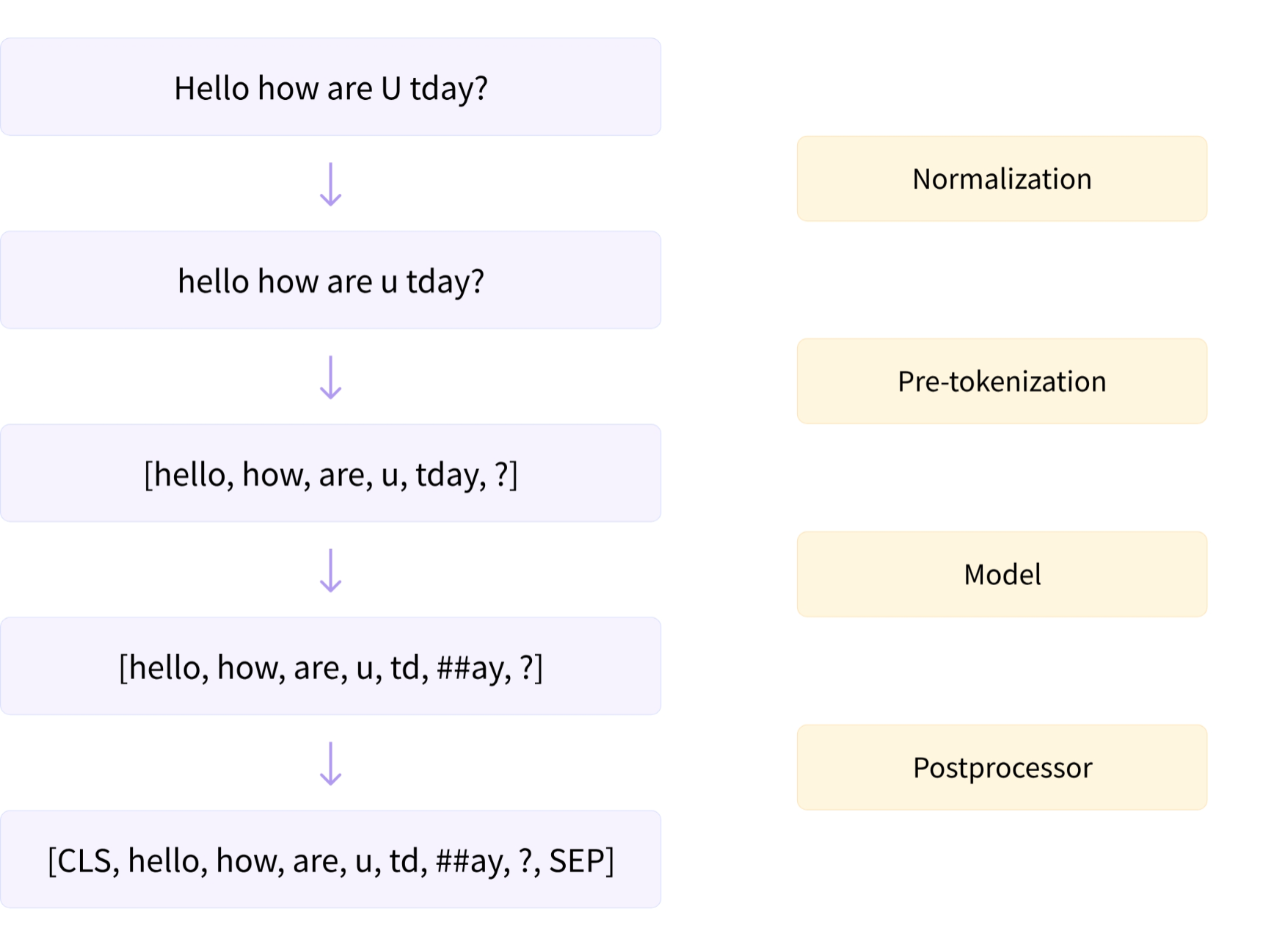
# How to tokenize

## 基本流程



Normalization阶段可以略过，直接从pre-tokenization开始。

## Pre-tokenization

我们首先需要将原始的数据分成更小的实体，比如单词，这就是word based tokenizer。

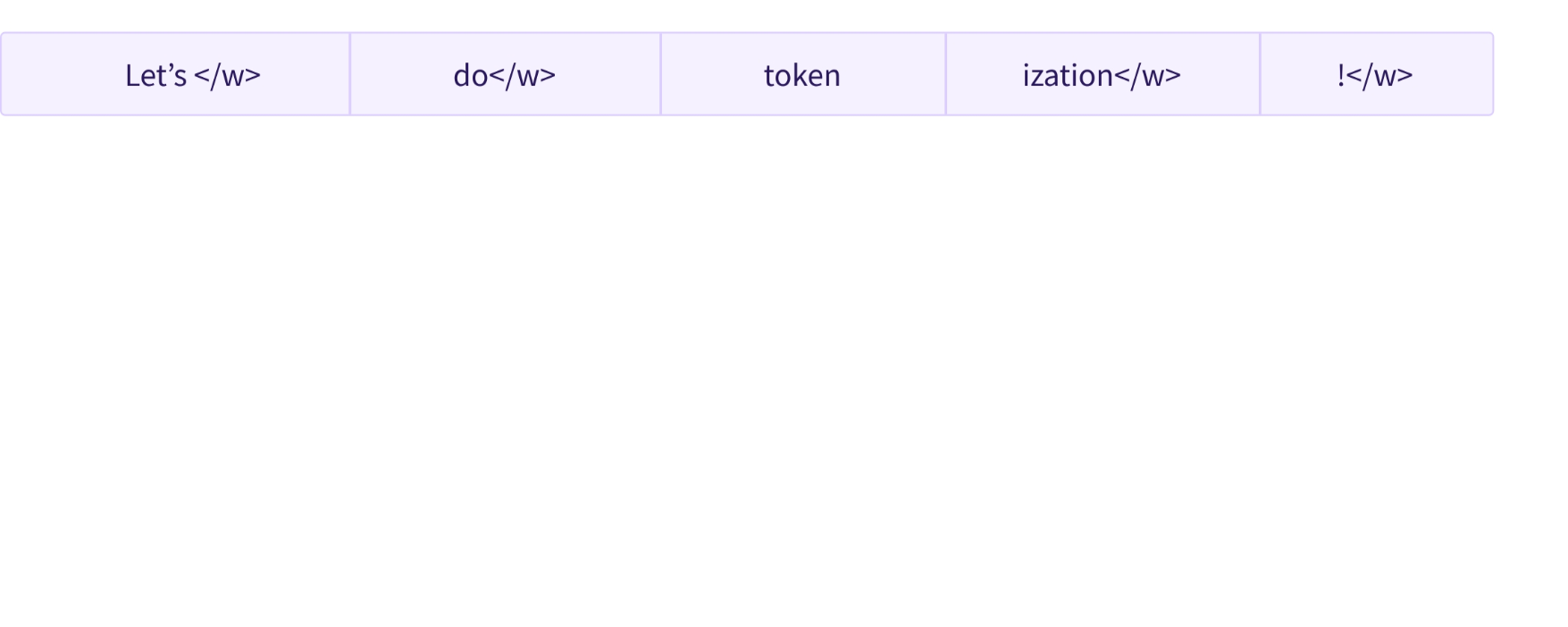
### Character-based tokenizer

由于现在的表达是基于字符而不是单词，从直觉上讲是不如按照word来tokenize的。从另一方面讲，这样对于空间的占用也会更大，也会增加模型的负担。



### Subword-based tokenizer

Subword-based tokenizer依赖于不应该将常用的词拆分为更小的子词，而应该将不常见的词分解为有意义的子词的原则。



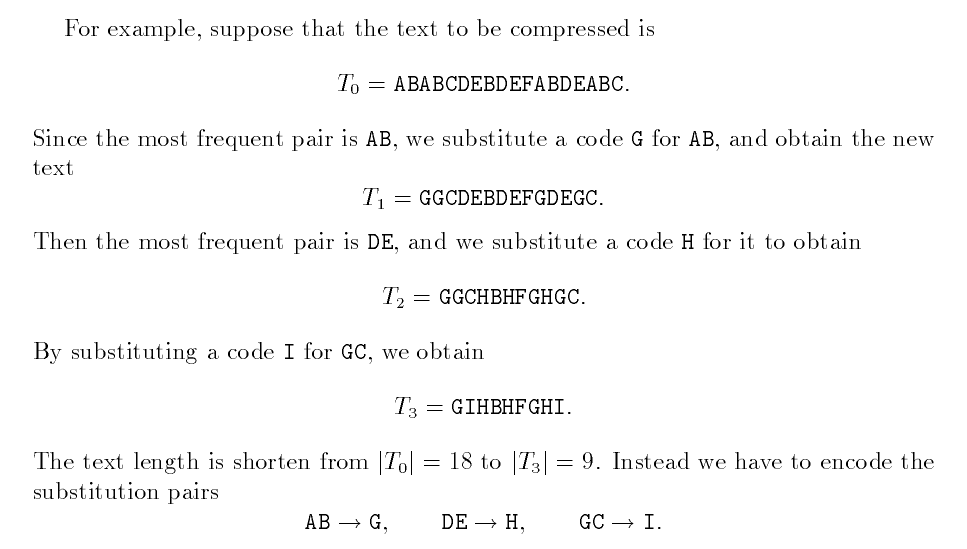
## Tokenization

参考：[Byte-Pair Encoding tokenization - Hugging Face NLP Course](https://huggingface.co/learn/nlp-course/chapter6/5?fw=pt)

常见的分词方法有Byte-Pair Encoding、WordPiece和Unigram三种。

### Byte-Pair Encoding

BPE最早是一种字符串压缩算法，后来才使用到分词中去，其压缩功能可以参考Byte pair encoding: A text compression scheme that accelerates pattern matching这篇文章。*其实也可以使用BPE对长序列DNA数据进行压缩，一篇参考文章Efficient data transfer scheme using word-pair-encoding-based compression for large-scale text-data processing）*



另外附一篇DNA信息化编码综述[Survey of Information Encoding Techniques for DNA | ACM Computing Surveys](https://dl.acm.org/doi/full/10.1145/3626233)

### Unigram

Unigram基于一个最强的假设，即每个词语出现的概率是完全独立，和之前的文本是没有关系的。因此我们只要统计出每一个词语在语料库中出现的概率，就可以估算出句子在某一种切分方式下成立的对应概率。接下来，遍历一个句子所有可能切分方式，找到概率最大的哪一种方式作为分词结果。

1. 建立subword概率词典

统计语料库中所有可能的subword（一般最大长度会限制在4个字符之内），遍历整个语料库，并统计每个subword出现的频数，同时计算每个subword的概率。

比如，good可以分为"g", "o", "d", "go", "oo", "od", "goo", "ood", "good"。

1. 计算每个句子的概率

在当前概率词典下，按照不同的分词方式，可以计算出句子的不同概率：比如，P(good) = P(go)P(od) 或者 P(good) = P(g)P(o)P(o)P(d)。取概率较大的一种作为句子的分词方式。

Unigram中单字子词的概率通常大于多字子词的概率，那么句子会倾向于被划分成一个字一个字的，解决方案有：

**Bigram和N-gram：**使用Bigram可以考虑上下文信息，具体的是利用联合概率：

**引入词频阈值：**在计算概率时对多字词的概率进行额外放大，比如引入长度权重：

另一方面，也可以删除一些频率较低的单字。

### SentencePiece

# 词汇的表示

在建立词汇表后，需要对词汇token进行数值化表示，也即转化成向量。最简单的方法就是使用独热编码，将每个词表示成一个高维**稀疏向量**，但是这种方式有很大的局限性。词汇的分布式表示（Distributed Representation of Words）则是将词汇表示到**密集向量**中去。

分布式表示相较于传统独热表示方式有两个有优点，一是可以捕捉语义相近的关系，这样的单词在向量空间中像距比较近；二是包含一定上下文依赖性，即一个词的意义可以通过他的的上下文来推断，这也是分布式表示的核心思想。

经典的方法包括：

* Word2Vec（基于推理的方法）：利用上下文窗口（如CBOW或Skip-Gram）捕捉词汇与上下文的关系。
* GloVe（基于计数的方法）：基于全局共现矩阵的统计信息生成词向量。
* 预训练模型：如BERT或GPT，这些模型生成分布式表示是上下文敏感的，即一个词在不同的语境中可能有不同的表示（称为动态表示，和前两种的静态表示有所区别）。

词向量平均：用于将一个序列的多个分词表示压缩成一个固定长度的向量，指将每个分词的词向量加和平均在一起，而非序列排布。