Направление Data Science, профессия ML-инженер NLP-направления MTS AI

Аннотация

Добро пожаловать на платформу карьерного тест-драйва **Shift + Enter**!
Здесь ты можешь решить задания от MTS AI — компании, где работа построена на сплоченных продуктовых командах, использующих agile[1] методологию. MTS AI работает над созданием и развитием продуктов на стыке речевых технологий, компьютерного зрения и edge computing[2].

Попробуй себя в роли ML-инженера направления NLP MTS AI, чтобы:

- Лучше узнать, над какими реальными проектами работает компания.
- Прокачать знания основных библиотек Python для работы с данными и построить предсказательную модель.

Развиваемые компетенции

По результатам выполнения заданий ты сможешь развить такие навыки, как:

- Обучение модели машинного обучения.
- Анализ данных.

А также научишься критически мыслить, работать с открытыми данными и делать выводы.

Описание подзадач

Миссия команды MTS AI – соединить потребности бизнеса с эффективными AI-технологиями. И сейчас у тебя появилась возможность виртуально присоединиться к работе департамента и помочь коллегам с классификацией интентов[3] для определения смысла, который пользователи вкладывают в свои запросы.

Рекомендуемый тайминг

Выполнение всего блока заданий займет у тебя не более 80-95 минут.

1. 50-60 минут на первое задание*.
2. 30-35 минут на второе задание.

Информация о загрузке решения

Стажировка содержит несколько подзадач. Можно загрузить файл, содержащий решение части заданий, но по возможности постарайся сделать их все.

Желаем успехов!

---

[1] Agile — это группа методик для гибкого управления проектами в команде разработки. Рабочий процесс при таком подходе разбивается на небольшие временные промежутки, их еще называют спринтами или итерациями. Во время каждого спринта команда разработки создает часть продукта, которую можно протестировать и оценить.

[2] Edge computing (периферийные или граничные вычисления) – концепция, суть которой заключается в том, чтобы максимально приблизить место обработки данных к их источнику. Для этого часть процессов обработки данных переносится из центра обработки данных на служебные устройства.

[3] Под интентом в информационном поиске подразумевается потребность пользователя. По каждому запросу могут присутствовать основной интент и дополнительные (дополнительные потребности).

* В рекомендуемый тайминг включено время на ознакомление с заданием и на написание кода без учета затрат времени на запуск и обучение модели.

Задание 1. Обучи классификатор, используя открытые данные

Рады, что ты присоединился к команде MTS AI!

Александр[4], руководитель NLP[5]-девелоперов и твой наставник, попросил тебя подключиться к первой задаче и помочь коллегам с классификацией интентов.

Привет!

В экосистеме МТС есть множество продуктов, и у пользователей иногда возникают вопросы по их использованию. Они могут обращаться в службу поддержки (чаты) или пытаться самостоятельно найти ответы, вводя поисковые запросы. При этом алгоритм машинного обучения оперативно подхватывает пользовательские запросы и пытается предложить решение. Если алгоритм верно уловил суть, он дает полезный ответ. То есть с помощью описанного выше алгоритма, мы решаем задачу под названием «классификация интентов».

Твоя задача — провести классификацию интентов на основе примеров из предложенного датасета.

Для этого:

1. **Обучи классификатор, используя набор фраз из обучающей выборки.**
2. **Помни, что классификатор должен уметь самостоятельно оценивать, насколько фразы из тестовой выборки похожи на примеры из обучающей, и принять решение, к какому классу отнести в итоге эти запросы.**

Hints. Для решения задания рекомендую использовать трансформеры, например BERT, чтобы достигнуть высокого качества работы модели на интентах, которые есть в обучении.

Спасибо!

Полезные материалы

- Ссылка на датасет MASSIVE[6] (на английском языке), содержащий большое количество высказываний виртуального помощника Alexa. Обрати внимание, что датасет содержит 11500 записей в обучающей, 2030 записей — в валидационной и 2970 записей — в тестовой выборках.
- Курс об основах обработки естественного языка на Hugging Face.

Формат конечного результата

Код с комментариями - вы можете самостоятельно выбрать предпочтительный формат: файл .ipynb (ноутбук) с описанием или скрипт .py с readme файлом. В любом случае нужны какие-то комментарии и выводы по полученным результатам.

---

[4] Все имена и названия вымышленные, любые совпадения случайны. Данные заданий могут быть изменены в целях конфиденциальности.

[5] NLP (Natural Language Processing или обработка естественного языка) – технология машинного обучения, которая дает компьютерам возможность интерпретировать, манипулировать и понимать человеческий язык.

[6] Цитирование – MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.

Задание 2. Доработай модель с учетом новых вводных

Пришло время поработать над вторым и заключительным заданием. Александр прислал тебе детали проекта, связанного с доработкой модели, созданной на предыдущем этапе.

Привет,

Спасибо за отлично проделанную работу по обучению модели!

Как ты знаешь, пользователи не всегда могут формулировать запросы, на которые у модели есть ответы. Чтобы не вызывать их недовольство, работая заведомо некорректно, мы должны отлавливать запросы, относящиеся к тем интентам, которые модель не знает, либо вообще не содержащие в себе интент.

Твоя задача — доработать модель для возможности учета out-of-scope запросов.

Для этого:

1. Используй любые данные (чувствую, что без обогащения нам не обойтись 🙂) и методы на твое усмотрение. Советую для начала ознакомиться с теми, которые указаны в полезных материалах во вложении.
2. Предложи варианты, как можно оповещать пользователей о том, что их запрос не относится ни к одному из тех классов, которые модель умеет определять.

Спасибо!

Полезные материалы

- Статья про Out of Scope (OOS) detection.
- Материал об оценке меток для Out-of-Domain (OOD) интентов.
- Ссылка, содержащая информацию о системе обучения K-ближайших соседей для обнаружения OOD интентов.

Формат конечного результата

Файл в формате .docx, содержащий описание данных и методов, с помощью которых будешь дорабатывать модель, а также варианты оповещения пользователей.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.