Московский государственный технический университет им. Н.Э. Баумана

Факультет «Головной учебно-исследовательский и методический центр

профессиональной реабилитации лиц с ограниченными возможностями здоровья (инвалидов)»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа 8**

**по дисциплине «Методы машинного обучения в АСОИУ»**

**"** **Предобработка текстов "**

СТУДЕНТ:

студент группы ИУ5Ц-21М

Москалик А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

Москва, 2024

# 

# Цель лабораторной работы: изучение методов предобработки текстов. Задание:

# Для произвольного предложения или текста решите следующие задачи:

# Токенизация.

# Частеречная разметка.

# Лемматизация.

# Выделение (распознавание) именованных сущностей.

# Разбор предложения.

# Ход работы

# Импорт библиотек

# Этот код выполняет подготовку к анализу текста, используя две популярные библиотеки для обработки естественного языка: nltk и spaCy.

# - nltk.download('punkt'): Загружает ресурсы для токенизатора punkt. Этот токенизатор предварительно обучен работать с многими европейскими языками и используется для разбиения текста на предложения и слова.

# - nltk.download('averaged\_perceptron\_tagger'): Загружает ресурсы для частеречного теггера, который использует алгоритм перцептрона. Это обеспечивает функциональность маркировки частей речи (POS tagging), позволяя определить, является ли слово существительным, глаголом и т. д.

# 

# Текст для анализа

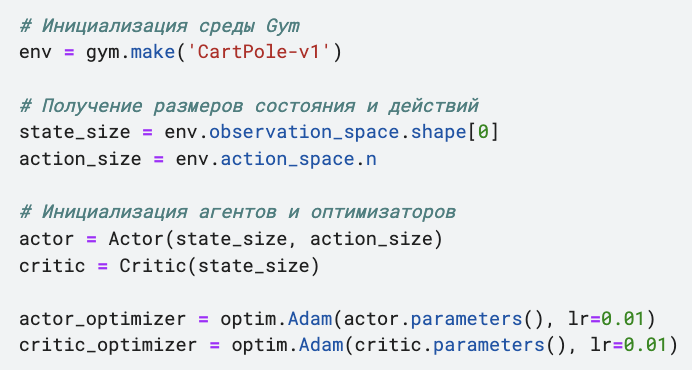
# Текст содержит информацию о компании Apple Inc.:

# 

# "Apple Inc." указывает на название компании.

# "is an American multinational technology company" описывает компанию как американскую мультинациональную технологическую компанию.

# "headquartered in Cupertino, California" указывает на местоположение главного офиса компании в городе Купертино, штат Калифорния.



1. **Токенизация**

Токенизация — это процесс разделения текста на более мелкие части, называемые токенами. Токены обычно представляют собой слова, числа или знаки пунктуации. Токенизация является одним из первых шагов в обработке и анализе текста, поскольку она помогает подготовить данные для дальнейшей обработки,такой как частеречная разметка, лемматизация и синтаксический разбор.



- word\_tokenize(text) — функция из библиотеки NLTK, используемая для токенизации строки text. Она разбивает текст на слова, учитывая знаки пунктуации как отдельные токены.

В переменную tokens сохраняется список токенов, полученных в результате токенизации текста.

- print("Tokens:", tokens) — выводит на экран список токенов. В данном случае вывод будет следующим:



1. **Частеречная разметка**

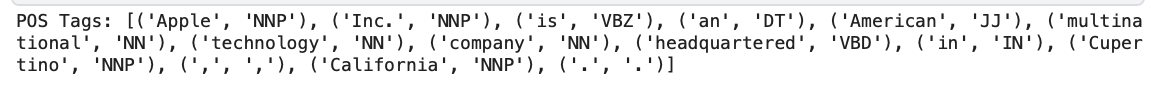
Частеречная разметка (POS tagging, part-of-speech tagging) — это процесс присваивания каждому слову в тексте соответствующей части речи, такой как существительное, глагол, прилагательное и так далее. Этот процесс является важным шагом в анализе текста, поскольку понимание роли каждого слова в предложении помогает в дальнейшем синтаксическом и семантическом анализе.



- nltk.pos\_tag(tokens) — функция, которая принимает список токенов и возвращает список кортежей, где каждый кортеж содержит токен и соответствующий ему тег части речи. Для определения частей речи используется предварительно обученная модель.

В переменной pos\_tags сохраняется список кортежей с токенами и их частеречными тегами.

- print("POS Tags:", pos\_tags) — выводит на экран список частеречных тегов. В данном случае вывод будет следующим:

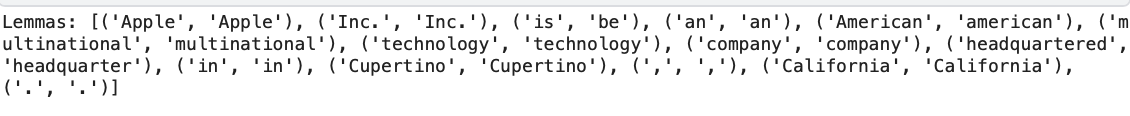


* NNP обозначает собственное имя (proper noun).
* VBZ обозначает глагол в настоящем времени, 3-е лицо единственного числа.
* DT обозначает артикль или указательное местоимение.
* JJ обозначает прилагательное.
* NN обозначает общее существительное (common noun) в единственном числе.
* VBD обозначает глагол в прошедшем времени.
* IN обозначает предлог или подчинительный союз.

1. **Лемматизация**

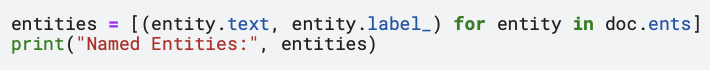
Лемматизация — это процесс приведения слова к его базовой форме, или лемме, который обычно включает сведение различных форм слова (например, множественное число, разные времена глаголов) к единой, канонической форме. Этот процесс помогает уменьшить морфологическую сложность текста и улучшить производительность многих задач обработки естественного языка, таких как семантический анализ, машинный перевод и поиск информации.





1. **Распознавание именованных сущностей**

Распознавание именованных сущностей (Named Entity Recognition, NER) — это процесс идентификации и классификации ключевых информационных элементов в тексте в категории, такие как имена людей, организаций, местоположений, дат, продуктов и других специфических групп. NER помогает в структурировании неструктурированного текста и является важной частью многих приложений обработки естественного языка, таких как извлечение информации, автоматическое резюмирование, поиск по именам и машинный перевод.



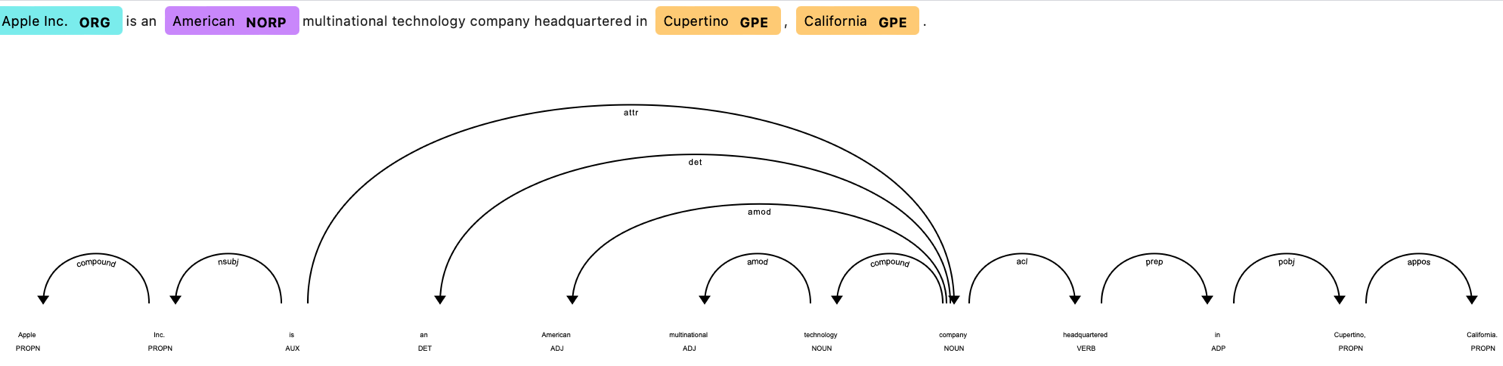


* 'Apple Inc.' классифицирована как 'ORG' (организация).
* 'American' классифицировано как 'NORP' (национальная или религиозная группа).
* 'Cupertino' и 'California' классифицированы как 'GPE' (геополитическая единица, такая как страна, город, штат).

1. **Визуализация анализа**

Визуализация анализа текста с использованием инструментов библиотеки spaCy, таких как displacy, позволяет наглядно представить результаты обработки естественного языка.





**Вывод:**

На изображении показаны:

1. Распознавание именованных сущностей (NER):

- Apple Inc. помечена как организация (`ORG`).

- American помечено как национальная принадлежность (`NORP`).

- Cupertino, California помечено как геополитическая единица (`GPE`).

2. Разбор предложения (Dependency Parsing):

- Структура зависимостей между словами в предложении визуализирована с помощью стрелок, указывающих отношения между словами (например, `nsubj`, `attr`, `amod`).

Этот вывод демонстрирует, как spaCy анализирует текст, выделяя ключевые структурные элементы и семантические свойства. Такой визуальный анализ очень полезен для понимания взаимосвязей в тексте и для проверки правильности обработки текста алгоритмами.