Лабораторная работа 8

Тема: **Текстовые эмбендинги. Работа с предобученными моделями эмбендингов.**

Эмбендинги – это методы векторизации слов с использованием нейросетей. В отличие от рассмотренных в прошлой работе простых методов векторизации слов, эмбендинги рассматривают контекст слова, это позволяет учитывать семантику слов и, таким образом, удобно для поиска семантически близких слов и текстов. Принцип эмбендингов используется во внутренних механизмах современных LLM.

Существуем множество моделей текстовых эмбендингов. Ваша задача поработать минимум с двумя моделями и сравнить эффективность их работы на одинаковых задачах.

Сравнение качества эмбеддингов

1. **Загрузите предобученные эмбеддинги, например,** Word2Vec от Google News, GloVe от Stanford), FastText от Facebook или иные.
Для работы используйте как минимум две различные модели.

2. С использованием библиотеки gensim реализуйте скрипт, который:
   - ✓ ищет семантически схожие слова для какого-либо заданного слова;
   - ✓ выполняет аналогию типа "Париж – Эйфелева башня, Минск - ?" или "Италия – пицца, Беларусь - ? " или др.
   - ✓ **визуализирует векторы 20–30 слов с помощью t-SNE или PCA (по желанию).

3. - Сделайте выводы: какая модель лучше справляется с семантическими аналогиями?

Вопросы:

1. Опишите принцип выделения эмбендингов выбранных вами моделей.
2. Какой параметр характеризует семантическую близость слов?
3. Как можно обучить свои текстовые эмбендинги и какие библиотеки предоставляют модели для обучения?