# Задание 1.

## Приведите классификацию формальных грамматик Хомского с примерами для категорий 2 и 3.

Классификация формальных грамматик Хомского:

* Тип 0 – неограниченные.
* Тип 1 – контекстно-зависимые.
* Тип 2 – контекстно-свободные. Примером может служить любой язык программирования. Программу можно проанализировать с точки зрения слов, то есть посмотреть, какие встречаются условия и т. д. (лексический анализ) и понять их взаимосвязь (синтаксический анализ).
* Тип 3 – регулярные. Например, регулярные выражения позволяют находить определенные шаблоны в тексте.

## Что такое бейзлайн, пайплайн, SOTA? Приведите примеры.

* **Бейзлайн** – самое простое решение задачи. Например, в контексте автоматического определения тональности текста, бейзлайн может заключаться в создании словаря положительных и отрицательных слов и анализе того, каких слов в тексте встретилось больше.
* **Пайплайн** – этапы обработки данных. Например, в NLP эти этапы включают в себя токенизацию, нормализацию, удаление стоп-слов, стемминг/лемматизацию, POS-tagging.
* **SOTA** – лучшее на текущий момент решение в определенной области. Примером SOTA может быть модель нейронной сети, которая достигает самых высоких показателей в задачах обработки естественного языка, например, в машинном переводе или распознавании речи.

## Какие элементы имплементации регулярного языка PCRE не являются собственно элементами грамматики регулярного языка по классификации Хомского?

Проверки (lookaheads и lookbehinds)

## Что такое языковая модель? Какие типы языковых моделей вы знаете?

Языковая модель – это алгоритм, позволяющий предсказать вероятность появления последовательности слов в тексте.

Типы языковых моделей:

* Статистические языковые модели (SLM)
* Нейронные языковые модели (NLM)
* Предобученные языковые модели (PLM)
* Большие языковые модели (LLM)

## Чем задача классификации отличается от задачи кластеризации?

* **Классификация** – распределение объектов по заранее определенным категориям. Например, определение части речи слова.
* **Кластеризация** – группировка объектов на основе сходства их характеристик, без заранее определенных категорий. Например, может использоваться для обнаружения аномалий.

# Задание 3.

*Выберите любую понравившуюся вам задачу NLP и исследуйте литературу по этой задаче. Какой у нее бейзлайн? Какая SOTA? В каком направлении ведутся современные исследования, связанные с этой задачей? Какие практические применения? Напишите коротенький конспект.*

**Автоматическое реферирование (Automatic summarization)** – это задача NLP, которая заключается в создании краткого и связного изложения текста без потери значимой информации. Данная задача является довольно сложной, поскольку для ее решения требуется как понимание текстов, так и их генерация.

**Бейзлайном** для автоматического реферирования является извлечение первых трех предложений из документа (“lead-3 baseline”).

На сегодняшний день **лучшей моделью** для автоматической суммаризации текста является Pegasus+DotProd. Она набирает 40.6 баллов по метрике ROUGE-1 на датасете GigaWord. ROUGE – это основная **метрика** оценки качества автоматического реферирования, которая вычисляет количество совпадающих n‑грамм, встречающихся и в образцовом, и в автоматическом рефератах.

**Современные исследования**:

* Сосредоточение на **абстрагирующих (генерирующих) подходах**. Существует два основных подхода к реферированию: *экстрагирующий* (выделение наиболее важных фрагментов из исходного текста) и *абстрагирующий* (переформулирование исходного текста для создания краткого содержания). Многие исследователи переключили свое внимание на абстрагирующие методы, поскольку они позволяют создавать более естественные и точные краткие изложения.
* **Семантический анализ и понимание текста**. Многие существующие системы по-прежнему ограничиваются выявлением часто встречающихся последовательностей слов без глубокой семантической обработки.
* Расширение **базы данных** для реферирования. Существует потребность в качественных базах данных для реферирования, особенно на языках, кроме английского.
* Улучшение **методов оценки** систем реферирования. Стандартно используемая метрика ROUGE далека от совершенства во многих аспектах.

**Практические применения**:

* Бизнес. Автоматическое реферирование отчетов, переписки и т. д.
* Новостные агрегаторы. Автоматическое создание кратких обзоров статей и новостей.
* Академические исследования. Автоматическое реферирование научных статей для быстрого обзора.

Автоматическое реферирование помогает анализировать большие объемы текстовых данных и выделять ключевую информацию, что может быть полезно в различных областях.