Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика, искусственный интеллект и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 8**

**по дисциплине «Методы машинного обучения»**

Предобработка текста

ИСПОЛНИТЕЛЬ:

студентка ИУ5-23М

Морозевич М.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

\_\_\_ "\_\_\_\_\_\_\_\_" 2024 г.

Москва, 2024

# **Задание**

Для произвольного предложения или текста решите следующие задачи:

1. Токенизация;
2. Частеречная разметка;
3. Лемматизация;
4. Выделение (распознавание) именованных сущностей;
5. Разбор предложения.

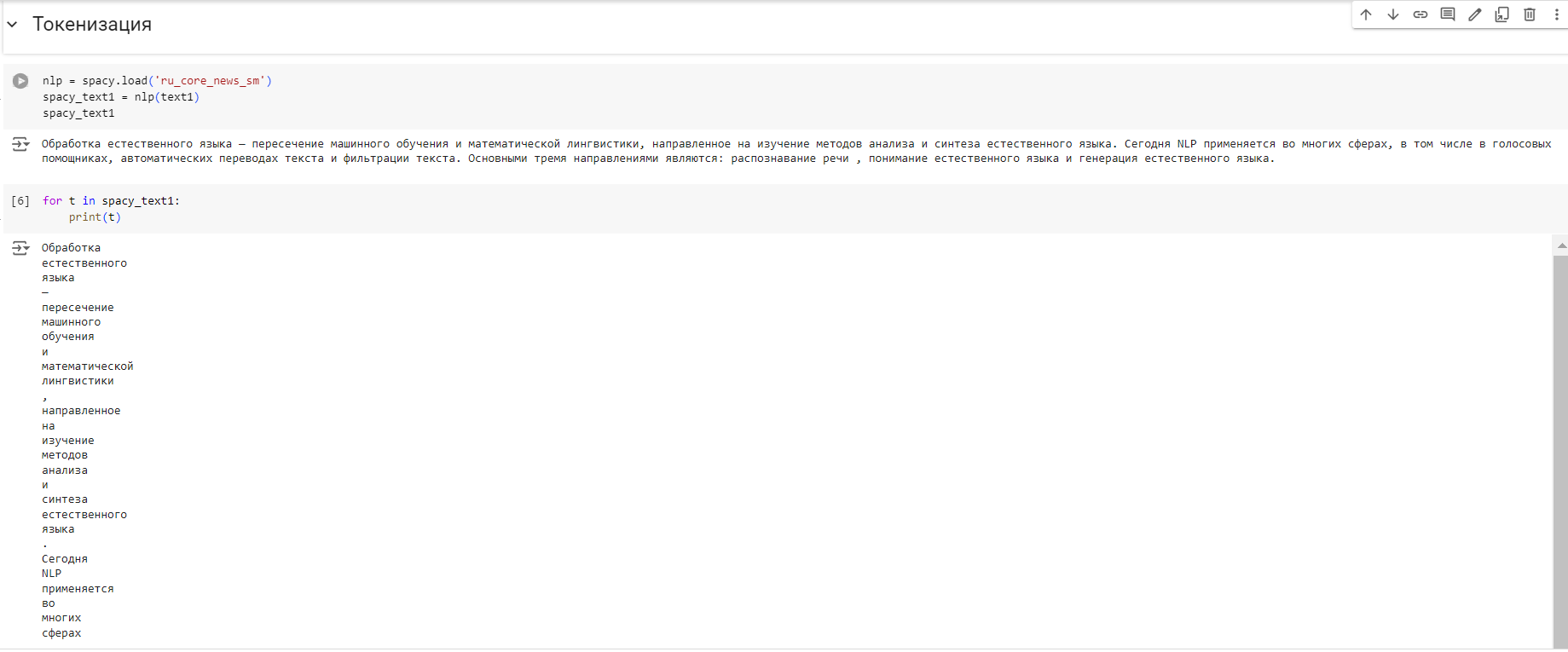
# **Описание данных и библиотеки Spacy**

Для разбора было выбрано предложение:

Обработка естественного языка — пересечение машинного обучения и математической лингвистики, направленное на изучение методов анализа и синтеза естественного языка. Сегодня NLP применяется во многих сферах, в том числе в голосовых помощниках, автоматических переводах текста и фильтрации текста. Основными тремя направлениями являются: распознавание речи , понимание естественного языка и генерация естественного языка.

# **Токенизация**

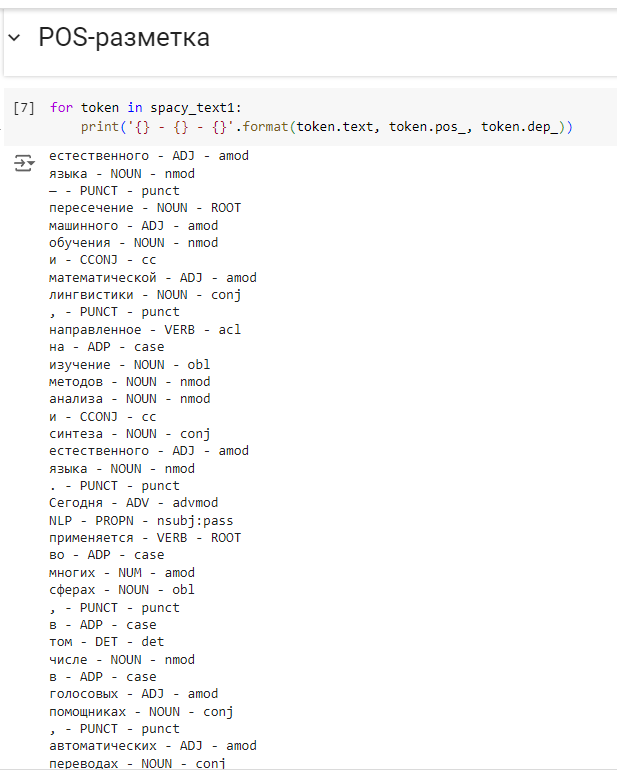
Первым шагом является разделение текста на отдельные слова или токены. Проведем токенизацию исходных данных.



Были получены токены текста, которые будем использовать на дальнейших шагах.

1. **Частеречная разметка**

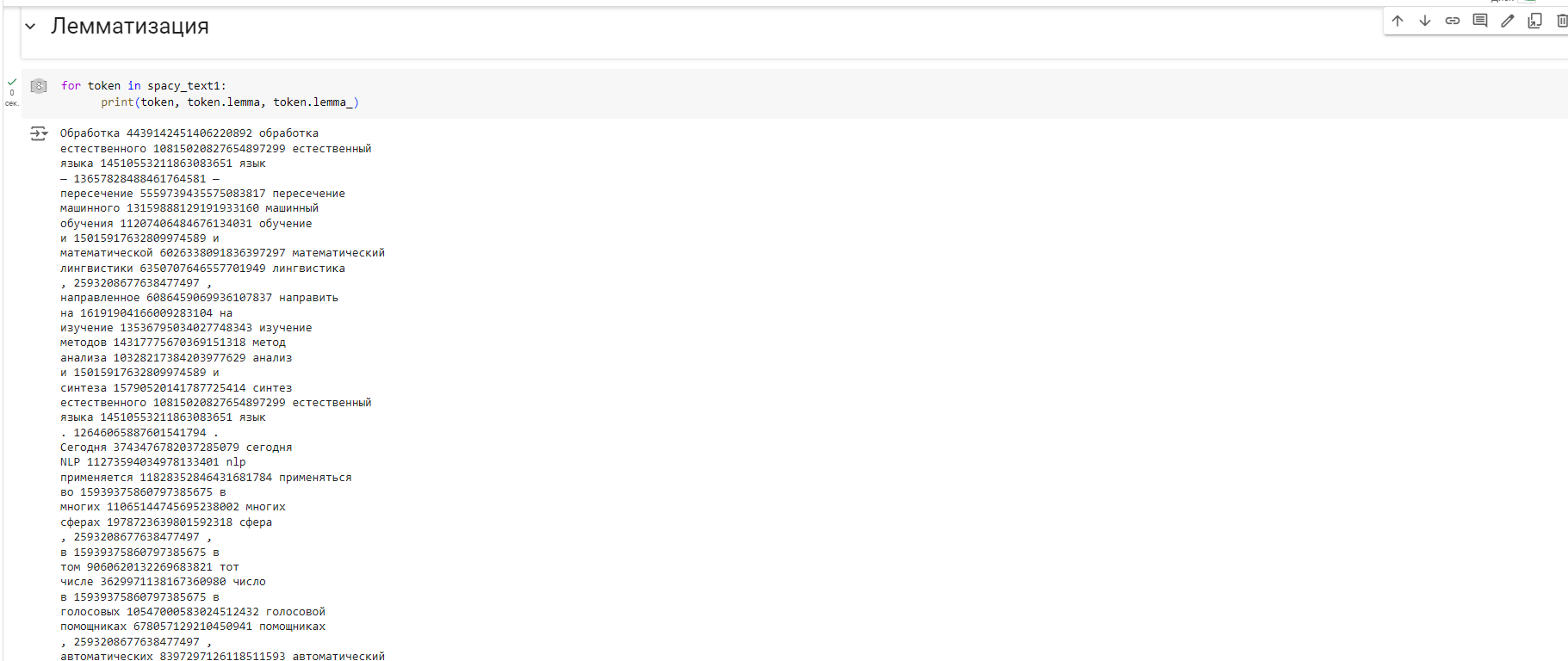
Дальше необходимо присвоить части речи каждому слову в тексте, это поможет понять грамматическую структуру предложения.



Были получены части речи или признак знака препинания для токенов текста.

# **Лемматизация**

Следующим шагом приведем слова к их нормальной (словарной) форме, что помогает уменьшить словарный запас и улучшать качество анализа текста.

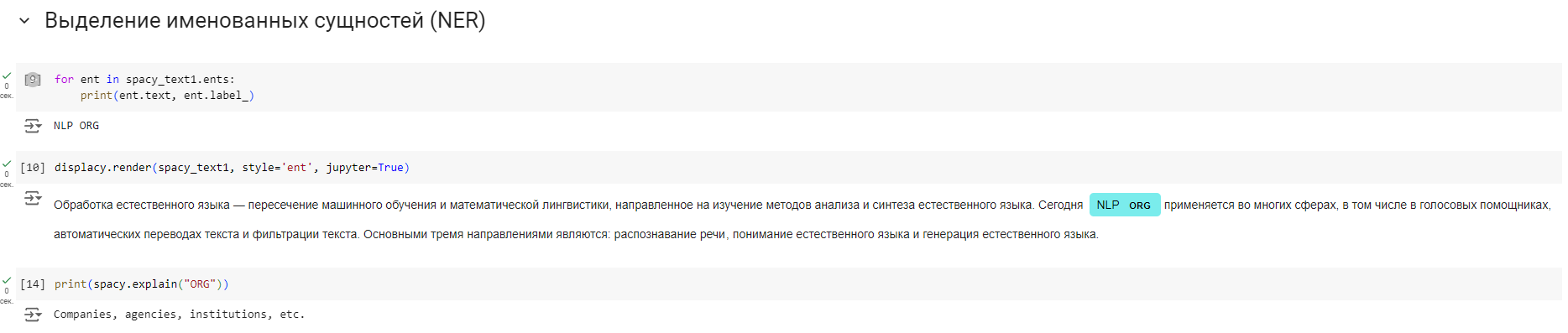


Слова текста были приведены к их леммам.

# **Выделение именованных сущностей**

Позволяет выделять из текста именованные сущности, такие как имена людей, названия организаций, географические места и т.д. Распознавание именованных сущностей полезно для извлечения информации и классификации текста.

Проверим существуют ли именованные сущности в исходном тексте.



В итоге, была выделена единственная сущность «NLP», которая была отнесена к ORG (Компании, агентства, учреждения и т.д.).

# **Разбор предложения**

Проанализируем грамматическую структуру предложения, включая выделение подлежащего, сказуемого, дополнений и других составляющих, для понимания связи между словами в предложении и его структуры.



Таким образом, были использованы методы предобработки текста. Были получены токены с их частями речи, их словарные формы и связи между ними в виде разбора предложения. Более того, в тексте была выявлена именованная сущность «NLP».