наш проект направлен на решение данной проблемы. Вместо поиска вопроса в поисковике, используя нейронную сеть, пользователь обращается напрямую к интересующим его документам, таких как Конституция РФ, Трудовой, Уголовный Кодексы и другие.

На данном этапе выполнения учебного проекта первостепенной задачей для нас является выполнение всех требования MVP к указанному в календарном плане сроку.

Наши основные задачи по проекту: составить реестр требований, провести аналитику, изучить аналоги, определиться со стеком технологий, разработать минимально работающий веб-сервис.

КОМАНДА

— Лебедев Егор Михайлович РИ-200003 – Тимлид

— Зенков Илья Дмитриевич РИ-200014 – Аналитик

— Сарапулов Матвей Дмитриевич РИ-200003 – Разработчик

— Куланчеев Евгений Анатольевич РИ-200016 – Разработчик

ЦЕЛЕВАЯ АУДИТОРИЯ

Люди возрастом 14+ любого пола, ищущие в сети Интернет актуальную информацию о том или ином законодательном акте.

КАЛЕНДАРНЫЙ ПЛАН ПРОЕКТА

Название проекта: Вопросно-ответная система для нормативно-правовых актов России

№	Название	Ответственный	Длительность	Дата начала	Временные рамки проекта(недели)							
					1	2	3	4	5	6	7	8
Анализ												
1.1	*Поиск и создание датасетов*	Зенков И.Д.	2 недели	01.04.2022								
1.2	*Изучение библиотеки Huggingface*	Куланчеев Е.А.	2 недели	01.04.2022								
1.3	*Изучение SQuAD, SberQuAD*	Зенков И.Д.	2 недели	01.04.2022								
1.4	*Анализ BERT моделей*	Лебедев Е. М.	2 недели	01.04.2022								
1.5	*Изучение методов нечёткого поиска фразы по тексту*	Куланчеев Е.А.	2 недели	01.04.2022								
1.6	*Формулировка цели*	Лебедев Е. М.	2 недели	01.04.2022								
1.7	*Формулирование требований к продукту*	Сарапулов М.Д.	2 недели	01.04.2022								
1.8	*Определение задач*	Вся команда	2 недели	01.04.2022								
Проектирование												
2.1	*Выбор технологий реализации проекта*	Сарапулов М.Д.	1 неделя	08.04.2022								
2.2	*Изучение темы doc2vec, sent2vec*	Сарапулов М.Д.	1 неделя	15.04.2022								
2.3	*Прототипирование user-story*	Куланчеев Е.А.	1 неделя	08.04.2022								
2.4	*Разработка веб-интерфейса*	Лебедев Е. М.	3 недели	08.04.2022								
2.5	*Разработка общей архитектуры проекта*	Зенков И.Д.	3 недели	08.04.2022								

Разработка												
3.1	*Разработка серверной части сервиса*	Зенков И.Д.	3 недели	06.05. 2022								
3.2	*Разработка микросервиса*	Зенков И.Д.	3 недели	15.04. 2022								
3.3	*Объединение основной логики сервиса с пользовательским интерфейсом*	Сарапулов М.Д.	6 недель	15.04. 2022								
3.4	*Перенос кодовой базы с jupyter ноутбука в github*	Лебедев Е. М.	6 недель	15.04. 2022								
3.5	*Разработать дизайн пользовательского интерфейса*	Куланчеев Е.А.	6 недель	15.04. 2022								
Тестирование												
4.1	*Запуск кода на локальном ПК, SberQuAD*	Зенков И.Д.	2 недели	13.05. 2022								
4.2	*Тестирование пользовательского интерфейса*	Зенков И.Д.	2 недели	13.05. 2022								
4.3	*Тестирование разных методов семантического поиска*	Сарапулов М.Д.	3 недели	06.05. 2022								
4.4	*Общее тестирование сервиса*	Лебедев Е. М.	3 недели	06.05. 2022								
4.5	*Тестирование вопрос-ответной модели*	Куланчеев Е.А.	3 недели	06.05. 2022								
Презентация												
5.1	*Написание отчёта*	Зенков И.Д.	1 неделя	20.05. 2022								
5.2	*Репетиция защиты проекта*	Куланчеев Е.А.	1 неделя	20.05. 2022								
5.3	Создание стиля презентации	Лебедев Е. М.	1 неделя	20.05. 2022								

5.4	*Подготовка текста презентации и доклада*	Сарапулов М.Д.	1 неделя	20.05. 2022								
5.5	*Создание презентации*	Куланчеев Е.А.	1 неделя	13.05. 2022								

ОПРЕДЕЛЕНИЕ ПРОБЛЕМЫ

Проблема, которую может решить наш проект, заключается в том, что для людей является трудозатратным поиск ответов на юридические вопросы.

ПОДХОДЫ К РЕШЕНИЮ ПРОБЛЕМЫ

1. Поиск по тексту из нормативно-правовых актов. Пользователь ищет ответ по ключевым словам, а не по смыслу. Основной минус заключается в том, что ответ может быть сформулирован другими словами.
2. Обращение к специалисту. Чаще всегда эти услуги бывают не бесплатными, и могут занять достаточно большое количество времени.
3. Поиск в сети Интернет. Зачастую поисковик не выдаёт конкретный ответ, а лишь даёт набор статей, в которых может быть ответ.

АНАЛИЗ АНАЛОГОВ

1) Haystack. Использует современные технологии семантического поиска, state-of-the-art языковые модели. Имеет открытый исходный код и удобный REST API. Пользователю, чтобы начать использовать сервис, нужно развернуть его на сервере, сконфигурировать и скачать нужные модели. Основной минус: плохая поддержка русского языка «из коробки».

2) Milvud. Может искать по картинкам, химическим структурам и текстам. Также имеет открытый исходный код. Хорошо подходит для семантического поиска, но не для задачи Q&A. Не поддерживает русский язык.

ТРЕБОВАНИЯ К ПРОДУКТУ И К MVP

- Наличие пользовательского интерфейса
- Возможность выбрать документ из заранее представленных
- Модель должна выдавать ответ на вопрос исходя из текста

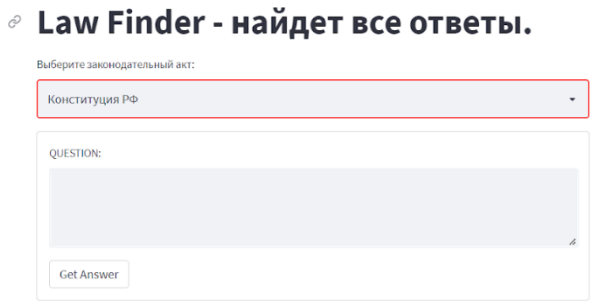
СТЕК ДЛЯ РАЗРАБОТКИ

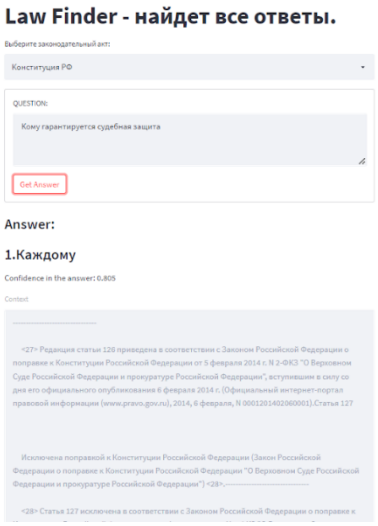
1) Python

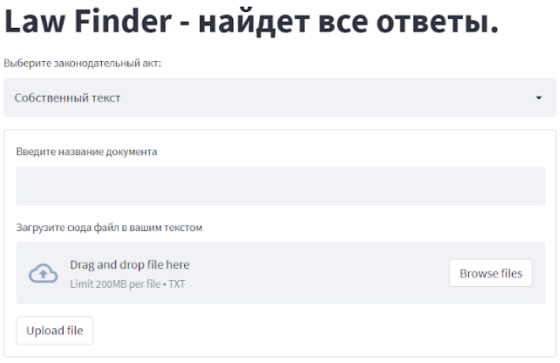
2) HuggingFace

3) Streamlit

4) fastText


Law Finder - найдет все ответы.
Выберите законодательный акт:
Конституция РФ
QUESTION:
Get Answer


Law Finder - найдет все ответы.
Выберите законодательный акт:
Конституция РФ
QUESTION:
Кому гарантируется судебная защита
Get Answer
Answer:
1.Каждому
Confidence in the answer: 0.805
Context


Law Finder - найдет все ответы.
Выберите законодательный акт:
Собственный текст
Введите название документа
Загрузите сюда файл в вашим текстом
Drag and drop file here
Limit 200MB per file • TXT
Browse files
Upload file

РАЗРАБОТКА СИСТЕМЫ

В самом начале запускается streamlit, затем подгружается модель XLM-RoBERTa[1]. Пользователь может выбрать нормативный акт или загрузить свой текст. Потом пользователь вводит свой вопрос и нажимает кнопку «Получить ответ». Затем Q&A модель даёт свой выводит свой ответ и контекст, в котором лежит ответ.

При загрузке собственного текста мы стемизируем слова с помощью MyStem3 от Яндекса и токенизируем на предложения с помощью библиотеки razdel, затем fasttext[2] даёт эмбеддинги (векторизованное представление) для всего массива данных. Далее при поиске он ищет ближайшее по смыслу предложение с помощью косинусного расстояния — так реализован семантический поиск. Мы берём топ 5 предложений и 2 предложения рядом с ним для контекста.

ЗАКЛЮЧЕНИЕ

Нами проведен анализ конкурентов, сделаны выводы о конкурентоспособности нашего продукта, выявлена целевая аудитория, подтверждена актуальность продукта, выбран технический стек реализации нашего продукта. Также нами реализован MVP. Изучили методы получения эмбеддингов для текста, научились использовать модели из библиотеки HuggingFace, научились создавать пользовательский интерфейс с помощью Streamlit.

СПИСОК ЛИТЕРАТУРЫ

1. SberQuAD -- Russian Reading Comprehension Dataset: Description and Analysis /// arXiv:1912.09723

2. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information // arXiv:1607.04606