Talent Case

СБЕР

Talent Case Contest 2023

В борьбе за уникальность: выявление рерайтинга

Используя предоставленный датасет, реализуйте алгоритм определения рерайтинга текста

Привет!

Talent Case | СБЕР

Поздравляем вас с началом отборочного этапа Talent Case Contest 2023!

В течение этой недели вам предстоит решить кейс по выявлению рерайтинга и отправить решение на почту.

Ключевые этапы кейс-чемпионата:

30.10 – 8.11	Отборочный этап
1.11 в 17:00	Брифинг и QA-сессия с экспертами
8.11 в 23:59	Дедлайн отправки решения
13.11	Объявление финалистов
13.11 – 22.11	Финальный этап
23.11	Очный финал в офисе Сбера в Москве

Напоминаем, что отборочный этап пройдет полностью онлайн с заочной оценкой решений – постарайтесь оформить презентацию и исходный код максимально читаемо и аккуратно!

Брифинг и QA-сессия с экспертами
1 ноября, 17:00, онлайн

Специально для вас составители кейса расскажут о задаче кейса, поделятся советами и ожиданиями, а также ответят на все ваши вопросы!

Задать вопрос эксперту

Ссылку на брифинг и всю важную информацию опубликуем в канале

Введение

Talent Case СБЕР

Рерайтинг – обработка исходных текстовых материалов в целях их дальнейшего использования. В отличие от копирайтинга, за основу берётся уже написанный текст, который переписывается своими словами с сохранением смысла.

Обработка естественного языка (NLP)

– это технология машинного обучения, которая дает компьютерам возможность интерпретировать, манипулировать и понимать человеческий язык.

Как работает NLP?

Обработка естественного языка сочетает в себе компьютерную лингвистику, машинное обучение и модели глубокого обучения для обработки человеческого языка. Компьютерная лингвистика – это наука о понимании и построении моделей человеческого языка с помощью компьютеров и программных инструментов. Исследователи используют методы компьютерной лингвистики, такие как синтаксический и семантический анализ, для создания платформ, помогающих машинам понимать разговорный человеческий язык. Такие инструменты, как переводчики языков, синтезаторы текста в речь и программное обеспечение для распознавания речи, основаны на компьютерной лингвистике.

Почему NLP играет такую важную роль?

Обработка естественного языка имеет решающее значение для эффективного анализа текстовых и речевых данных. Таким образом, можно преодолевать различия в диалектах, сленге и грамматических нарушениях, типичных для повседневных разговоров. Компании используют этот метод для нескольких автоматизированных задач, таких как:

- Обработка, анализ и архивирование больших документов
- Анализ отзывов клиентов или записей колл-центра
- Запуск чат-ботов для автоматизированного обслуживания клиентов
- Ответы на вопросы «кто, что, когда и где»
- Классификация и извлечение текста

Полезные ссылки:

ML: Введение в машинное обучение ↘

ML: N-граммы ↘

ML: Embedding слов ↘

ML: Embedding слов v2 ↘

Постановка задачи

Talent Case

СБЕР

Цель

Используя предоставленный датасет, реализовать алгоритм выявления рерайтинга:

1. Проанализировать основные приёмы рерайтинга;
2. Проанализировать различные методы обработки текстов;
3. Подобрать техническое решение для обработки файла;
4. Реализовать поиск дубликатов одним из найденных методов
5. Придумать не менее 2-х методов, которые позволят выявить переписанные тексты;
6. Получить список рерайтов на основе реализации алгоритмов.

Решение должно содержать:

- Презентацию с подробным отчетом о подходе, средствах и результатах
- Ссылку на репозиторий с исходным кодом решения и инструкциями README

Обязательное ограничение

Необходимо использовать исключительно **алгоритмический подход.** Применение нейронных сетей и готовых моделей машинного обучения *запрещено*.

→ Решение необходимо отправить до 23:59 8 ноября на почту:

cases@talentcase.ru

→ В текст письма добавьте ссылку на репозиторий и прикрепите презентацию до 10 слайдов в формате .pdf или .pptx

→ Назовите письмо и файл с решением по шаблону:

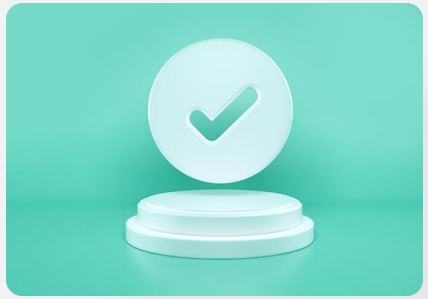
TCC2023_*название команды*

Это сильно упростит нам обработку решений и повысит ваши шансы на выход в финал :)

Критерии оценки решений

Talent Case

СБЕР



Письмо с вашим решением должно содержать:

1. Презентацию в формате *.pdf или .pptx* с подробным описанием подхода к решению и полученных результатах:
   - Объем не более 10 слайдов;
   - Титульный слайд и резюме команды (включает фото, имена и ваш опыт)
2. Ссылку на репозиторий *GitHub* с воспроизводимым алгоритмом, результатами обработки датасета и файлом с инструкциями README.

Критерии оценки презентации с отчетом

Критерии качества анализа:

- Качество проведенного анализа входных данных
- Описание способа обработки данных
- Логика вывода решений

Критерии качества презентации:

- Соблюдены формальные требования
- Все слайды имеют четкую структуру
- Нет технических ошибок и опечаток

Критерии оценки реализации алгоритма

Критерии качества реализации:

- Инструкция содержит исчерпывающие сведения по использованию решения
- Программная реализация покрывает описанные алгоритмы
- Решение воспроизводимо и чисто оформлено

Техническая точность:

- Процент корректно выявленных дубликатов в датасете

Talent Case Contest 2023

Команда организатора кейс-чемпионата ООО «Талант Кейс» подготовила данный кейс по заказу ПАО «Сбербанк».

Все решения, полученные в процессе проведения Кейс-чемпионата Talent Case Contest 2023, являются собственностью и могут быть использованы в коммерческих целях. Частные данные, используемые для подготовки кейса, могли быть изменены или сгенерированы с целью сохранения конфиденциальности данных.

Информационный канал кейс-чемпионата