****Московский государственный технический университет им. Н.Э. Баумана  
Кафедра «Системы обработки информации и управления»

Лабораторная работа №8  
по дисциплине  
«Методы машинного обучения»  
на тему

# «Предобработка текста»

Выполнил:  
студент группы ИУ5И-22М  
Лу Жуньда

Москва — 2024 г.

1. **Цель лабораторной работы**

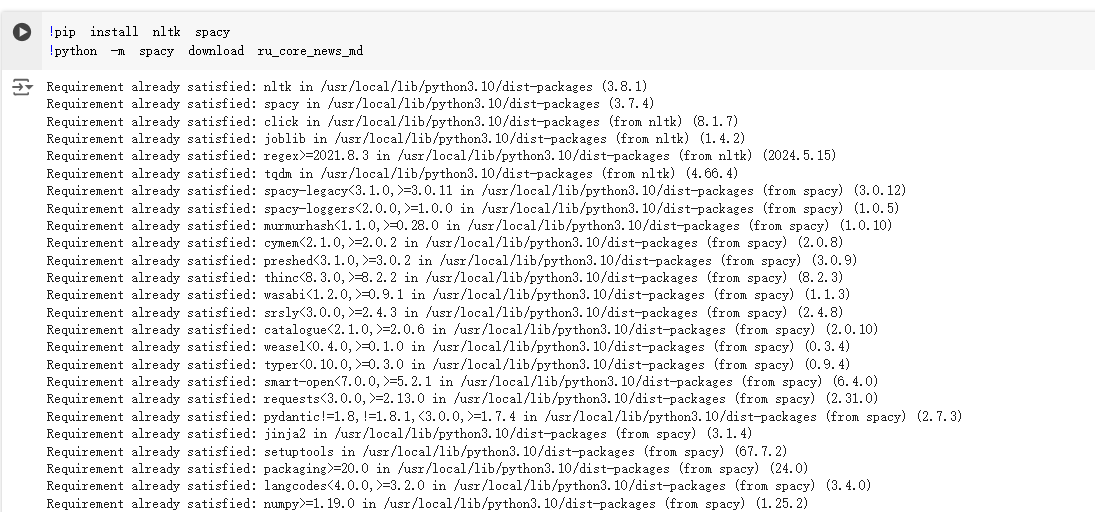
Изучение методов предобработки текстов.

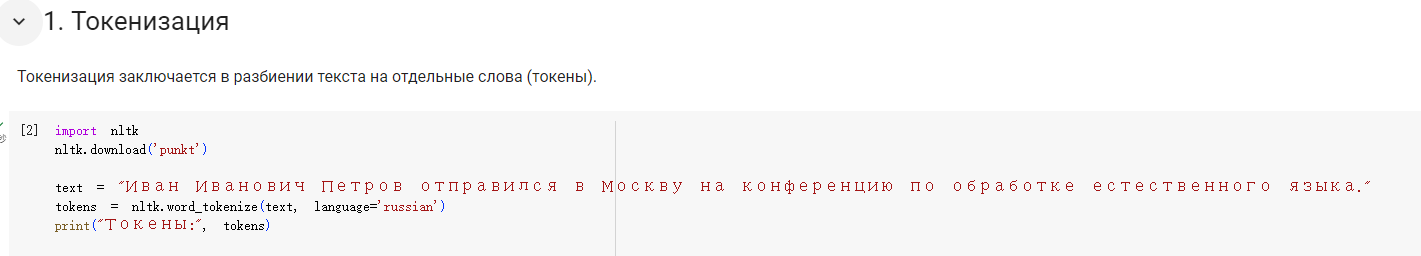
1. **Задание**

Для произвольного предложения или текста решите следующие задачи:

1. Токенизация.
2. Частеречная разметка.
3. Лемматизация.
4. Выделение (распознавание) именованных сущностей.
5. Разбор предложения.
6. **Текст программы**

Для выполнения задач предобработки текста на примере произвольного предложения, мы воспользуемся библиотекой nltk и spacy в Python.

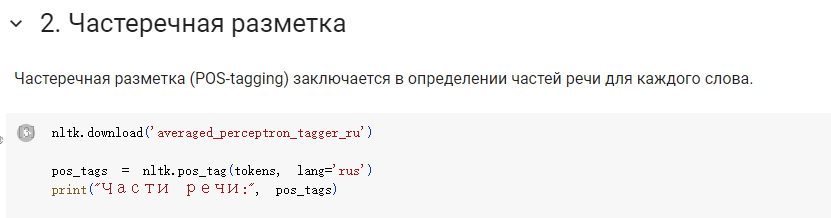




Токены: ['Иван', 'Иванович', 'Петров', 'отправился', 'в', 'Москву', 'на', 'конференцию', 'по', 'обработке', 'естественного', 'языка', '.']

[nltk\_data] Downloading package punkt to /root/nltk\_data...

[nltk\_data] Package punkt is already up-to-date!



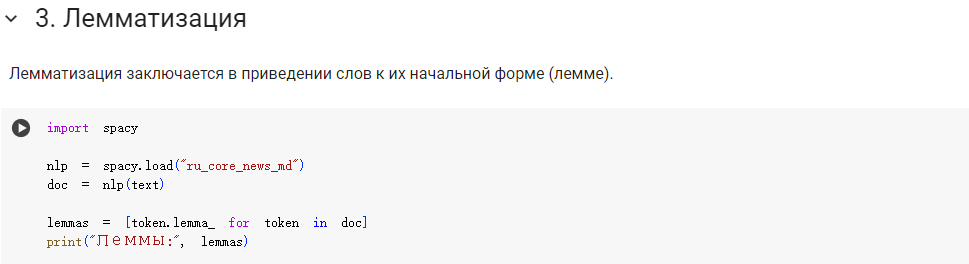
[nltk\_data] Downloading package averaged\_perceptron\_tagger\_ru to

[nltk\_data] /root/nltk\_data...

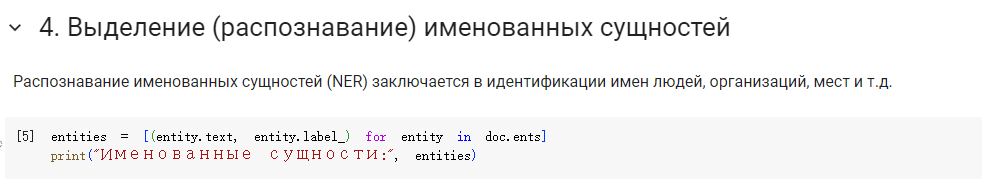
[nltk\_data] Package averaged\_perceptron\_tagger\_ru is already up-to-

[nltk\_data] date!

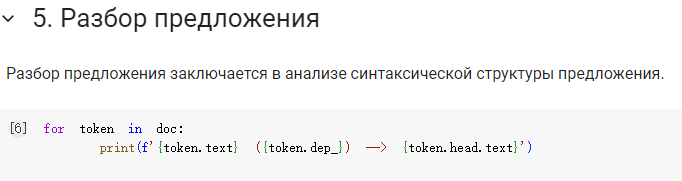
Части речи: [('Иван', 'S'), ('Иванович', 'S'), ('Петров', 'S'), ('отправился', 'V'), ('в', 'PR'), ('Москву', 'S'), ('на', 'PR'), ('конференцию', 'S'), ('по', 'PR'), ('обработке', 'S'), ('естественного', 'A=m'), ('языка', 'S'), ('.', 'NONLEX')]



Леммы: ['иван', 'иванович', 'петров', 'отправиться', 'в', 'москва', 'на', 'конференция', 'по', 'обработка', 'естественный', 'язык', '.']



Именованные сущности: [('Иван Иванович Петров', 'PER'), ('Москву', 'LOC')]



Иван (nsubj) --> отправился

Иванович (appos) --> Иван

Петров (appos) --> Иван

отправился (ROOT) --> отправился

в (case) --> Москву

Москву (obl) --> отправился

на (case) --> конференцию

конференцию (obl) --> отправился

по (case) --> обработке

обработке (nmod) --> конференцию

естественного (amod) --> языка

языка (nmod) --> обработке

. (punct) --> отправился