МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

**Лабораторная работа №\_\_8\_\_**

по дисциплине«Методы машинного обучения в автоматизированных системах обработки информации и управления»

Тема: « Изучение методов предобработки текстов»

ИСПОЛНИТЕЛЬ: \_\_Се Цзявэнь\_\_\_\_

ФИО

Группа ИУ5И-22М\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_30\_"\_\_\_\_\_04\_\_\_\_2024\_ г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Задания

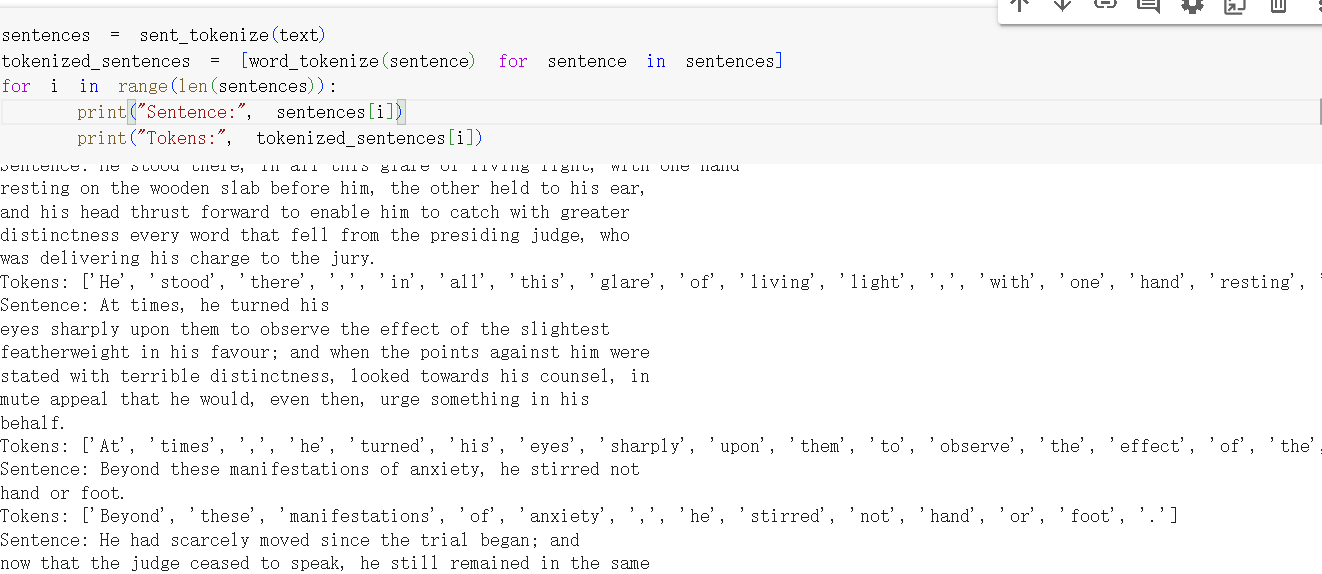
Для произвольного предложения или текста решите следующие задачи:

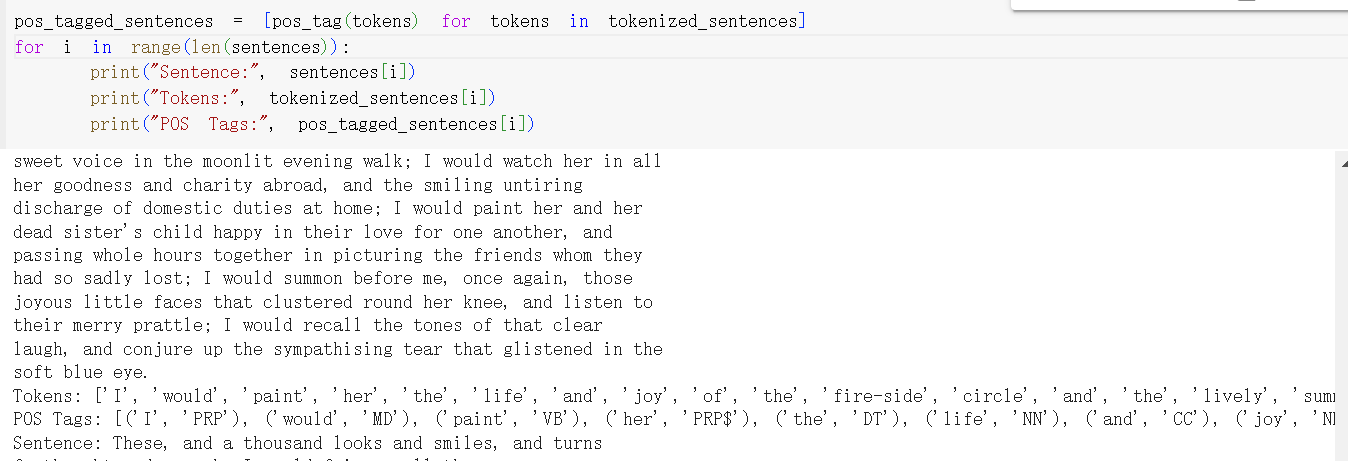
* Токенизация.
* Частеречная разметка.
* Лемматизация.
* Выделение (распознавание) именованных сущностей.
* Разбор предложения.

# Программы

* Загрузите текст:

Текст, который я здесь использую, представляет собой текстовую версию книги Оливера Твиста на английском языке..

* Токенизация. 
* Частеречная разметка



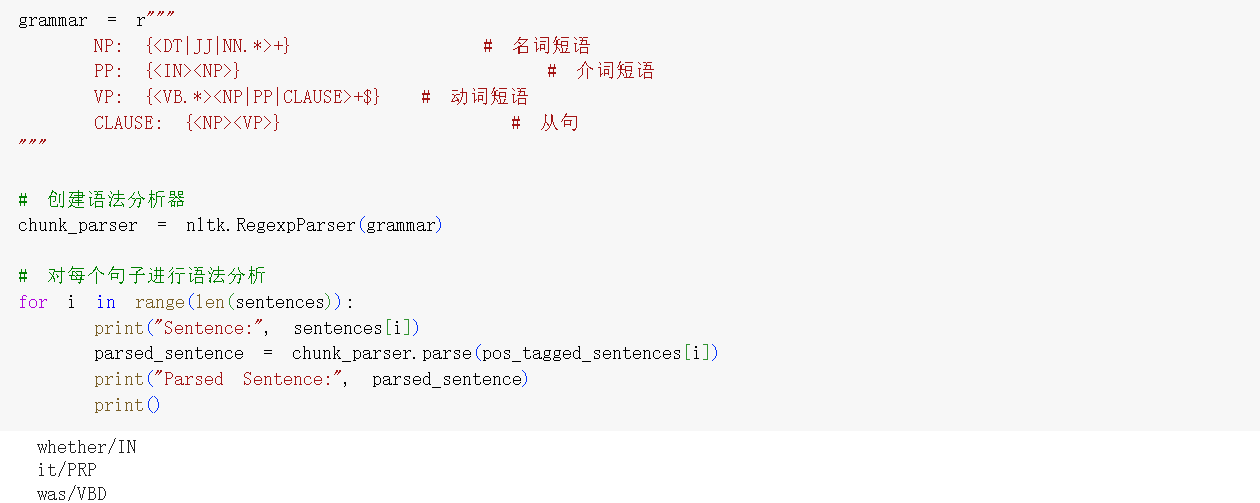
* Лемматизация.



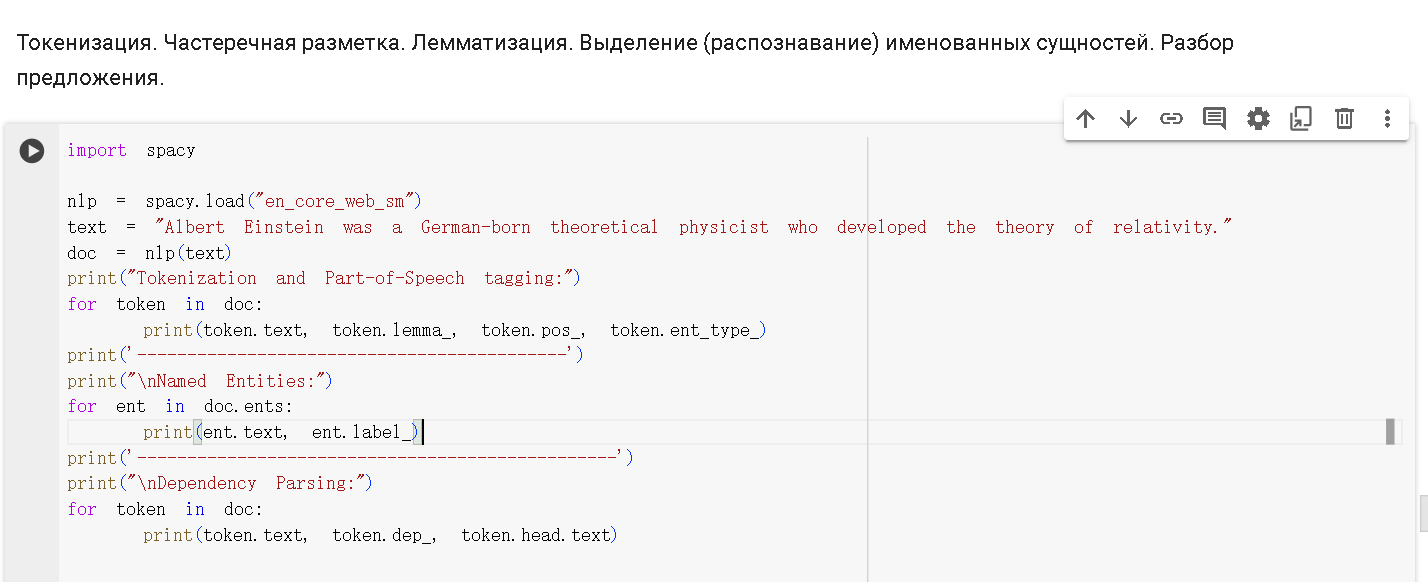
* Выделение (распознавание) именованных сущностей.

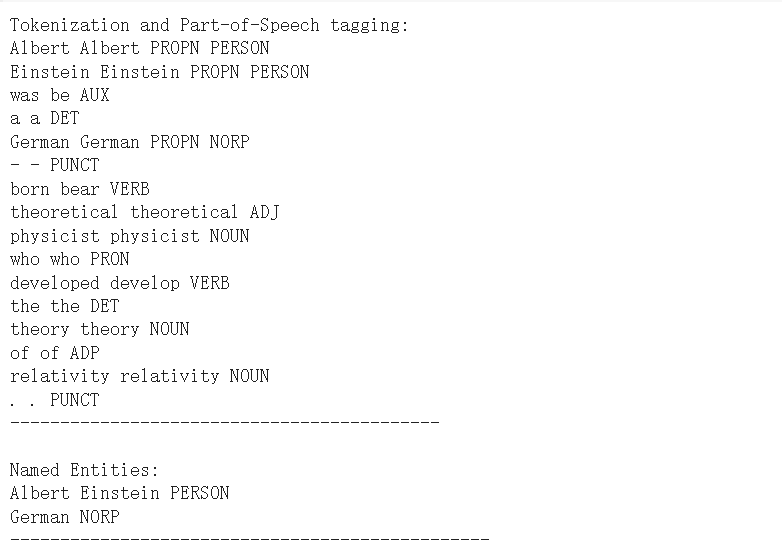


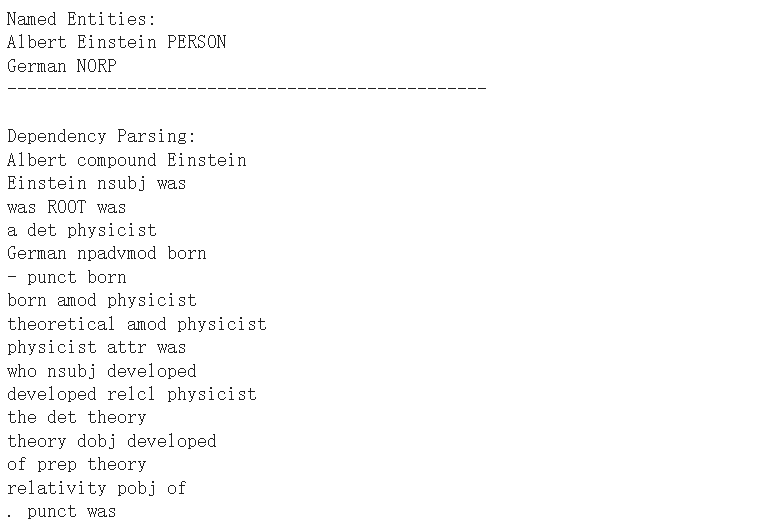
* Разбор предложения

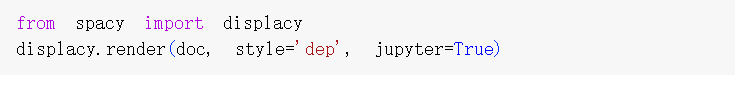


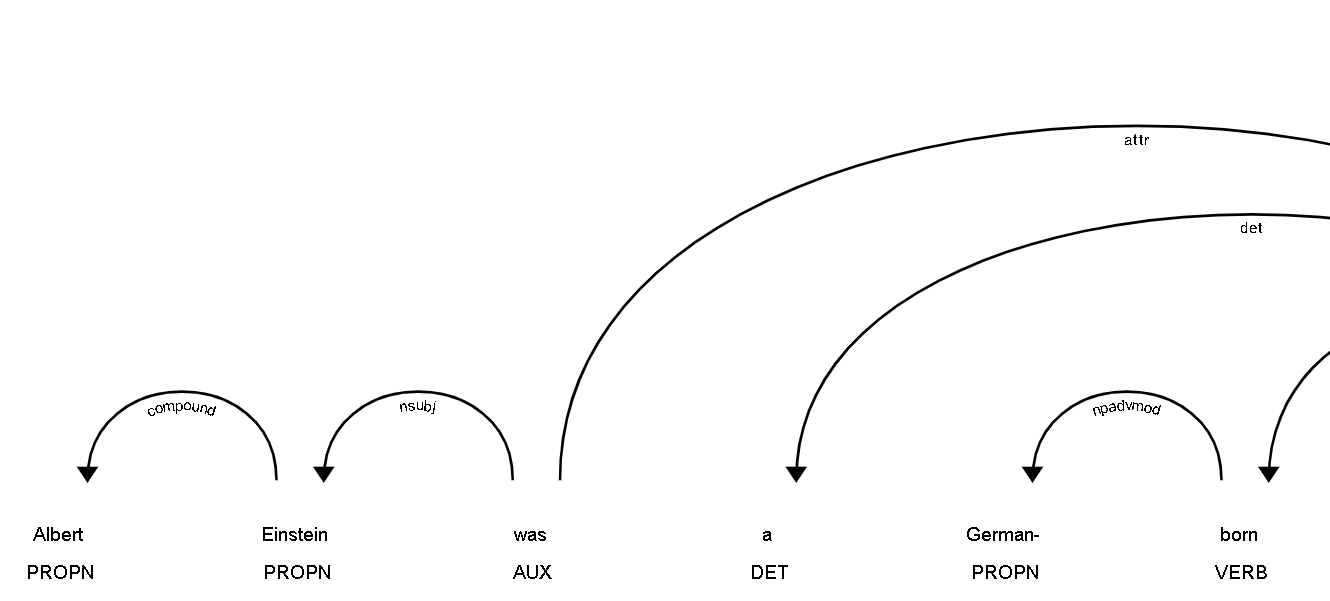
* **Используйте библиотеку spacy для реализации простого однострочного текста:**

****

****



****

****

# Вывод

Используйте инструменты обработки естественного языка, такие как NLTK (Natural Language Toolkit), SpaCy и т. д. для обработки и анализа текста. Эти инструменты предоставляют множество функций, включая сегментацию слов, разметку частей речи, лемматизацию, распознавание именованных объектов и синтаксический анализ.