**Assignment №2**

**Overview**

Over the past decade, there has been a noticeable trend toward the digitalization and automation of tasks that were previously performed manually. This trend has also affected expert areas such as planning for the development of oil and gas fields. While previously the task of creating work plans in this field was solved manually by domain experts, nowadays, many of its components are automated and based on optimization algorithms and expert systems built using real data from completed projects. However, the lack of unified data representation formats can make working with data in digital form challenging. This is especially true for task names in natural language, as different experts may use various terms and synonymous expressions. In this assignment, you will need to cluster standard task names into common groups and implement an algorithm for semantic mapping of task names (together with measurement units) to standardized names.

**Task Description and Data**

To complete the tasks, you will have a dataset with 1,242 standardized task names performed as part of oil and gas field development. For each standard task, the following information is provided:

* **task\_id** – a unique identifier for the task;
* **task\_name** – the standardized name of the task in natural language;
* **measurement** – the unit of measurement used for the task's volume;
* **subtasks** – additional information about the composition of tasks included in the standard task and additional conditions for performing the task (unstructured, in natural language).

**Task 1. Clustering and Generalization of Standard Tasks**

The first part of the assignment involves working with only the ‘task\_name’ and ‘measurement’ columns. Many tasks in the standardized dataset differ only by additional parameters, such as the pipeline diameter (Figure 1) or the specifics of the work process (type of soil, method, and place of development, as shown in Figure 2). Such tasks can be generalized to the following names: "Pipeline Trench Development" for tasks in Figure 1 and "Soil Development" for tasks in Figure 2.



**Figure 1** – Example of tasks belonging to the "Pipeline Trench Development" group

*Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание***Figure 2** – Example of tasks belonging to the "Soil Development" group

Therefore, in the first part of the assignment, you need to cluster all the presented standard task names and determine a generalized name for each. After completing the assignment, add a column named **general\_name** to the dataset and include the obtained generalized names, which will be used in the second part of the assignment.

**Task 2. Semantic Mapping of Task Names**

In the second part, you need to implement an algorithm that performs the mapping (matching) of given task names and measurement units to the standardized names in the provided dataset and the generalized names obtained from the first part of the assignment. The algorithm should take a CSV file with the columns **task\_name** and **measurement** as input and return an extended dataset that adds the columns **general\_name** (with the corresponding generalized names) and **standard\_name** (with the standardized names from the original dataset). If a detailed standardized name cannot be matched due to the absence of additional information in the text, the **standard\_name** field should duplicate the value from **general\_name**.

**Assessment**

The defense of this assignment will take place on **December 11, 2024**, in the same format (online presentation) as the first assignment defense.

Two weeks before – on **November 27, 2024**, we will release a test dataset on which you will need to run the algorithm you developed and obtain the corresponding generalized name (**general\_name**) and specific standardized name (**task\_name**) for each presented task name and its unit of measurement, when additional parameters in the task name allow such mapping.

One week before – on **December 4, 2024**, you will need to **submit the results** from running the test dataset so that we can review them.

**Лабораторное задание №2**

**Аннотация**

За последние 10 лет активно наблюдается тренд на цифровизацию и автоматизацию задач, которые ранее выполнялись только вручную. Это затронуло и экспертные области, такие как планирование обустройства месторождений нефти и газа. Если раньше задача построения планов выполнения работ в этой области решалась вручную экспертами предметной области, сейчас многие ее составные части автоматизированы и основываются на алгоритмах оптимизации и экспертных системах, построенных на основе реальных данных уже выполненных проектов. Тем не менее, из-за отсутствия единых форматов представления данных работа с ними в цифровом виде может быть затруднена. Особенно это касается названий работ и на естественном языке, поскольку разные эксперты могут использовать в них разные термины и синонимичные выражения. В данном задании вам предстоит кластеризовать нормативные названия задач по общим группам и реализовать алгоритм семантического маппинга названий работ (в паре с единицами измерения) на нормативные названия.

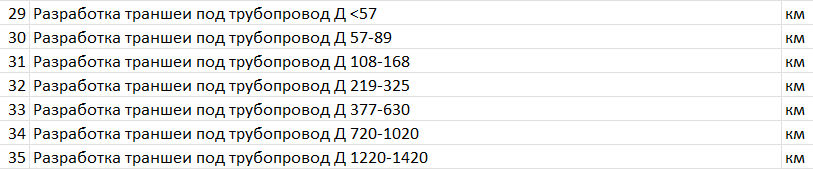
**Задание и описание данных**

Для выполнения заданий у вас будет датасет с 1242 нормативными наименованиями работ, которые выполняются в рамках обустройства месторождений нефти и газа. Для каждой нормативной работы представлены:

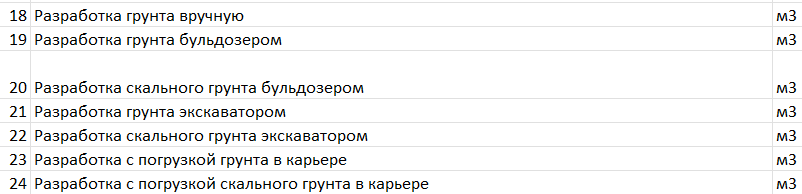
* task\_id – порядковый идентификатор работы;
* task\_name – нормативное название работы на естественном языке;
* measurement – единица измерения, в которой измеряется объем работы;
* subtasks – дополнительная информация по составу задач, входящих в нормативную работу, и дополнительным условиям выполнения работы (неструктурированная, на естественном языке).

**Задание 1. Кластеризация и обобщение нормативных работ**

Первая часть задания подразумевает работу только с колонками **task\_name** и **measurement**. Большое количество задач, представленных в датасете нормативных наименований, отличаются только дополнительными параметрами, такими как диаметр трубопровода (рисунок 1) или особенности процесса выполнения работ (вид грунта, способ и место разработки на рисунке 2). При этом такие задачи можно обобщить до следующих названий: «Разработка траншеи под трубопровод» для задач с рисунка 1 и «Разработка грунта» для задач с рисунка 2.



**Рисунок 1** – Пример задач, относящихся к группе «Разработка траншеи под трубопровод»



**Рисунок 2** – Пример задач, относящихся к группе «Разработка грунта»

Поэтому в рамках первого задания вам необходимо будет провести кластеризацию всех представленных нормативных названий работ и для каждого такого названия определить обобщенное название, которое ему соответствует.

После выполнения задания добавьте в датасет колонку **general\_name** и включите туда полученные обобщенные наименования. Они будут использоваться вами при выполнении второй части задания.

**Задание 2. Семантический маппинг наименований работ**

Во второй части задания вам необходимо будет реализовать алгоритм, выполняющий маппинг (сопоставление) переданных названий работ и единиц измерения к нормативным наименованиям в представленном датасете и обобщенным наименованиям, полученным по результатам выполнения первой части задания.

На вход алгоритм должен принимать csv-файл с колонками task\_name и measurement и возвращать расширенный датасет, в который будут добавлены колонки **general\_name** (с соответствующими обобщенными наименованиями) и **standard\_name** (с нормативными наименованиями – task\_name из изначального датасета). В случае, если нельзя сопоставить детализированное нормативное наименование для работы из-за отсутствия в тексте дополнительной информации, в поле standard\_name нужно продублировать значение из general\_name.

**Оценивание**

**Защита** данного задания будет происходить **11.12.2024** в таком же формате (онлайн презентация), как и защита первого задания.

**За две недели – 27.11.2024** мы выложим тестовый набор данных, на котором нужно будет запустить разработанный вами алгоритм и получить для каждой представленной в нем пары из наименования задачи и ее единицы измерения значение соответствующего ей обобщенного названия (general\_name) и конкретного нормативного названия (task\_name) для случаев, когда в названии задачи содержатся дополнительные параметры, позволяющие сделать такое сопоставление.

**За неделю – 04.12.2024** нужно будет отправить результаты запуска на тестовом датасете, чтобы у нас была возможность с ними ознакомиться.

**这是该实验任务的中文翻译：**

**实验任务 №2**

**摘要**

**在过去的10年里，数字化和自动化的趋势在许多以前仅由人工完成的任务中得到了显著体现。这一趋势也影响了像石油和天然气田开发规划这样的专家领域。过去，这一领域的工作计划通常由领域专家手动完成，而现在，许多组成部分已经通过优化算法和基于已完成项目的实际数据构建的专家系统实现了自动化。然而，由于缺乏统一的数据表示格式，处理数字化数据可能会遇到困难，尤其是在任务名称和自然语言描述上，因为不同的专家可能使用不同的术语和同义表达。在本次任务中，您需要将标准化任务名称聚类到相同的组中，并实现一个算法，执行任务名称（与测量单位一起）到标准名称的语义映射。**

**任务和数据描述**

**为完成任务，您将获得一个包含1242个石油和天然气田开发过程中执行的标准任务名称的数据集。对于每个标准任务，提供以下信息：**

* **task\_id – 任务的顺序标识符；**
* **task\_name – 任务的标准名称（自然语言）；**
* **measurement – 任务量的测量单位；**
* **subtasks – 有关任务组成和执行条件的附加信息（非结构化的自然语言）。**

**任务1：标准任务的聚类与归类**



**任务的第一部分仅涉及处理“task\_name”和“measurement”两列。数据集中许多任务的标准名称仅在附加参数上有所不同，例如管道直径（见图1）或工作过程的具体细节（如土壤类型、方法和开发地点，见图2）。这些任务可以归类为以下名称：“管道沟槽开发”（见图1中的任务）和“土壤开发”（见图2中的任务）。**

**图1 – 属于“管道沟槽开发”组的任务示例**

**图2 – 属于“土壤开发”组的任务示例**

**因此，在任务的第一部分，您需要对所有标准任务名称进行聚类，并确定每个任务对应的归类名称。完成任务后，向数据集中添加一个名为 general\_name 的新列，并将所获得的归类名称包含在其中，这将在任务的第二部分中使用。**

**任务2：任务名称的语义映射**

**在任务的第二部分，您需要实现一个算法，将给定的任务名称和测量单位映射到提供的数据集中的标准名称以及第一部分任务中获得的归类名称。该算法应接受一个包含 task\_name 和 measurement 列的CSV文件作为输入，并返回一个扩展的数据集，新增 general\_name 列（包含相应的归类名称）和 standard\_name 列（包含原始数据集中的标准任务名称）。如果由于文本中缺乏附加信息而无法匹配详细的标准任务名称，则 standard\_name 字段应与 general\_name 字段的值相同。**

**评估**

**本任务的答辩将在2024年12月11日进行，答辩形式与第一次任务答辩相同（在线展示）。**

* **提前两周 – 2024年11月27日，我们将发布一个测试数据集，您需要运行开发的算法，并为其中每一对任务名称和测量单位获取对应的归类名称 (general\_name) 和具体标准任务名称 (task\_name)，特别是当任务名称中包含附加参数时，可以进行这样的映射。**
* **提前一周 – 2024年12月4日，您需要提交运行测试数据集的结果，以便我们进行审查。**

**如果您需要进一步的帮助，请告诉我！**