

Компютърна Лингвистика в Областта на Биоинформатиката

гл. ас. д-р Георги Петров Пашев

<http://gpashev.com> § georgepashev@uni-plovdiv.bg

Кратко описание на дисциплината:

Този курс се фокусира върху приложението на компютърна лингвистика в областта на биоинформатиката, като се разглеждат техники за обработка на естествен език и машинно обучение, приложими в анализа на биологични данни. Курсът съчетава теоретични знания с практически умения в програмиране на Python и Rust, като изследва интеграцията на тези езици в биоинформатични приложения.

Списък с теми за лекциите (7 седмици):

00001. Въведение в Компютърната Лингвистика и Биоинформатика

- Основни понятия и приложения
- Примерен код: Обработка на текстови данни с Python

00002. Машинно Обучение в Биоинформатиката

- Основи на машинното обучение
- Примерен код: Създаване на базов класификатор в Python

00003. Естествен Езиков Обработка (NLP) и Геномни Данни

- NLP техники за анализ на геномни последователности
- Примерен код: NLP алгоритми в Rust

00004. Биоинформатични Базии Данни и Тяхната Интеграция

- Работа с биологични бази данни
- Примерен код: Интегриране на бази данни в Python

00005. Използване на GPT-4 и LLAMA2 за Анализ на Биологични Данни

- Приложения на модели за естествен език в биоинформатиката
- Примерен код: Интегриране на GPT-4 API в Python

00006. Развитие на Уеб и Десктоп Интерфейси за Биоинформатични Приложения

- Основи на уеб и десктоп програмирането
- Примерен код: Създаване на прост UI в Rust

00007. Проектна Работа и Научно Писане

- Техники и методи за разработване на научни публикации
- Разработване на проектни идеи и начало на проектната работа

Оценяване:

Студентите работят в малки групи, разработвайки проект, който включва създаване на чат бот, използващ GPT-4 API, LLAMA2, или машинно обучение с Python и sklearn. Всеки проект включва научна публикация и разработка на потребителски интерфейс.

Възможно е студентите да изберат да разработват само обзорно-научната част на проекта, без практическата, ако не се чувстват достатъчно силни в областта на разработката на приложения. В такъв случай, максималната оценка, която могат да получат, е **Много Добър (5)**.

Списък с Упражнения за Всяка Седмица

Седмица 1: Въведение в Компютърната Лингвистика и Биоинформатика

- **Упражнение:** Анализиране и обработка на текстови данни с Python. Студентите трябва да напишат скрипт за извличане на информация от научни публикации, свързани с биоинформатиката.

Седмица 2: Машинно Обучение в Биоинформатиката

- **Упражнение:** Създаване на базов класификатор в Python за разпознаване на видове ДНК последователности. Студентите използват предоставени данни за трениране на модела.

Седмица 3: Естествен Езиков Обработка (NLP) и Геномни Данни

- **Упражнение:** Имплементиране на NLP алгоритми в Rust за анализ на геномни последователности. Студентите трябва да адаптират традиционни NLP методи за работа с генетичен код.

Седмица 4: Биоинформатични Базии Данни и Тяхната Интеграция

- **Упражнение:** Интегриране на биологични бази данни в Python програма. Студентите създават интерфейс за достъп и анализ на данни от биоинформатични бази данни.

Седмица 5: Използване на GPT-4 и LLAMA2 за Анализ на Биологични Данни

- **Упражнение:** Работа с GPT-4 API или LLAMA2 за генериране на предсказания или анализи въз основа на биологични данни. Студентите разработват прототип на приложение, използващо тези модели.

Седмица 6: Развитие на Уеб и Десктоп Интерфейси за Биоинформатични Приложения

- **Упражнение:** Създаване на прост уеб или десктоп базиран UI в Rust, който да показва резултати от биоинформатични анализи. Студентите трябва да интегрират своя код в потребителски интерфейс.

Седмица 7: Проектна Работа и Научно Писане

- **Упражнение:** Начална разработка на проекта и написване на научен обзор за избраната тема. Студентите трябва да определят обхвата на своя проект и да изготвят план за научна публикация.

Тези упражнения са насочени към развитие на практически умения и подготовка за финалния проект, като същевременно се засилва теоретичната основа на дисциплината.

Литература:

1. Mount, D. W. (2004). Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press.
2. Pevsner, J. (2015). Bioinformatics and Functional Genomics. Wiley-Blackwell.
3. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
4. Klabnik, S., & Nichols, C. (2019). The Rust Programming Language. No Starch Press.
5. McKinney, W. (2018). Python for Data Analysis. O'Reilly Media.
6. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.