Лабораторное задание №2

Аннотация

Применение ML-подходов к решению практически значимых индустриальных и промышленных задач – одно из наиболее перспективных направлений в современном мире. Большое количество накопленных производственных данных формирует ценный для анализа экспертный пласт, который может служить основой как для создания автоматизированных инструментов, использующих эти данные, так и для создания широкого спектра систем поддержки и принятия решений для людей-экспертов в различных областях. Однако, ключевой проблемой создания моделей на таких данных является необходимость проведения их комплексной предобработки. Зачастую такая предобработка не ограничивается классическими методами, такими как устранение шумов и выбросов, заполнение пропусков, нормализация и базовая обработка текстовых данных и требует разработки отдельной полноценной модели для эффективной подготовки данных к последующей работе и конструированию предметных моделей.

Задание

Вам будет предоставлен набор данных о задачах, которые выполнялись в рамках строительства капитальных объектов на месторождениях нефти и газа. Набор содержит информацию о примерно 716 тысячах задач. Для каждой из задач доступна информация о ее названии в строительном плане, а также частично заданная информация об иерархии задач и обобщенных классах наименований, к которым относятся эти задачи (двух разных степеней детализации).

Используя эти данные, вам необходимо будет разработать семантическую модель, которая позволяла бы эффективно определять обобщенные классы для задач, у которых эта информация не представлена.

Каждая задача описывается следующими атрибутами.

- **work_name** – Текстовое название задачи в строительном плане (без предобработки).
- **upper_works** – Информация об иерархии названий объектов и блоков работ, в рамках которых выполнялась эта задача. Если задачи имеют одинаковое значение этого атрибута – это означает, что они выполнялись в рамках одного блока работ над одним объектом (может быть пустым).
- **generalized_work_class** – Информация об обобщенном классе наименований работ, к которому относится задача (может быть пустым).
- **global_work_class** – Информация о самом высоком уровне обобщения названия задачи (может быть пустым).

Примечания

- Классификацию неразмеченных данных необходимо проводить на те классы, которые присутствуют среди размеченных.
- Стоит обратить внимание на то, что в названиях задач присутствует большое количество опечаток и грамматических ошибок.
- Выбор метрики оценки качества выполнения задания остается на усмотрение команды и презентуется на защите.

Тестовая выборка будет выложена за неделю до защиты – 18.10.2023

Laboratory task №2

Introduction

Application of ML-approaches for solving practically significant industrial tasks is one of the most promising directions in the modern world. A large amount of accumulated industrial data forms an expert layer valuable for analysis, which can serve as a basis both for the creation of automated tools that use this data and for the creation of a wide range of support and decision-making systems for human experts in various fields. However, the key problem of creating models on such data is the need for complex preprocessing. Often such preprocessing is not limited to classical methods, such as noise and outliers' removal, skip filling, normalization, and basic text data processing, and requires development of a separate full-fledged model to effectively prepare the data for further work and construction of subject models.

Assignment

You will be provided with a dataset of tasks that were performed as part of the construction of capital facilities at oil and gas fields. The set contains information about approximately 716 thousand tasks. For each of the tasks, information about its name in the construction plan is available, as well as partially specified information about the task hierarchy and the generalized name classes to which the tasks belong (two different levels of detail).

Using this data, you will need to develop a semantic model that allows you to efficiently define generalized classes for tasks that do not have this information.

Each task is described by the following attributes.

- **work_name** – Text name of the task in the construction plan (without preprocessing).
- **upper_works** – Information about the hierarchy of names of objects and work blocks within which this task was performed. If tasks have the same value of this attribute - it means that they were performed within one block of works on one object (may be empty).
- **generalized_work_class** – Information about the generalized work name class to which the task belongs (may be empty).
- **global_work_class** – Information about the highest level of generalization of the task name (may be empty).

Notes

- Classification of unlabeled data should be done into those classes that are present among the labeled data.
- It should be noted that there are a lot of typos and grammatical errors in the task names.
- The choice of a metric for assessing the quality of the assignment is left to the team's discretion and will be presented at the defense.
- All the works names are in Russian, so please use NLP models specialized for working with Russian data (there are a lot of them).

The test dataset for will be published the week before the defense – on 18.10.23