**Министерство науки и высшего образования Российской Федерации**

**федеральное государственное автономное образовательное учреждение высшего образования**

**“НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО”**

**Факультет** Программной инженерии и компьютерной техники

**Направление подготовки (специальность)** Системное и прикладное программное обеспечение

### **ЭСсе** на тему:

# Курс “Machine Learning: Natural Language Processing in Python”

Обучающийся Орловский М.Ю P4116

(Фамилия И.О.) (номер группы)

Санкт-Петербург

2023 г.

Этот курс предназначен в первую очередь для обучения методам обработки текста и способам применения векторных моделей для анализа извлеченных данных. Основной упор делается на методы преобразования текста в числовые представления, будь то метрики или просто преобразование. В курсе также рассматриваются основные задачи обработки естественного языка, будь то обнаружение спама или генерация статей.

Особое внимание уделяется векторизации текста, где описываются методы подсчета частотности и применение TF-IDF для определения важности терминов в документе. Эти навыки позволяют эффективно представлять текстовые данные в виде числовых векторов. Кроме того, курс знакомит с различными векторными представлениями текста, включая Word2vec, GloVe и FastText.

Также в рамках курса представлены различные методы предварительной обработки текста, такие как токенизация, удаление стоп-слов, стемминг и лемматизация. Все эти методы демонстрируются на практике.

Не обошли стороной и нейронные сети и их применение в обработке естественного языка. Хотя основное внимание курса не сосредоточено на нейронных сетях, он дает общее представление об их применении и помогает систематизировать знания. Описывается устройство моделей CNN, RNN и Transformer. Заинтересовали модели CodeBERT и GraphCodeBERT, т. к. они ориентированы на работу с программным кодом, что могло бы пригодиться в моем проекте.

Для моей научной работы данный курс является отличным вводным курсом, который помогает изучать научные статьи, связанные с анализом кода при использовании методов NLP. Один из интересных материалов, с которым я ознакомился, это статья [hindle2016naturalness](https://www.researchgate.net/publication/254041552_On_the_naturalness_of_software), где авторы представляют доказательства, подтверждающие, что программное обеспечение, несмотря на свою теоретическую сложность, в действительности обладает множеством закономерностей, которые можно отразить с помощью даже простой статистической модели. Они утверждают, что большинство программных продуктов также имеют естественные черты, поскольку они создаются людьми на работе с учетом всех сопутствующих ограничений. Поэтому, аналогично естественному языку, программный код также может быть подвержен анализу с использованием методов обработки естественного языка.

По итогам прохождения данного курса я пришел к выводу, что область обработки естественного языка имеет большое практическое применение в моей предметной области. Она достойна внимания и, вероятно, может найти свое применение в проекте, хотя и не в основной роли.

Также, прикладываю сертификат о выполнении

A picture containing text, screenshot, font

Description automatically generated