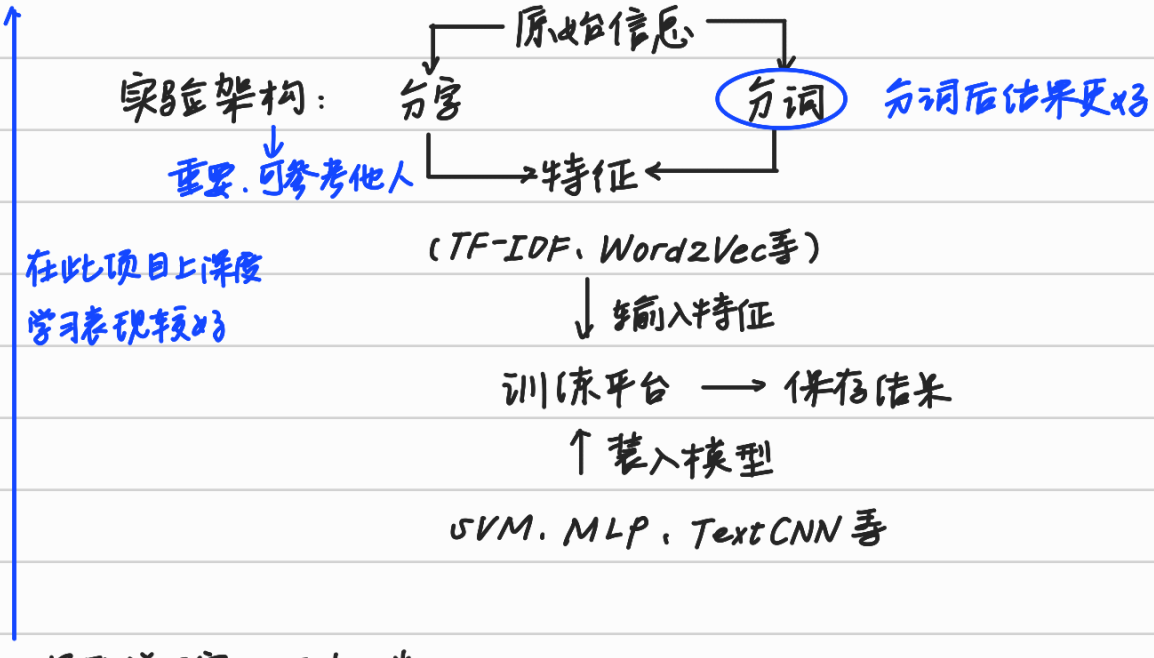
1. 文本数字化

- 向量空间模型
  - 词袋模型：向量中每个位置的值为该编码对应的词的出现次数
  - TF-IDF模型：这个词的出现次数/所有词的出现次数之和
  - → 将一行文本转换成一个向量，没有考虑词与词之间的关系
- 分布式表示—词向量
  - 基于计数的：共现矩阵，经常一起出现的词的词向量相近
  - 基于预测的：Word2Vec，给定前n个词预测第n+1个词
  - → 将一个词表示为低维、稠密的向量

2. 机器学习分类器



3. 深度学习实现文本分类

→ 在小数据集上使用预训练效果不错

(1) FastText：模型简单，在简单文本分类任务中效果不错，直接softmax

(2) TextCNN：用卷积提取局部特征，用池化对特征进行选择
→ 局部信息，并行度好

(3) TextRNN：将上一个词的隐藏层也作为输入传给下一层，"整个文本形成链"
→ 全局信息，并行度差

(4) Attention机制
→ 既有全局信息，并行度也好。

4. 模型融合：关键是多个"好而不同"的模型

4) 模型的多样性

结论 ①分字和分词的差异性最大 ②传统ML和DL差异性较大

③数据增强也有差异性

4.1.3 模型多样性度量

模型多样性度量是指度量模型融合中单模型的多样性，用来刻画单模型的多样化程度，它在模型融合和集成学习中是一个比较重要的问题。比较典型的做法是，考虑单模型的两两相似性或不相似性。总体上，模型多样性度量主要分为两大类：成对的多样性度量和非成对的多样性度量。

首先，假设二分类任务有m个样本，分类器h_j和分类器h_i对样本的预测结果组合情况如表4-1所示。

表4-1 分类器预测结果组合情况

	$h_i=1$	$h_i=0$
$h_j=1$	a	c
$h_j=0$	b	d

成对多样性度量主要有如下参数。

不一致度量：

$$\mathrm{dis}_{i,j}=\frac{b+c}{m} \tag{4.12}$$

相关系数：

$$\rho_{ij}=\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} \tag{4.13}$$

5. BERT：效果很好，只是训练耗费较大