**Проектное предложение**

|  |  |
| --- | --- |
| Тип проекта | *Исследование, реализация* |
| Название проекта | *Кластеризатор данных о компании* |
| Подразделение инициатор проекта | ООО Яндекс, группа разработки бэкофиса  |
| Руководитель проекта | *Сухов Павел* |
| **Подробное описание содержания проектной работы** | Данная проектная работа является модулем **Справочника Организаций** а Яндексе. В справочнике организаций данные о каждой организации(компании) представляются в индивидуальной структуре данных – **Карточке Организации**. У справочника организаций есть различные критерии качества, такие как:* **Точность** – карточке, находящейсе в справочнике, соответствует некоторая существующая в реальности.
* **Полнота** – существующая организация есть в справочнике
* **Задублированность** – на существующую организацию существует ровно одна карточка организаций.

Одной из самых больших проблем хранения данных об организациях, является выявление карточек-**дублей**. Данные о карточках постоянно меняются, в результате чего дубли могут как придти извне, так и образоваться внутри самого справочника. Задача студентов – собрать механизм по выявлению дублей, и доказать его эффективность на данных, близких к реальным. К нему относится: **Формула** - машиннообученный классификатор, который “достаточно достоверно” может сказать, являются ли две карточки дублями. Считается, что у этого классификатора при очень высокой точности недостаточно высокая производительность для вычислений на масштабах порядка всего Справочника. **Эвристика** для поиска “гипотез” на задублированность – эвристический алгоритм, который позволит из всех данных справочника выбрать только те организации, которые гипотетически могут быть дублями. Здесь критичны ошибки второго рода (не найти карточку, если она дубль), и количесво карточек переданных “формуле” (много он обработать не сможет). **Инфраструктура** – механизм, позволяющий отслеживать изменения некоторого рода в справочнике и анализировать появление дублей “реал-тайм”, с отправкой данных на ручную проверку и автоматическим дообучением формулы.  |
| **Цель и задачи проекта** | Создание механизма по выявлению дубликатов. |
| **Проектное задание (виды деятельности, выполняемые студентом в проекте)** | * Исследование предметной области.
* Создание алгоритма машинного обучения сравнения «похожести» карточек организаций.
* Проведение тестов и доказательства эффективности алгоритма.
* Создание эвристического алгоритма подборки «гипотез на задублированность».
* Проведение тестов качества эвристического алгоритма.
* *Создание инфраструктуры, позволяющей проводить постоянные проверки на появлении дублей в Справочнике Организаций.\**

\* - факультативное задание, либо может быть заданием для отдельной группы. |
| Сроки реализации проекта | 1 сентября 2016 – 31 мая 2017 г. |
| Количество кредитов | *2* |
| Форма итогового контроля | *Защита алгоритма. Прохождение внутреннего тестирования качества.* |
| Тип занятости студента | *Работа дома, периодические встречи в компании* |
| Интенсивность (часы в неделю) | *8 с каждого, ну или больше, если кого затянет.* |
| Вид проектной деятельности | *Групповая*  |
| Требования к студентам, участникам проекта | *Хорошее знание одного из языков: С++, Python, Java, Scala. (могу проверить при первой встречи).* Приличные знания основ математического анализа, матстатистики, базовых алгоритмов. Знание машинного обучения хотябы одного из членов группы, или желание потратить огромное количество сил по самостоятельному изучению этой области.Приветствуется знание git, опыт использования github.com.  |
| **Планируемые результаты проекта** | *Исходный код по анализу данных о компания,* *результаты оценки качества алгоритмов.* |
| **Формат представления результатов, который подлежит оцениванию**  | *результаты оценки качества алгоритмов* |
| **Критерии оценивания результатов проекта** | *Формальные критерии алгоритмов: ошибки 1-ого и 2-ого рода, скорость работы, и т.п.* |
| **Возможность пересдач при получении неудовлетворительной оценки** | *Нет* |
| Количество вакантных мест на проекте | *4* |
| Критерии отбора студентов в проект (применяются в случае большого количества заявок на проект) | *Знание математики и математической статистики**Знания машинного обучения**Глубокое знание языков программирования и умение оптимизировать скорость работы программы.*  |
| Образовательные программы |  |
| Территория |  |