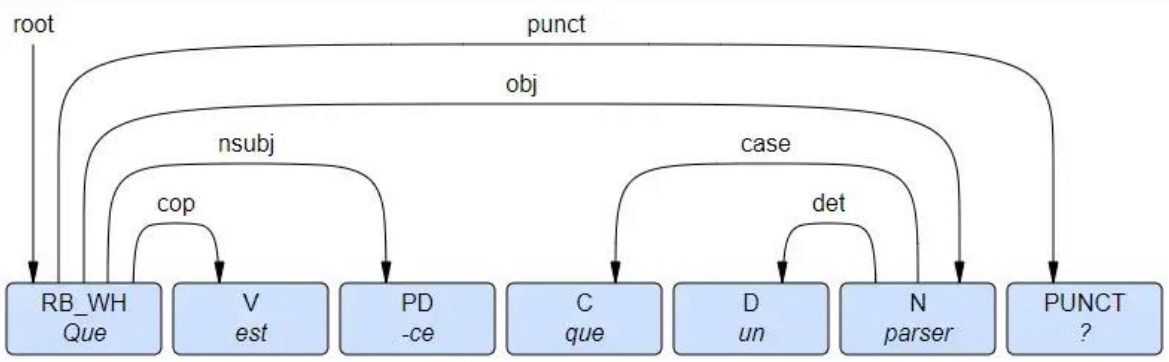
Teach Me Skills

Занятие 43. Введение в NLP

Talk about…

- История и развитие NLP
- Классификация текстов: объяснение задачи и примеры.
- Извлечение информации: объяснение задачи и примеры.
- Генерация текста: объяснение задачи и примеры.
- NLTK, spaCy, Gensim.



1. **Введение в NLP и история**

Teach Me Skills

1.1950-1960 годы: Зарождение NLP
2.В это время начались первые попытки создания компьютерных программ, способных анализировать и генерировать текст на естественном языке.
3.Проекты, такие как "Georgetown-IBM experiment" и "ELIZA," были важными ранними шагами в этом направлении.
4.1970-1980 годы: Грамматические подходы
5.Этот период был ознаменован развитием грамматических формализмов и компьютерных грамматик для анализа и синтеза текста.
6.Проекты, такие как "Shrdlu" и "GUS" использовали формальные грамматические правила для обработки текста.

1.1990-2000 годы: Статистические методы и машинное обучение
2.В это время стали активно применяться статистические методы и машинное обучение для решения задач NLP.
3.Словоизвлечение и морфологический анализ начали основываться на статистических моделях.
Появление корпусов текстов и методов машинного обучения, таких как скрытые марковские модели (HMM) и условные случайные поля (CRF), улучшило результаты.
2000-2010 годы: Подходы, основанные на данных
1.Появление методов, таких как WordNet, и стандартизация оценки качества систем NLP через соревнования, такие как SemEval и CoNLL, стали важными.
2.Применение глубокого обучения к NLP, такое как рекуррентные нейронные сети (RNN) и сверточные нейронные сети (CNN), стало популярным.

С 2010 года: Взрывной рост глубокого обучения и BERT

1.Появление глубоких архитектур, таких как рекуррентные нейронные сети с долгой краткосрочной памятью (LSTM) и BERT (Bidirectional Encoder Representations from Transformers), привело к значительному улучшению результатов в NLP.

BERT, представленный в 2018 году, стал одним из ключевых моментов в развитии NLP, и его предобученные модели применяются для ряда задач NLP.

1.2020 и далее: Эволюция и расширение применений

2.NLP продолжает развиваться, и исследователи работают над расширением применения методов обработки текста в различных областях, включая медицину, финансы, автоматический перевод и многие другие.


SPAM
SPAM
SPAM FILTER

2. Основные задачи NLP

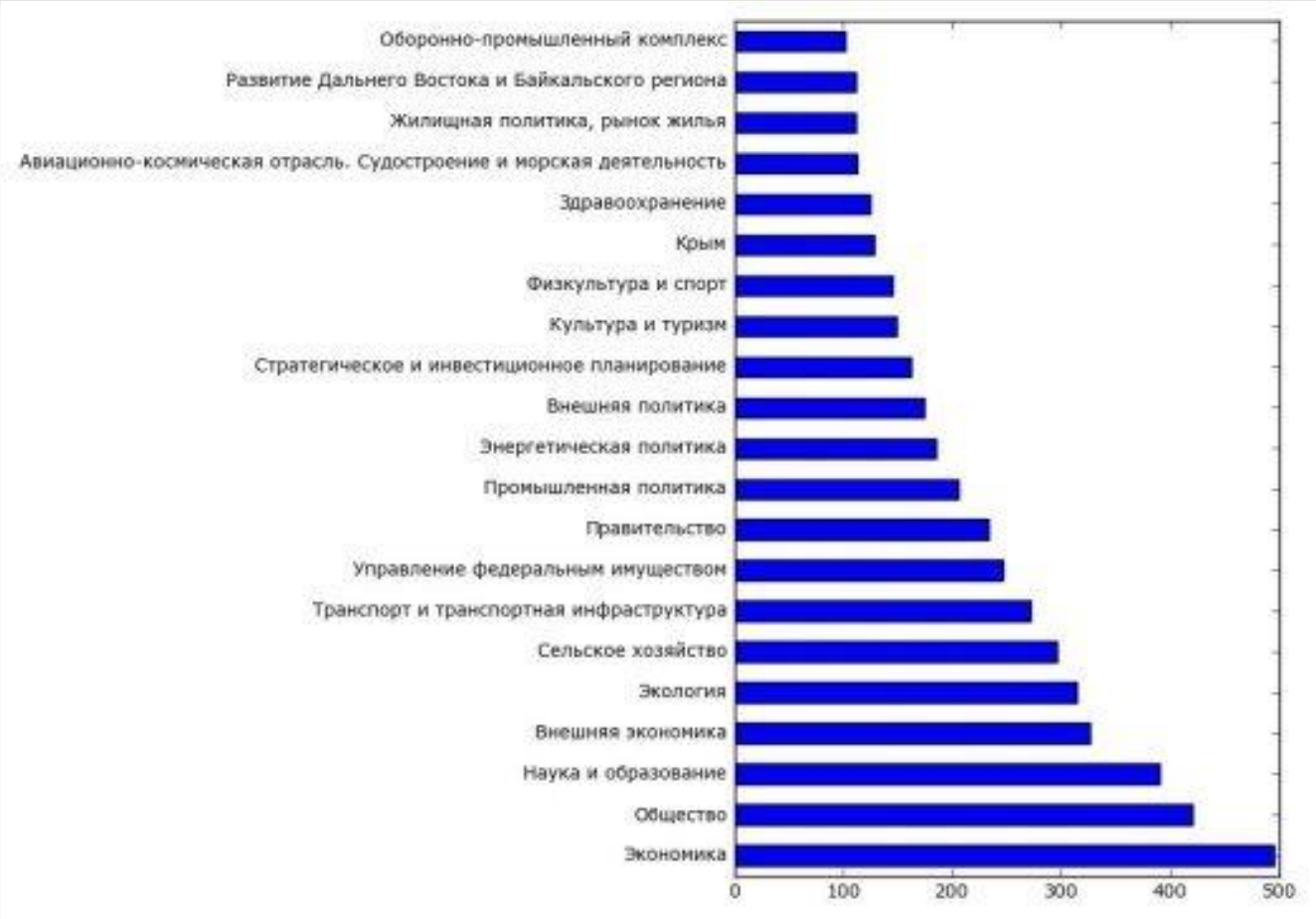
Teach Me Skills

Спам-фильтрация: Одной из самых известных задач классификации текстов является фильтрация спама в электронной почте. Система классифицирует входящие сообщения как "спам" или "не спам" на основе их содержания и характеристик.

Анализ тональности: Задача анализа тональности заключается в определении эмоциональной окраски текста. Например, система может классифицировать отзывы о продуктах или фильмах как "положительные," "отрицательные" или "нейтральные".

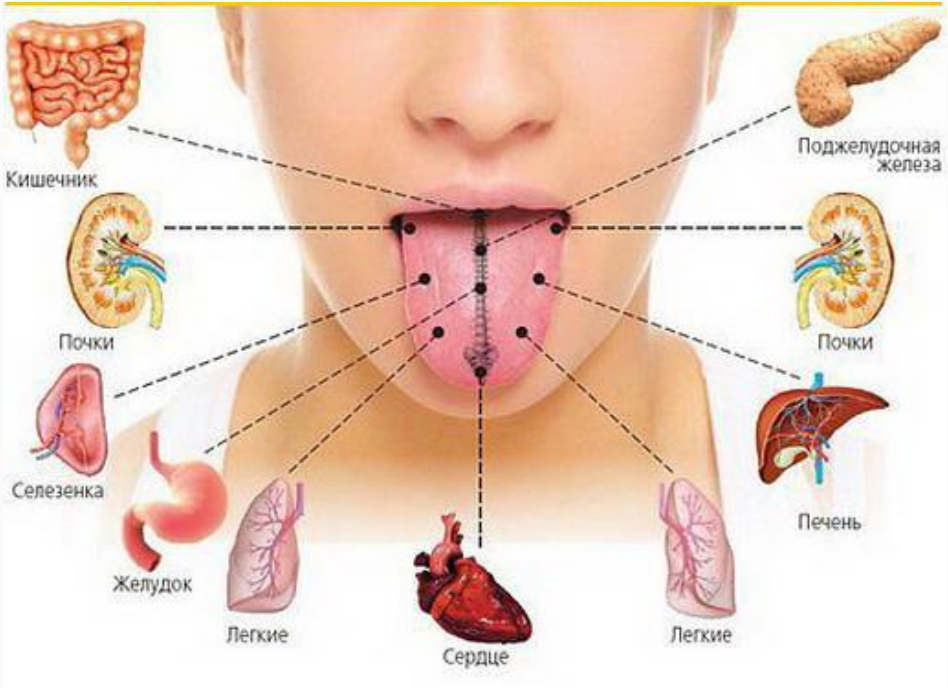
Классификация категорий новостей: Системы классификации могут быть использованы для автоматической категоризации новостных статей в различные рубрики, такие как "политика," "спорт," "наука" и т. д.

2. Основные задачи NLP



Выявление тем в текстах: Классификация текстов может также использоваться для определения темы или категории текста. Например, определение, к какой теме относится статья в блоге или на новостном сайте.

Определение языка: Задача определения языка текста также является видом классификации, где система определяет, на каком языке написан данный текст.



Бинарная классификация: В данном случае тексты классифицируются на два класса, например, "положительный" и "отрицательный" отзыв.

Многоклассовая классификация: В этой задаче тексты могут быть отнесены к нескольким классам. Например, новостная статья может быть отнесена к категориям "политика," "спорт" и "наука" одновременно.

Многомерная классификация: Здесь каждый текст может быть отнесен к нескольким категориям, но каждая категория может иметь разную степень принадлежности. Это используется, например, в задаче классификации тематики статей.

Извлечение именованных сущностей (NER): В этой задаче система выявляет и извлекает именованные сущности, такие как имена людей, местоположения, организации и даты, из текстов. Например, в предложении "Барак Обама родился 4 августа 1961 года в Гонолулу" система должна извлечь имена "Барак Обама," "Гонолулу" и дату "4 августа 1961 года."

Извлечение фактов и событий: Задача заключается в извлечении структурированных данных о событиях или фактах из текстов. Например, из текста новостной статьи о спортивном событии система может извлечь информацию о дате, местоположении и участниках события.

2. Основные задачи NLP

Teach Me Skills

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Извлечение именованных сущностей (NER)

Извлечение отношений между сущностями: Задача состоит в определении отношений или связей между именованными сущностями в тексте. Например, из текста "Стив Джобс был основателем Apple Inc." система должна извлечь отношение "основатель" между "Стивом Джобсом" и "Apple Inc."

Извлечение информации из табличных данных: IE может применяться для извлечения данных из табличных форматов, таких как расписания, цены товаров или данные о компаниях, чтобы обогатить базы данных или автоматически обновлять информацию.

Системы на основе правил: Извлечение информации может быть реализовано с использованием набора правил и шаблонов, которые описывают структуру интересующей информации. Это позволяет выявлять факты и сущности на основе явных паттернов в тексте.

Машинное обучение: Методы машинного обучения, такие как маркированные корпуса и алгоритмы классификации, могут быть применены для автоматического обучения систем извлечения информации из больших объемов текста.

Глубокое обучение: Нейронные сети, включая рекуррентные нейронные сети (RNN) и трансформеры, такие как BERT, могут использоваться для задачи извлечения информации. Эти методы обычно достигают хороших результатов на задачах NER и извлечения отношений.

Генерация автоматических описаний изображений (Image Captioning): Задача заключается в создании текстовых описаний изображений. Например, для фотографии льва система может сгенерировать описание "Лев, стоящий на саванне."

Генерация текстовых рецензий и отзывов: Системы могут генерировать текстовые отзывы о продуктах, фильмах или ресторанах, основываясь на анализе текстовых данных и обратной связи от пользователей.

Автоматическое создание контента для сайтов и блогов: Генерация статей, новостных заголовков или блог-постов на определенные темы или ключевые слова.

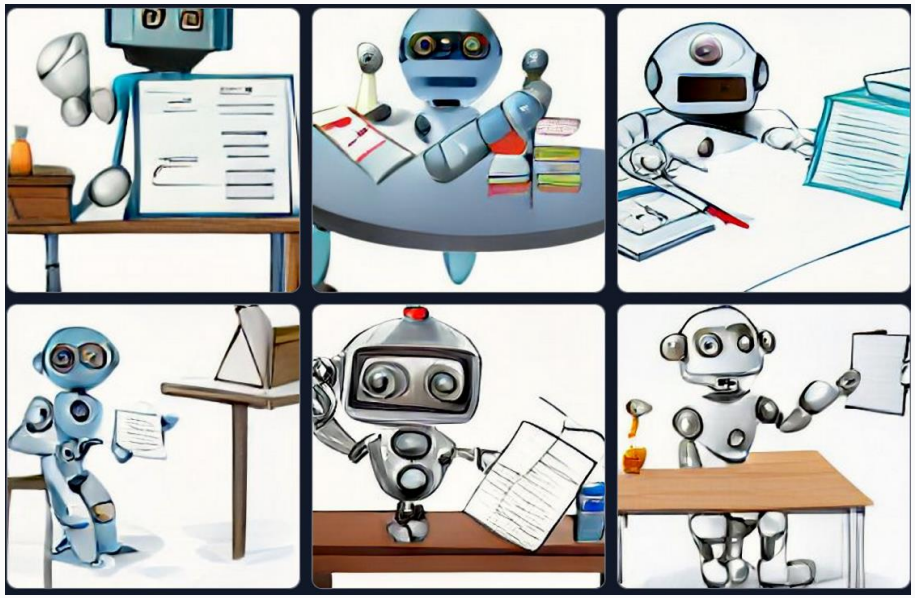
2. Основные задачи NLP

Teach Me Skills

Генерация автоматических описаний изображений (Image Captioning

Генерация диалоговых систем (чат-ботов): Создание ответов и диалоговых сообщений, которые могут использоваться в чат-ботах или виртуальных ассистентах.

Генерация музыки и поэзии: Задачей может быть создание текстов песен или стихов.



Шаблонные методы: Это простые методы, которые могут создавать текст с использованием заранее определенных шаблонов или правил. Например, система может заменять заполнители в шаблоне, чтобы создать уникальные тексты.

Модели на основе правил: Задачу генерации текста можно решать с использованием систем, которые содержат множество правил и логику для генерации текста. Например, система автоматической генерации текстовых отчетов может использовать правила для создания текста на основе данных.

Методы машинного обучения: Многие задачи генерации текста решаются с использованием методов машинного обучения, таких как генеративные модели, рекуррентные нейронные сети (RNN) и сверточные нейронные сети (CNN).

Глубокое обучение: Глубокие нейронные сети, такие как рекуррентные сети с долгой краткосрочной памятью (LSTM) и трансформеры, стали популярными для задач генерации текста. Например, модель GPT-3 может генерировать текст на основе контекста и заданных ключевых слов.

Перевод текстов: Задача перевода текстов с одного языка на другой язык. Например, перевод с английского на французский, с испанского на немецкий и так далее.

Автоматический перевод в реальном времени: Машинный перевод может быть использован для автоматического перевода разговоров или текстов в реальном времени, например, в чатах или телефонных звонках.

Перевод веб-страниц и документов: Системы машинного перевода могут переводить веб-страницы, документы и мультимедийный контент, чтобы сделать их доступными для аудитории на разных языках.

Перевод субтитров и озвучки: Перевод субтитров и озвучки фильмов и видеороликов на другие языки.

Статистический машинный перевод (SMT): Этот метод основан на статистическом анализе параллельных текстовых корпусов на двух языках. Модель SMT определяет, какие слова или фразы соответствуют друг другу на разных языках на основе вероятности и частоты встречаемости.

Модели на основе нейронных сетей: Современные методы машинного перевода, такие как Seq2Seq (Sequence-to-Sequence) и трансформеры, используют глубокое обучение и нейронные сети для более точного и контекстно-зависимого перевода. Например, модель "Transformer" используется в таких системах, как Google Translate.

Системы с гибридными подходами: Некоторые системы машинного перевода используют комбинацию методов SMT и нейронных сетей, чтобы совместить преимущества обоих подходов и улучшить качество перевода.

Анализ тональности (Sentiment Analysis), также известный как определение тональности, - это задача в области обработки естественного языка (NLP), которая заключается в определении эмоциональной окраски текста или выраженной в нем субъективной оценки. В данной задаче система анализирует текстовые данные и определяет, является ли эмоциональное состояние текста положительным, отрицательным или нейтральным.

Оценка отзывов и комментариев: В веб-магазинах, на сайтах ресторанов и в социальных сетях системы анализа тональности могут автоматически определять, являются ли отзывы о продуктах или услугах положительными или отрицательными.

Мониторинг общественного мнения: Системы анализа тональности могут использоваться для отслеживания общественного мнения и реакции на события, продукты или бренды в социальных медиа и новостных источниках.

Анализ новостных статей: Системы могут определять, какой тональности являются новостные статьи, чтобы классифицировать их как позитивные, негативные или нейтральные.

Мониторинг чувств публики: В рамках мероприятий и медиа-кампаний анализ тональности может использоваться для измерения реакции аудитории и определения эффективности кампаний.

Модели на основе правил: Системы могут использовать набор правил и лексические ресурсы, чтобы определить эмоциональную окраску слов и фраз. Например, негативные слова или выражения могут указывать на негативную тональность.

Машинное обучение: Методы машинного обучения, такие как классификация, используются для обучения моделей анализа тональности на размеченных данных. Модели могут классифицировать тексты как "положительные," "отрицательные" или "нейтральные" на основе признаков, извлеченных из текста.

Глубокое обучение: Нейронные сети, включая сверточные нейронные сети (CNN) и рекуррентные нейронные сети (RNN), могут использоваться для анализа тональности. Например, рекуррентные сети могут учитывать контекст и последовательность слов в тексте для более точной оценки тональности.

Поддержка клиентов и обслуживание запросов: Чат-боты могут отвечать на вопросы клиентов, предоставлять информацию о продуктах или услугах, а также помогать в решении технических или административных вопросов.
Заказ товаров и услуг: Виртуальные ассистенты могут принимать заказы от клиентов, рассчитывать стоимость, управлять резервированием и даже оформлять покупки.
Проведение опросов и анкетирование: Чат-боты могут задавать вопросы и собирать информацию от пользователей в ходе опросов, анкетирования и сбора обратной связи.
Развлечения и информационные услуги: Чат-боты могут предоставлять пользовательский контент, такой как анекдоты, новости, гороскопы и другие развлекательные или информационные услуги.
Управление задачами и напоминания

Правила и скрипты: Простые чат-боты могут быть созданы с использованием набора правил и скриптов, которые определяют, как бот реагирует на определенные фразы и команды пользователя.

Машинное обучение: Чат-боты, обученные с использованием методов машинного обучения, могут распознавать и анализировать текстовые запросы пользователей, чтобы предоставлять более точные ответы и улучшать интерактивное взаимодействие.

Нейронные сети и обработка естественного языка (NLP): Современные чат-боты могут использовать глубокое обучение и NLP для понимания смысла текстовых запросов и генерации естественных ответов.

NLTK (Natural Language Toolkit):

Назначение: NLTK - это библиотека для обработки текстовых данных на языке Python. Она предоставляет множество инструментов и ресурсов для анализа текста, включая методы для токенизации, морфологического анализа, синтаксического анализа и многие другие задачи NLP.
Основные возможности: NLTK включает в себя множество корпусов текста, лексических ресурсов и алгоритмов, а также модели машинного обучения для различных задач NLP. Она широко используется в учебных целях и для прототипирования NLP-приложений.

spaCy:

Назначение: spaCy - это библиотека для обработки естественного языка с акцентом на производительность и эффективность. Она предназначена для решения различных задач NLP, таких как токенизация, морфологический анализ, выделение именованных сущностей и синтаксический анализ.

Основные возможности: spaCy предоставляет высокооптимизированные алгоритмы для анализа текста, что делает ее одной из самых быстрых библиотек для NLP. Она также имеет предобученные модели для разных языков и поддерживает создание собственных моделей машинного обучения.

Gensim:

Назначение: Gensim - это библиотека для работы с векторным представлением текста, включая модели Word2Vec, Doc2Vec и другие методы для создания и использования векторных представлений слов и документов.

Основные возможности: Gensim предоставляет инструменты для обучения векторных моделей на больших корпусах текста. Она позволяет выполнять операции с векторами слов, такие как поиск похожих слов, кластеризацию и даже создание векторных представлений для документов.

Teach
Me
Skills

Домашнее задание:
Реализация, обучение и оценка NLP моделей