****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 ИУ5-23М  
Чжэн Сяохуэй

Москва — 2024 г.

**1. Цель лабораторной работы**

изучение методов предобработки текстов.

**2. Задание**

Для произвольного предложения или текста решите следующие задачи:

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

**3. текст программы**

import nltk

from nltk.tokenize import word\_tokenize

from nltk.tag import pos\_tag

from nltk.stem import WordNetLemmatizer

from nltk.chunk import ne\_chunk

from nltk.parse import CoreNLPParser

# 下载必要的资源

nltk.download('punkt')

nltk.download('averaged\_perceptron\_tagger')

nltk.download('wordnet')

nltk.download('maxent\_ne\_chunker')

nltk.download('words')

# 示例文本

text = "Apple is looking at buying U.K. startup for $1 billion."

# 1. 分词

tokens = word\_tokenize(text)

print("分词结果：", tokens)

# 2. 词性标注

tagged\_tokens = pos\_tag(tokens)

print("词性标注结果：", tagged\_tokens)

# 3. 词形还原

lemmatizer = WordNetLemmatizer()

lemmatized\_tokens = [lemmatizer.lemmatize(token) for token in tokens]

print("词形还原结果：", lemmatized\_tokens)

# 4. 命名实体识别

named\_entities = ne\_chunk(tagged\_tokens)

print("命名实体识别结果：")

print(named\_entities)

# 5. 句法解析

parser = CoreNLPParser(url='https://corenlp.run/')

parse\_tree = next(parser.raw\_parse(text))

print("句法解析结果：")

print(parse\_tree)

4. экранные формы с примерами выполнения программы.

[nltk\_data] Downloading package punkt to /home/ubuntu/nltk\_data...

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

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

[nltk\_data] /home/ubuntu/nltk\_data...

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

[nltk\_data] date!

[nltk\_data] Downloading package wordnet to /home/ubuntu/nltk\_data...

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

[nltk\_data] Downloading package maxent\_ne\_chunker to

[nltk\_data] /home/ubuntu/nltk\_data...

[nltk\_data] Package maxent\_ne\_chunker is already up-to-date!

[nltk\_data] Downloading package words to /home/ubuntu/nltk\_data...

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

分词结果： ['Apple', 'is', 'looking', 'at', 'buying', 'U.K.', 'startup', 'for', '$', '1', 'billion', '.']

词性标注结果： [('Apple', 'NNP'), ('is', 'VBZ'), ('looking', 'VBG'), ('at', 'IN'), ('buying', 'VBG'), ('U.K.', 'NNP'), ('startup', 'NN'), ('for', 'IN'), ('$', '$'), ('1', 'CD'), ('billion', 'CD'), ('.', '.')]

词形还原结果： ['Apple', 'is', 'looking', 'at', 'buying', 'U.K.', 'startup', 'for', '$', '1', 'billion', '.']

命名实体识别结果：

(S

(GPE Apple/NNP)

is/VBZ

looking/VBG

at/IN

buying/VBG

U.K./NNP

startup/NN

for/IN

$/$

1/CD

billion/CD

./.)

句法解析结果：

(ROOT

(S

(NP (NNP Apple))

(VP

(VBZ is)

(VP

(VBG looking)

(PP

(IN at)

(S

(VP

(VBG buying)

(NP (NNP U.K.) (NN startup))

(PP (IN for) (NP (QP ($ $) (CD 1) (CD billion)))))))))

(. .)))

**Список литературы**

[1] Гапанюк Ю. Е. COURSE\_MMO\_SPRING\_2024// GitHub. –– 2024. –– Режим доступа: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_MMO\_SPRING\_2024

[2] <https://www.kaggle.com/datasets>