Q1 [30 points]:

从NLP领域选择三个方向，每个方向选择一个benchmark dataset。
简单介绍其输入、输出、评估方
法，并列出在该benchmark dataset上表现最佳的前两个模型/方法，简单介绍该模型/方法以及其表现。

1. 文本分类

基准数据集：

AG News

输入：

新闻文章的文本内容。

输出：

四个类别之一（例如：
World, Sports, Business, Science/Technology）。

评估方法：

使用准确率（Accuracy）、宏平均F1分数（Macro F1 Score）等指标评估模型性能。

表现最佳的模型/方法：

BERT（Bidirectional Encoder Representations from Transformers）

BERT 是一种预训练的语言表示模型，通过双向训练来学习词汇的上下文关系。
它在多种NLP任务上表现出色。

BERT 在 AG News 上取得了非常高的准确率，因其对上下文的理解能力强大。

RoBERTa（A Robustly Optimized BERT Pretraining Approach）

RoBERTa 是 BERT 的一个优化版本，通过改进的预训练方法和增加的训练数据，进一步提升了性能。

相较于 BERT，RoBERTa 在 AG News 上展示了更优的性能，特别是在文本分类任务上。

2. 情感分析

基准数据集：

IMDb

输入：

电影评论的文本（正面或负面）。

输出：

情感极性（Positive 或 Negative）。

评估方法：

采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1分数（F1 Score）来评估模型。

表现最佳的模型/方法：

DistilBERT

DistilBERT 是 BERT 的轻量级版本，使其在保持性能的同时提高了推理速度和效率。

在 IMDb 数据集上，DistilBERT 展示了优越的情感分类能力，且推理速度更快。

XLNet

XLNet 是一种自回归模型，结合了BERT和Transformer-XL的优点，能够更好地捕获句子中的依赖关系。

在 IMDb 数据集上表现良好，XLNet 能够处理长文本并提供更准确的情感分类。

3. 机器翻译

基准数据集：

WMT (Workshop on Machine Translation)

输入：

一段文本（源语言，如英语）。

输出：

翻译后的文本（目标语言，如法语）。

评估方法：

使用 BLEU（Bilingual Evaluation Understudy）分数来评估翻译的质量，越高越好。

表现最佳的模型/方法：

方向展示了 NLP 中的文本分类、情感分析和机器翻译领域，分别介绍了对应的基准数据集、输入、输出和评估方法，以及当前表现最佳的模型和其性能表现。
通过这些信息，可以更好地理解 NLP 中的最新进展和技术应用。

Q2 [70 points]:

请你自行搜索并学习以上算法（BPE、BBPE、Unigram、WordPiece、SentencePiece），然后从它们的原理、计算方法、优化目标等方面展开介绍。

注意一：你不需要列举出所有关于算法的细节，我们的评分标准在于你的介绍是否全面且准确
注意二：你可能会对这些算法之间的关系感到困惑。
他们之间的关系并不是完全平行的，例如
WordPiece与BPE的原理十分相似，而SentencePiece采用BPE和Unigram作为训练算法

BPE

BPE（Byte Pair Encoding）

原理与计算方法：
BPE（Byte Pair Encoding） 是一种用于文本处理的算法，最初用于数据压缩，但在自然语言处理（NLP）领域被广泛用于文本的分词和编码。BPE 通过迭代地合并频率最高的字符对，以减少词汇的大小并提高模型的处理效率

1.准备足够大的训练语料，确定期望的subword词表大小；

2.准备基础词表：比如英文中26个字母加上各种符号；

3.基于基础词表将语料中的单词拆分为字符序列并在末尾添加后缀“ </ w>”；本阶段的subword的粒度是字符。例如单词“ low”的频率为5，那么我们将其改写为“ l o w </ w>”：5；

4.统计每一个连续字节对的出现频率，选择最高频的字符对合并成新的subword；

5.重复第4步直到达到第1步设定的subword词表大小或下一个最高频的字节对出现频率为1；

优化目标： BPE 的优化目标是最大限度地减小模型在处理未登录词（out-of-vocabulary words）时的困境，同时在保证编码效率的前提下，尽可能保留语言的语义信息。

BBPE

BBPE（Byte-Level BPE）

原理与计算方法：
BBPE（Byte Pair Encoding） 是字符级分词的一种变体，最初旨在提高文本数据准备的效率。在传统的 BPE 基础上，BBPE 通过在训练语料库中结合字节和字符信息，以实现更高效的编码。在自然语言处理（NLP）领域，BBPE 是一种为模型提供更强表达能力和更好处理未登录词能力的方法。

初始化：将输入文本按照字节进行分割，而非字符。
例如，Unicode 字符串被转换为其对应的字节形式。

统计频率：
统计每个字节的频率。

统计字节对的频率，生成字节对频率表。

合并操作：
找出最常见的字节对（如 'ab'），并将其合并为一个新字节（如 'ab'）。

将新字节添加到词汇表。

重复合并：
继续统计和合并字节对，直至达到指定的词汇表大小或没有更多的字节可合并。

处理多样化文本：
考虑到不同语言的特殊字符，BBPE 可以有效处理多种语言和编码。

优化目标： BBPE 的目标同样是减少未登录词问题，并提高对不同语言和编码格式的适应性。

Unigram

Unigram Language Model

原理与计算方法：
Unigram Language Model 是一种最简单的语言模型，旨在为一系列单词（或词）分配概率。它假设文本中的每个单词都是独立的，且只依赖于单词本身。这种假设使得 Unigram 模型相对简单而且易于实现。

词汇表构建：首先，基于训练文本统计所有单词的出现次数，构建词汇表。

计算概率：
对于每个单词，计算其出现的概率：。

基于独立性假设：
假设每个词是独立选择的，给定一个句子 ，可以计算句子的生成概率为：。

优化目标：
通过最大化训练数据的似然函数，优化单词的选择，从而提高模型的表现。

WordPiece

原理与计算方法：
WordPiece 是一种用于处理自然语言文本的分词技术，最初由 Google 提出，主要用于改进语言模型和词嵌入的效果。WordPiece 特别适用于处理大量未登录词（OOV，即 Out-Of-Vocabulary 词）的情况。
WordPiece 通过一种基于字符的分词算法，将单词分解为更小的单元（如字符或子词），从而降低词汇表的大小，提高模型的泛化能力。

输入：训练语料； 词表大小V

1.准备足够大的训练语料，确定期望的subword词表大小；

2.准备基础词表：比如英文中26个字母加上各种符号；

3.基于基础词表将语料中的单词拆分为最小单元;

4.基于第3步数据训练语言模型，可以是最简单的unigram语言模型，通过极大似然进行估计即可;

5.从所有可能的subword单元中选择加入语言模型后能最大程度地增加训练数据概率的单元作为新的单元;

6.重复第5步直到达到第2步设定的subword词表大小或概率增量低于某一阈值;

优化目标：

比较 BPE，WordPiece 主要目标是最大化在给定上下文下的似然函数，使得生成的子词对齐更有效。

SentencePiece

原理与计算方法：
SentencePiece 是一种用于自然语言处理中的文本分词和处理的工具，最初由 Google 开发。与传统的分词方法不同，SentencePiece 不依赖于语言学的知识，而是将文本视为一个字节序列，使用基于统计的方法自动产生子词（subword）单元。
SentencePiece 的核心思想是通过建模的方式将输入文本切分为一个固定大小的词汇表，允许处理未登录词（OOV，Out-Of-Vocabulary）并提升模型的表现。

数据预处理：
对输入数据进行文本标准化，包括去除标点、分句等处理。

子词训练：
使用无监督的方法，将整个输入文本视为一串字符，构建词汇表。
可以选择用 BPE 或 Unigram 中的任意一种或两者结合。

合并和分割：
通过 BPE 操作逐渐合并字符到子词，而 Unigram 模型则根据每个子词的概率进行选择。

每一轮的合并操作后，调整概率分布，以使得使用的子词组合能更好地表示原始序列。

优化目标:
目标是使用最少的符号表示最多的信息，提高压缩比和处理多种语言的能力，特别是在处理未登录词时非常有效。

4. More to Explore[10 points]

由于numpy手动实现的分词的逻辑和embeding生成的逻辑简单，导致生成的效果不好，和调用官方的库的效果相差甚远，还有待优化