一．NLP概念

NLP（自然语言处理）是一种让计算机理解、解释和生成人类语言的技术

核心任务：通过计算机程序来模拟人类对语言的认知和使用过程

二．NLP发展历程

早期探索（1940-1960）：这一时期的机器翻译系统非常简单，主要依赖字典查找和基本的词序规则来进行翻译，效果并不理想。

符号主义与统计方法（1970-1990）：研究者分为符号主义（或规则基础）和统计方法两大阵营。符号主义研究者关注于形式语言和生成语法，而统计方法的研究者更加关注于统计和概率方法。

机器学习与深度学习（2000至今）：深度学习模型如循环神经网络（Recurrent Neural Network，RNN）、长短时记忆网络（Long Short-Term Memory，LSTM）和注意力机制等技术被广泛应用于 NLP 任务中，取得了令人瞩目的成果。

三．NLP任务

1.中文分词：中文语言词与词之间没有像英文那样的明显分隔（如空格），所以无法直接通过空格来确定词的边界。因此，中文分词成为了中文文本处理的首要步骤，其目的是将连续的中文文本切分成有意义的词汇序列。

输入：雍和宫的荷花开的很好。

正确切割：雍和宫 | 的 | 荷花 | 开 | 的 | 很 | 好 | 。

错误切割 1：雍 | 和 | 宫的 | 荷花 | 开的 | 很好 | 。 （地名被拆散）

错误切割 2：雍和 | 宫 | 的荷 | 花开 | 的很 | 好。 （词汇边界混乱）

若分词不准确，则文本处理效果会被极大影响

2.字词切分：子词切分特别适用于处理词汇稀疏问题，即当遇到罕见词或未见过的新词时，能够通过已知的子词单位来理解或生成这些词汇。

输出：unhappiness

不使用子词切分：整个单词作为一个单位：“unhappiness”

使用子词切分（假设BPE算法）：单词被分割为：“un”、“happi”、“ness”

3.词性标注：它的目标是为文本中的每个单词分配一个词性标签，如名词、动词、形容词等。通过词性标注，计算机可以更好地理解文本的含义，进而进行信息提取、情感分析、机器翻译等更复杂的处理。

词性标注通常依赖于机器学习模型，如隐马尔可夫模型（Hidden Markov Model，HMM）、条件随机场（Conditional Random Field，CRF）或者基于深度学习的循环神经网络 RNN 和长短时记忆网络 LSTM 等。这些模型通过学习大量的标注数据来预测新句子中每个单词的词性。

4.文本分类：文本分类的关键在于理解文本的含义和上下文，并基于此将文本映射到特定的类别。

文本分类任务的成功关键在于选择合适的特征表示和分类算法，以及拥有高质量的训练数据。

5.实体识别：也称为命名实体识别，是 NLP 领域的一个关键任务，旨在自动识别文本中具有特定意义的实体，并将它们分类为预定义的类别。

假设有一个实体识别任务，目的是从文本中识别出人名、地名和组织名等实体。

输入：李雷和韩梅梅是北京市海淀区的居民，他们计划在2024年4月7日去上海旅行。

输出：[("李雷", "人名"), ("韩梅梅", "人名"), ("北京市海淀区", "地名"), ("2024年4月7日", "日期"), ("上海", "地名")]

6.关系抽取：它的目标是从文本中识别实体之间的语义关系。

输入：比尔·盖茨是微软公司的创始人。

输出：[("比尔·盖茨", "创始人", "微软公司")]

7.文本摘要：目的是生成一段简洁准确的摘要，来概括原文的主要内容。根据生成方式的不同，文本摘要可以分为两大类：抽取式摘要和生成式摘要。

抽取式摘要：抽取式摘要通过直接从原文中选取关键句子或短语来组成摘要。优点是准确性较高。然而，有时候生成的摘要可能不够流畅。

生成式摘要：生成式摘要需要对文本片段进行重新组织和改写，并生成新内容。生成式摘要通常需要复杂模型，如基于注意力机制的序列到序列模型（Seq2Seq）。

Exp. 2021年5月22日，国家航天局宣布，我国自主研发的火星探测器“天问一号”成功在火星表面着陆。此次任务的成功，标志着我国在深空探测领域迈出了重要一步。“天问一号”搭载了多种科学仪器，将在火星表面进行为期90个火星日的科学探测工作，旨在研究火星地质结构、气候条件以及寻找生命存在的可能性。

抽取式摘要：我国自主研发的火星探测器“天问一号”成功在火星表面着陆，标志着我国在深空探测领域迈出了重要一步。

生成式摘要：“天问一号”探测器成功实现火星着陆，代表我国在宇宙探索中取得重大进展。

8.机器翻译：指使用计算机程序将一种自然语言（源语言）自动翻译成另一种自然语言（目标语言）的过程。

9.自动问答：旨在使计算机能够理解自然语言提出的问题，并根据给定的数据源自动提供准确的答案。

自动问答大致可分为三类：检索式问答（Retrieval-based QA）、知识库问答（Knowledge-based QA）和社区问答（Community-based QA）。检索式问答通过搜索引擎等方式从大量文本中检索答案；知识库问答通过结构化的知识库来回答问题；社区问答则依赖于用户生成的问答数据，如问答社区、论坛等。

四．文本表示的发展历程

文本表示的目的是将人类语言的自然形式转化为计算机可以处理的形式，也就是将文本数据数字化，使计算机能够对文本进行有效的分析和处理。

1.词向量

向量空间模型（Vector Space Model, VSM）是 NLP 领域中一个基础且强大的文本表示方法。向量空间模型通过将文本（包括单词、句子、段落或整个文档）转换为高维空间中的向量来实现文本的数学化表示。

# "雍和宫的荷花很美"

# 词汇表大小：16384，句子包含词汇：["雍和宫", "的", "荷花", "很", "美"] = 5个词

vector = [0, 0, ..., 1, 0, ..., 1, 0, ..., 1, 0, ..., 1, 0, ..., 1, 0, ...]

# 16384维中只有5个位置为1，其余16379个位置为0

# 实际有效维度：仅5维（非零维度）

# 稀疏率：(16384-5)/16384 ≈ 99.97%

词汇表是一个包含所有可能出现的词语的集合。在向量空间模型中，每个词对应词汇表中的一个位置，通过这种方式可以将词语转换为向量表示。例如，如果词汇表大小为 16384 ，那么每个词都会被表示为一个 16384 维的向量，其中只有该词对应的位置为 1，其他位置都为 0。

2.语言模型

N-gram 模型是 NLP 领域中一种基于统计的语言模型。N-gram模型的核心思想是基于马尔可夫假设，即一个词的出现概率仅依赖于它前面的N-1个词。这里的N代表连续出现单词的数量，可以是任意正整数。例如，当N=1时，模型称为unigram，仅考虑单个词的概率；当N=2时，称为bigram，考虑前一个词来估计当前词的概率；当N=3时，称为trigram，考虑前两个词来估计第三个词的概率，以此类推N-gram。

N-gram的优点是实现简单、容易理解，在许多任务中效果不错。但当N较大时，会出现数据稀疏性问题。

尽管存在局限性，N-gram模型由于其简单性和实用性，在许多 NLP 任务中仍然被广泛使用。

3.Word2Vec

Word2Vec是一种流行的词嵌入（Word Embedding）技术。Word2Vec的核心思想是利用词在文本中的上下文信息来捕捉词之间的语义关系，从而使得语义相似或相关的词在向量空间中距离较近。

相比于传统的高维稀疏表示（如One-Hot编码），Word2Vec生成的是低维（通常几百维）的密集向量，有助于减少计算复杂度和存储需求。Word2Vec模型能够捕捉到词与词之间的语义关系，比如”国王“和“王后”在向量空间中的位置会比较接近，因为在大量文本中，它们通常会出现在相似的上下文中。Word2Vec模型也可以很好的泛化到未见过的词，因为它是基于上下文信息学习的，而不是基于词典。

4.ELMo

ELMo（Embeddings from Language Models）实现了一词多义、静态词向量到动态词向量的跨越式转变。ELMo模型的主要优势在于其能够捕捉到词汇的多义性和上下文信息，生成的词向量更加丰富和准确，适用于多种 NLP 任务。