

Методические материалы к хакатону «Zakupki.Hack»

1. Условие задачи

На основе данных ЕЭТП «Росэлторг» об участиях и победах поставщиков в торгах за 2019 и 2020 год по реализовать рекомендательную систему, которая будет составлять релевантные рекомендации из актуальных тендеров для поставщиков. (далее тендер – процедура). Для одного поставщика из обучающей выборки необходимо рекомендовать не более 35 актуальных процедур. Участники получают обезличенные данные о процедурах и участниках. Train data – содержит описание для уникальных процедур, в которых участвовали поставщики. Train labels – содержит пары, которые позволяют связать поставщика и процедуру, в которой он участвовал. Test data – актуальные процедуры, из которых не более 35 необходимо рекомендовать поставщику. Команда формирует .csv файл (формат описан ниже) с рекомендациями и отправляет организаторам.

2. Данные

train_data – Уникальные процедуры за период обучения

test_data – Уникальные актуальные процедуры (которые мы рекомендуем)

pn_lot_anon – Анонимизированный номер связи процедура-лот

fz – Федеральный закон, к которому относится процедура

region_code – Код региона (справочник регионов приложен)

min_publish_date – Дата первой публикации извещения

purchase_name – Название закупки

lot_name – Название лота

lot_price – Цена лота (в руб.)

okpd2_code – Код ОКПД2

okpd2_names – Название кода ОКПД2 (разделитель “|”)

additional_code – Добавочный код (если нет ОКПД2 кода, используется КТРУ для 44 ФЗ или ОКВЭД2 для 223 ФЗ)

additional_code_names – Название добавочного кода (разделитель “|”)

item_descriptions – Описание товаров (разделитель “|”)

Код ОКПД2 имеет несколько уровней вложенности. Например, 24.20.31.000 – Трубы сварные для нефте- и газопроводов, наружным диаметром не более 406,4 мм, стальные. В поле okpd2_names – название кода ОКПД2 всегда соответствует реальному уровню вложенности, а в okpd2_code – код ОКПД2 верхнеуровневый. В нашем примере okpd2_names сохраняется, но okpd2_code будет 24.2 (не больше одного знака после первой точки). Аналогично для additional_code и additional_code_names.

train_labels – Разметка, связь поставщика с процедурой, в которой он участвовал

pn_lot_anon – Анонимизированный номер связи процедура-лот из train_data

participant_inn_kpp_anon – Анонимизированный ИНН_КПП поставщика

is_winner – 1 - если поставщик победил в соответствующей процедуре

0 – поставщик участвовал в соответствующей процедуре

fz – Федеральный закон, к которому относится процедура

Участники получают файлы:

train_data.csv – Уникальные процедуры за период обучения

train_labels.csv – Разметка

test_data.csv - Уникальные актуальные процедуры

team_name.csv – Шаблон файла для отправки решения

Коды регионов – Справочник для связи кода региона с названием

Справочник ОКПД2 – Справочник для связи кода ОКПД2 с названием

3. Метрики

1. Полнота – кол-во фактических поданных заявок из рекомендаций / кол-во участия
2. Точность – кол-во фактических поданных заявок из рекомендаций / кол-во рекомендаций
3. Процент рекомендованных актуальных процедур
4. Процент покрытия поставщиков из обучающей выборки

Основной бизнес метрикой является полнота, при условии, что не более 35 актуальных процедур мы рекомендуем одному поставщику. Метрики 2-4 важны при защите проектов на 1 и 2 этапе подведения итогов.

4. Формат решений

В специальном формате участники формируют .csv файлы с рекомендациями для тестовой выборки и отправляют их организаторам конкурса. (см. team_name.csv – Шаблон файла для отправки решения). С помощью тестирующей системы происходит проверка файла и оценка результатов. Если файл не прошёл проверку, то участники уведомляются о причинах, им даётся не более 15 мин на их устранение (06.02.2021)

Формат файла – csv.

Разделитель – “;”.

Название файла – название команды.

Поля – inn_kpp, actual_recommended_pn_lot, similarity_score. (inn_kpp – Анонимизированный ИНН_КПП поставщика, actual_recommended_pn_lot – Анонимизированный номер связки процедура-лот, similarity_score – числовой критерий показывающий релевантность актуальной процедуры для поставщика)

Пример файла будет приложен.

Целевая метрика бизнеса – это полнота, по ней формируется рейтинговая таблица, команды сортируются по убыванию метрики. Таблица публикуется в открытом доступе и демонстрируется участникам.

5. Этапы и критерии оценивания

Предварительная защита проектов (максимум 20 баллов)

Критерии предварительной защиты:

1. Уровень понимания участником сферы тендерных закупок
2. Учёт ОКПД2, регионов, Ф3 и других характеристик тендера в решении
3. Корректность методологии рекомендательной системы
4. Корректность обработки и подготовки данных для рекомендательной системы
5. Понимание проблемных мест решения и возможностей улучшения

Оценка показаний метрик проектов (максимум 80 баллов)

Происходит по формуле:

$$points_i = \frac{score_i}{score_{max}} * 80, \text{ где}$$

$points_i$ – баллы i – ого участника,

$score_i$ – значением метрики i – ого участника,

$score_{max}$ – максимальное значение метрики в рейтинговой таблице,

i – номер участника в рейтинговой таблице.

После проведения полуфинала баллы всех команд складываются и выявляются до 5 команд с наилучшими результатами. В рамках финала команды с наилучшими результатами презентуют свои решения членам жюри. Победителем признается тот проект и команда, который набрал наибольшее суммарное количество баллов у всех членов жюри (максимальное количество баллов 100).

Критерии оценки жюри:

1. Работоспособность прототипа - 20
2. Оригинальность идеи - 20
3. Масштабируемость - 20
4. Уровень понимания участником сферы тендерных закупок – 20
5. Учёт ОКПД2, регионов, Ф3 и других характеристик тендера в решении – 20

6. Полезные ссылки

<https://colab.research.google.com> – Среда разработки от Google

<https://scikit-learn.org/> - Python библиотека для машинного обучения

https://radimrehurek.com/gensim_3.8.3/index.html – Python библиотека для работы с текстом и создания текстовых моделей

<https://fasttext.cc/> - Python библиотека для создания текстовых моделей

<https://docs.python.org/3.8/> - Документация Python

<https://www.anaconda.com/> - Дистрибутив языков Python и R, включающий библиотеки для машинного обучения

<https://jupyter.org/> - Среда разработки для анализа данных