NLP решает большой набор задач, который можно разбить по уровням (в скобках). Среди этих задач, можно выделить следующие:

* Распознавание текста, речи, синтез речи (сигнал);
* Морфологический анализ, канонизация (слово);
* POS-тэгирование, распознавание именованных сущностей, выделение слов (словосочетание);
* Синтаксический разбор, токенизация предложений (предложение);
* Извлечение отношений, определение языка, анализ эмоциональной окраски (абзац);
* Аннотация документа, перевод, анализ тематики (документ);
* Дедубликация, информационный поиск (корпус).

Датасет представляет из себя нормально распределённый(близко к норме) jsonl таблицу, с вопросами, лейблами и ответами.

Подготовка датасета(перевод вопроса к основному тексту, привести к нижнему регистру, удаление пунктуации, токенезация)

* Векторизация слов
* Обучение

Ресурсы для решения задач - huggingface.co - , docs.deeppavlov.ai



Работа программы осуществляется следующим образом: word2vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе.

Сначала он генерирует словарь корпуса, а затем вычисляет векторное представление слов, «обучаясь» на входных текстах.

Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а следовательно, имеющие схожий смысл), будут иметь близкие (по косинусному расстоянию) векторы.

Полученные векторные представления слов могут быть использованы для обработки естественного языка и машинного обучения.

**BERT** — двунаправленная модель с transformer-архитектурой, заменившая собой последовательные по природе рекуррентные нейронные сети (LSTM и GRU), с более быстрым подходом на основе механизма внимания.

Модель также предобучена на двух задачах без учителя — моделирование языковых масок и предсказание следующего предложения.

Это позволяет нам использовать предобученную модель BERT, настраивая её под конкретные задачи, такие как определение эмоциональной окраски текста, вопросно-ответные системы и многие другие.

Но проблема – Bert показал результаты по метрике accuracy не сильно лучше, что честно говоря повергло в шок.



Базовый принцип работы t-SNE заключается в сокращении попарных расстояний между точками при сохранении их относительного расположения. Иными словами, алгоритм отображает многомерные данные на пространство более низкой размерности, при этом сохраняя структуру соседства точек.