Московский государственный технический университет

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 8**

**По курсу «методы машинного обучения в АСОИУ»**

**«Предобработка текста»**

**Выполнил:**

студент ИУ5-24М

Ширшов А.С.

**Проверил:**

Гапанюк Ю.Е.

Подпись:

29.02.2024

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.

**Ход работы**

Возьмём библиотеку Spacy и напишем случайное предложение. Код на рисунке ниже

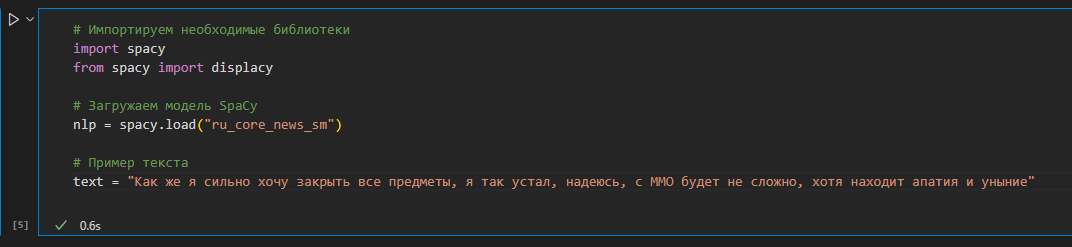


Рисунок 1 - Начало работы с предложением

Токенизация представлена ниже:

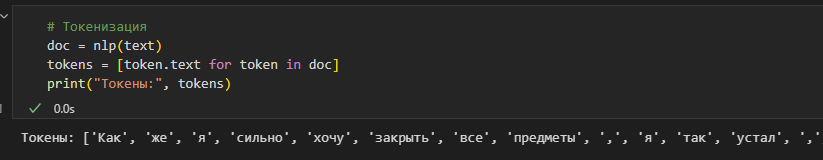


Рисунок 2 - Токенизация предложения

Частеречная разметка представлена ниже:

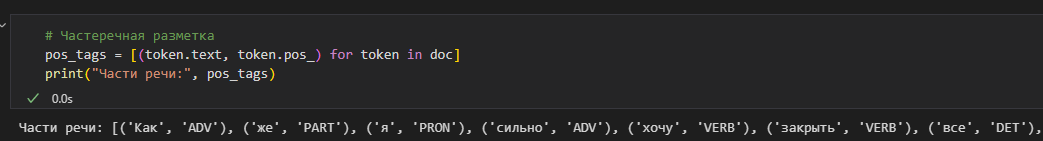


Рисунок 3 - Частеречная разметка

Лемматизация текста представлена ниже:

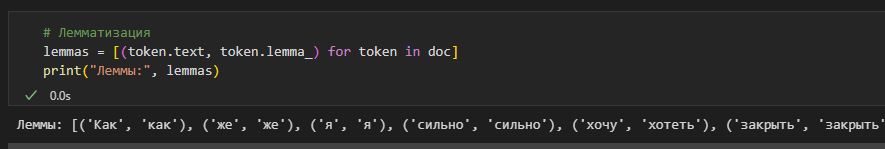


Рисунок 4 - Лемматизация текста

Выделение именованных сущностей представлено ниже:

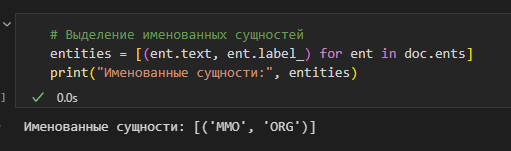


Рисунок 5 - Выделение именованных сущностей

Разбор предложения представлен на рисунке ниже:

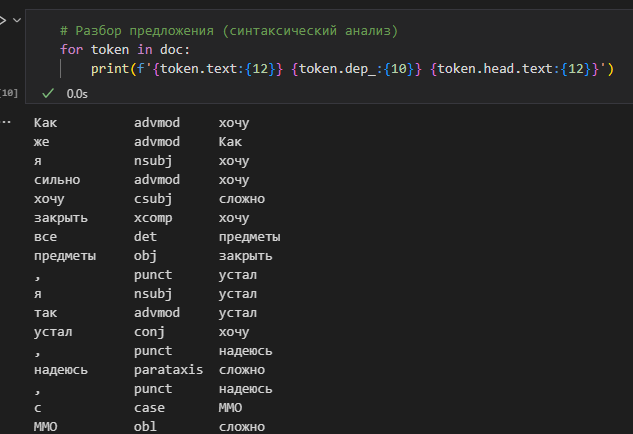


Рисунок 6 - Разбор предложения

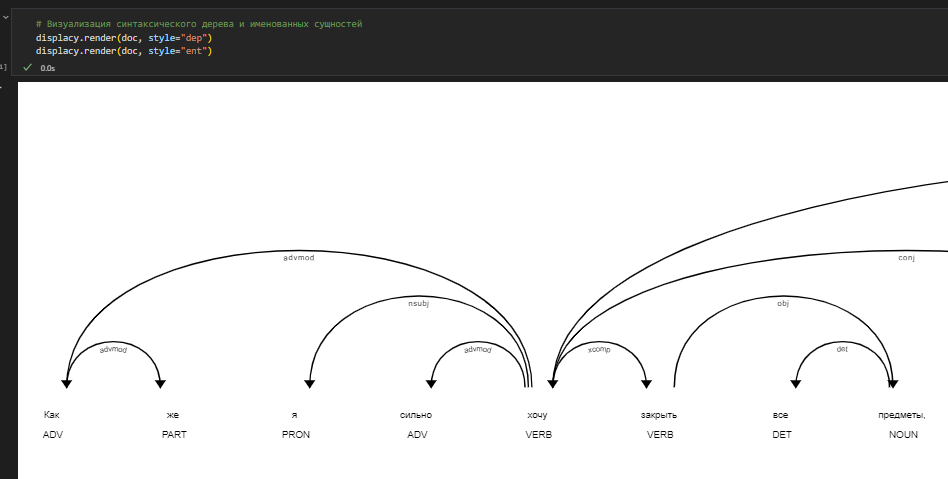


Рисунок 7 - Визуализация синтаксического дерева

**Вывод**

В ходе выполнения лабораторной работы были решены задачи по обработке произвольного предложения или текста.

1. Токенизация: Было проведено разделение текста на отдельные лексические единицы, или токены. Этот процесс является необходимым шагом для последующего анализа текста.
2. Частеречная разметка: Для каждого токена был определен его грамматический класс, или часть речи. Это позволяет лучше понять структуру предложения и роль каждого слова в нем.
3. Лемматизация: Была проведена лемматизация, или восстановление основной формы слова. Этот процесс позволяет сгруппировать разные формы одного и того же слова и улучшить результаты последующего анализа.
4. Выделение именованных сущностей: Были выделены именованные сущности, или объекты, которые можно однозначно идентифицировать, такие как имена людей, названия организаций, географические объекты и т.д. Этот процесс является важным шагом для многих задач обработки текстов.
5. Разбор предложения: Был проведен синтаксический разбор предложения, или определение грамматических связей между словами. Этот процесс позволяет лучше понять смысл предложения и роль каждого слова в нем.