цифровой прорыв

сезон: ИИ

КЕЙС

Семантическая классификация документов

Общество с ограниченной ответственностью
«Акселератор Возможностей»
при ИНТЦ МГУ «Воробьевы горы»

Министерство экономического развития Российской Федерации

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Кейсодержатель

Общество с ограниченной ответственностью «Акселератор Возможностей» при ИНТЦ МГУ «Воробьевы горы»

01 Сфера деятельности

Организация технологических и инвестиционных мероприятий, курирование инновационной деятельности внутри ИНТЦ МГУ «Воробьевы горы»

02 Краткое описание кейса

Создание программного модуля для семантической классификации документов по тексту.

Сайт организации

https://ac-vo.ru/

Постановка задачи

На основе входящего документа или его части, с применением технологий искусственного интеллекта, создать MVP в виде программного (программно-аппаратного) модуля определения типа документа с максимально возможной точностью.

Проблематика

Каждый день мы сталкиваемся с документами, как в работе, так и в жизни. Либо мы загружаем, либо нам их присылают и необходимо каждый раз проверять то ли нам прислали.
Например, для подачи документов (гос. органы, банк, контрагенты) всегда есть перечень документов, которые необходимо загрузить. На каждую загрузку, прежде чем отправить в работу, необходимо сотруднику проверить по списку все ли есть. Если нет, то повторить все действия: запрос - загрузка - проверка. Каждая итерация занимает определенное время, где-то час, а где-то и пару дней. Мы можем решить эту задачу, определяя необходимый перечень документов при загрузке и автоматически проверять содержимое. А если мы знаем содержимое, то можем сразу раскладывать необходимые документы автоматически в архив, помечая где какой.

Проблематика

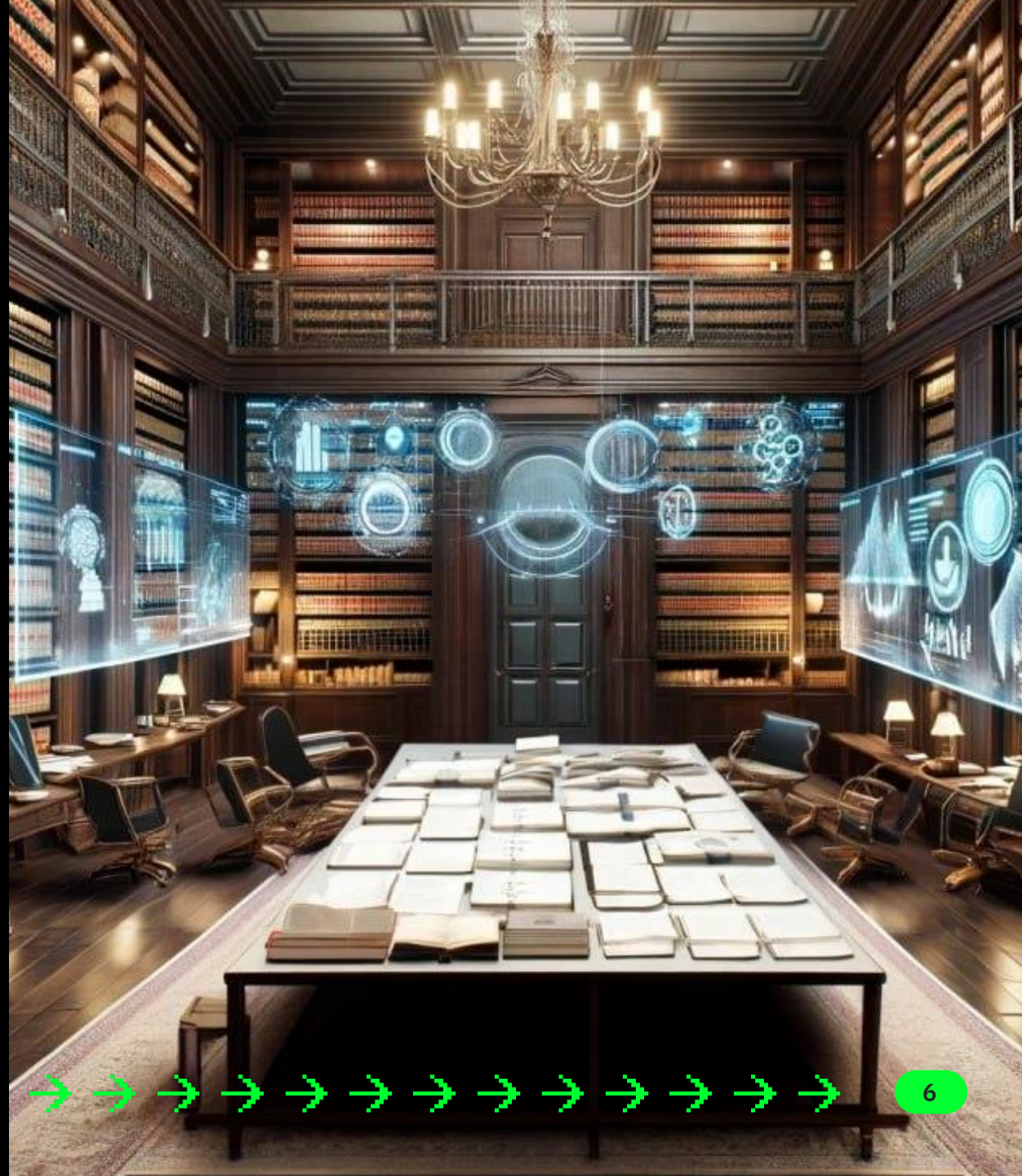
1) При обмене файлами в сети интернет необходимо тратить время на проверку загружаемых файлов пользователями. После каждой загрузки необходимо скачать файл и задействовать сотрудника чтобы посмотреть, тот ли документ загружен или пользователь перепутал и прислал другой.

2) Когда файлы уже имеется архив документов и необходимо навести порядок. Используя стандартные возможности системы мы имеем метаданные, по которым можем сортировать (дата, тип), но а как сортировать по содержимому. Счет может быть и в формате Word, Excel, PDF и т.п.

Решение

На входе мы имеем окно загрузки файлов (.pdf, .docx, .xls) с необходимым перечнем документов. После загрузки мы получаем результат, какие файлы были загружены, а каких не хватает.

Необходимо обратить внимание на то, что документы могут быть похожи друг на друга по тексту и структуре но семантически отличаться. На финальной проверке результатов, это будет учтено.



Стек технологий, обязательных к использованию

01

Python

Необходимые данные, дополнения, пояснения, уточнения

02

nlp-course-20/multiclass.ipynb at master · Metafiz/nlp-course-20 · GitHub
Многоклассовая и многозадачная классификация / Хабр (habr.com)
Открытый курс машинного обучения. Тема 3. Классификация, деревья решений и метод ближайших соседей / Хабр (habr.com)
Кластеризация и классификация больших Текстовых данных с помощью машинного обучения на Java. Статья #2 — Алгоритмы / Хабр (habr.com)
What is Text Classification? - Hugging Face
Classification — DeepPavlov 1.5.0 documentation
1.5. Stochastic Gradient Descent — scikit-learn 1.4.1 documentation
Извлечение фактов. Синонимия и омонимия / Хабр (habr.com)

Оценка

- Для оценки решений применяется метод экспертных оценок и автоматизированные средства оценивания.
- Жюри состоит из отраслевых и технических членов жюри.
- На основании описанных ниже характеристик, жюри выставляет оценки.
- Итоговая оценка определяется как сумма баллов всех членов жюри, умноженная на оценку автоматизированной системы.

Технический член жюри оценивает решение по следующим критериям:

01

Запускаемость кода

02

Обоснованность выбранного метода (описание подходов к решению, их обоснование и релевантность задаче)

03

Интегрируемость решения

04

Масштабируемость решения

05

Выступление команды (умение презентовать результаты своей работы, строить логичный, понятный и интересный рассказ для презентации результатов своей работы)

Автоматизированные средства оценивания точности работы предложенных участниками алгоритмов (решений) выставляют оценку в диапазоне 0-1, где 1 равно 100% точности работы решения.

Итоговая оценка определяется как итоговый балл жюри, умноженный на оценку автоматизированной системы.

Отраслевой член жюри оценивает решение по следующим критериям:

01

Релевантность поставленной задаче (команда погрузилась в отрасль, проблематику; предложенное решение соответствует поставленной задаче; проблема и решение структурированы)

02

Уровень реализации (концепция/ прототип и т.д.)

03

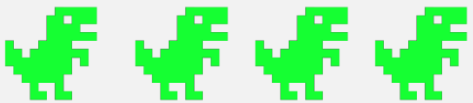
Пользовательский интерфейс

04

Адаптивность к разнообразию документов

05

Выступление команды (умение презентовать результаты своей работы, строить логичный, понятный и интересный рассказ для презентации результатов своей работы)



цифровой прорыв

сезон: ИИ

Министерство
экономического развития
Российской Федерации

РОССИЯ –
СТРАНА
ВОЗМОЖНОСТЕЙ