- NLP方向总结-分词
  - 分词方法BPE和WordPiece
    - 子词单元，代表形式是BPE, Byte-Pair Encoding 字节对编码，即把一个词可以拆分成多个部分，例如单词This被拆分成, [Th] [#is], 例如工>具subword-nmt
    - BPE与Wordpiece都是首先初始化一个小词表，再根据一定准则将不同的子词合并。词表由小变大
    - BPE与Wordpiece的最大区别在于，如何选择两个子词进行合并：BPE选择频数最高的相邻子词合并，而WordPiece选择能够提升语言模型概率最大的相邻子词加入词表。
    - 都是sub-word，子词型
    - BPE Tokenization Byte Pair Encoding (BPE) , GPT, GPT-2
      Step0: 初始化字典
      Step1: 用单词组成的常用组合的字母表示这个单词，然后在单词末尾追加</w>这个token
      Step2: 循环计数所有的字母组成的字符
      Step3: 合并最频繁出现的组成的字符，加入到字典
    - WordPiece: 是在自然语言处理中使用的子词分割算法。BERT用的此方法。子词分词的一种方法。 用该语言中的各个字符初始化单词表，然后将
      单词表中最常见的符号组合迭代添加到单词表中。 该过程是：1.用文本中的所有字符初始化单词清单。2.使用来自上一步的清单在训练数据上>构建语言模型。3. 通过组合当前单词清单中的两个单元, 将单词组装一个单词单元。 在添加到模型中时，从所有可能增加训练数据可能性中>选择一个新的词单元。 4. 转到2，直到达到预定义的词单元限制或可能性增加低于某个特定阈值。