第二单元　文本的智能生成

第 5 课　文本处理的基本单位

学校名称：

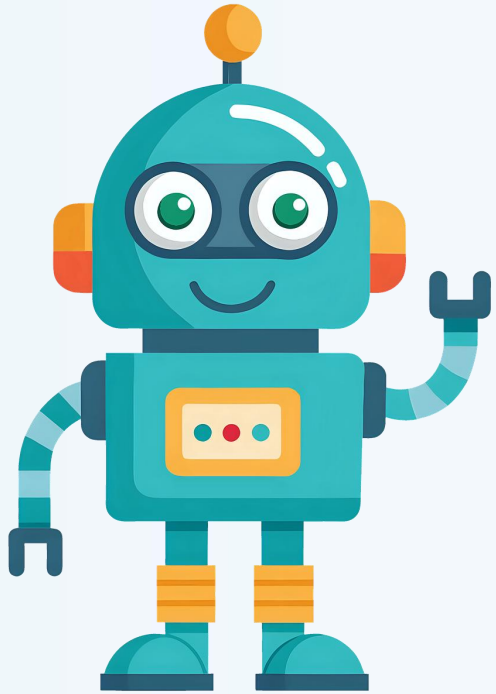
教师姓名：

学习目标

1. 理解人工智能处理自然语言时的基本单位。
2. 了解人与机器分词的差异。
3. 了解一种自动构建词表的算法思想。
4. 了解一种自动分词算法的实现过程。

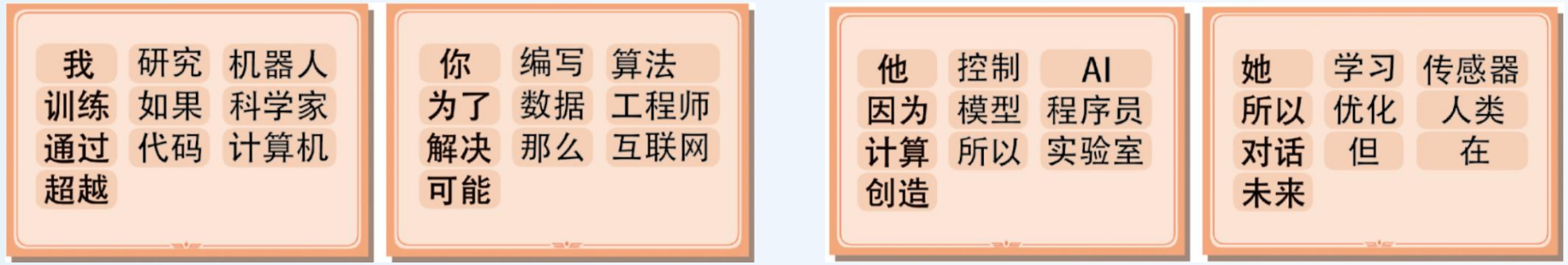
情境导入

人工智能是如何理解人类语言的呢？要理解这一问题，就要先来了解人工智能理解人类语言的基本单位。



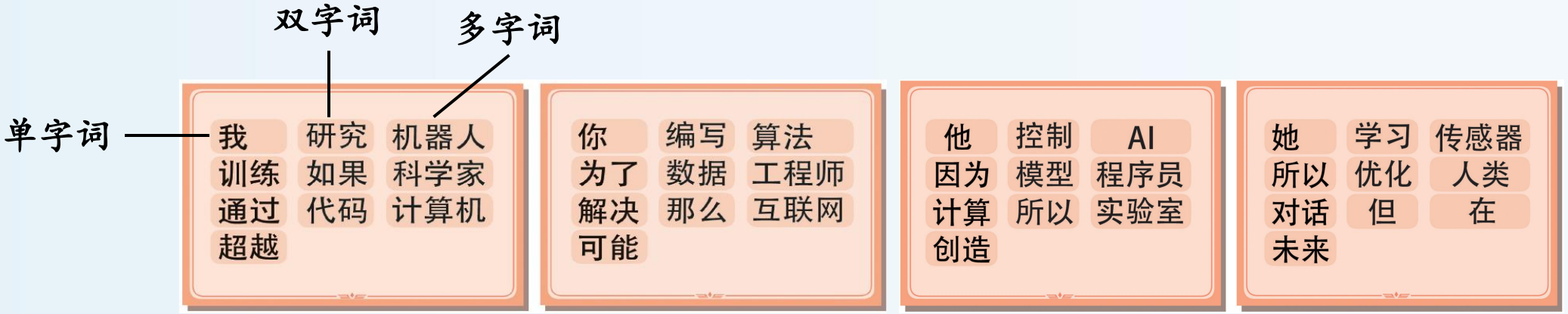
情境导入

尝试利用下面卡片中的词语，组成不同的句子。

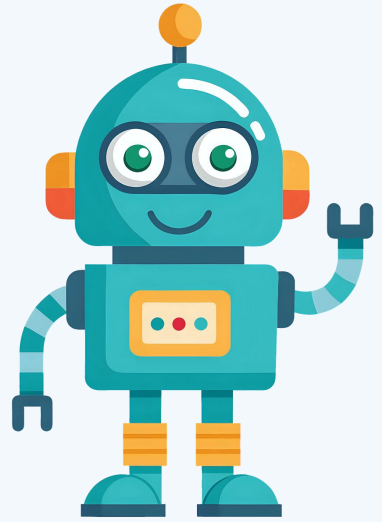


1. 基础组句（如用3到5个词造句）。
2. 逻辑扩展（如用“因为”“所以”等连词造句）。

情境导入



卡片中的单字词、双字词、多字词等，就可以视为模型处理自然语言的基本单位——记号，即token，日常称其为词或词元。



学习内容

1. 词与分词
2. 自动构建词表
3. 自动分词

词与分词

把文本切分成基本单位的过程，就是分词。人对句子分词，主要依靠自身的语言能力。

文本：像刚睡醒一样，眼都张开了，一切欣欣然的样子。

分词：像/ 刚/ 睡醒 / 一样/，/ 眼/ 都/ 张开/ 了/，/ 一切/ 欣欣然/ 的/ 样子/。

词与分词

机器分词的结果可能差异很大。

- 得到适合人理解的词，如绘制词云时的词；
- 得到子词，如 unhappy 分为 un / happy；
- 得到字符，如“子 / 非 / 鱼 / 安 / 知 / 鱼 / 之 / 乐”。

这些都是人工智能需要的词。

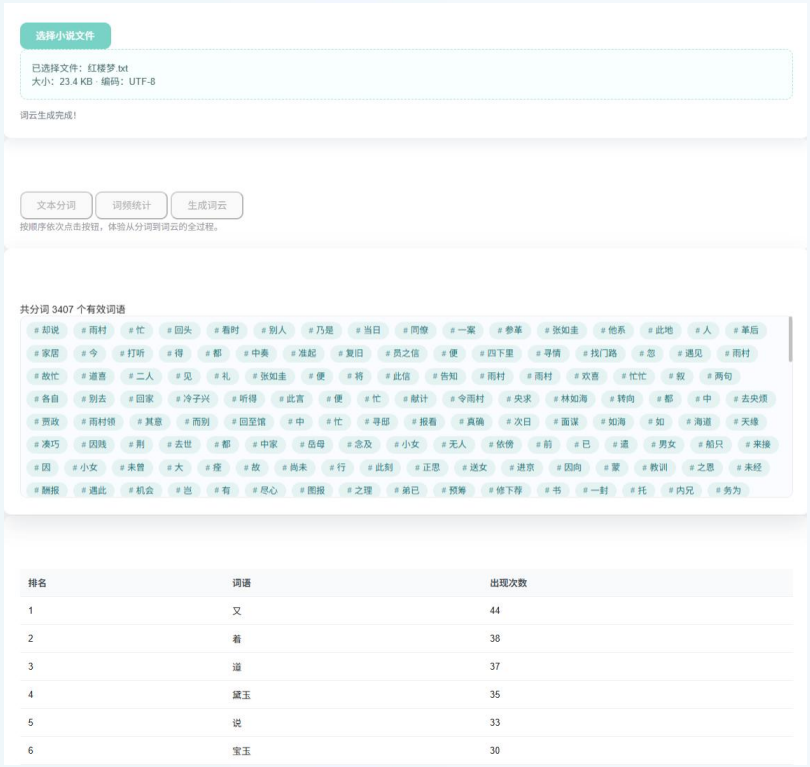
词与分词

体验分词

1. 运行Mixly AI 学习平台中的《小说文本分析》软件。

2. 单击“选择小说文件”按钮，导入小说文档。

3. 再依次单击“文本分词”“词频统计”和“生成词云”按钮进行体验，感受分词。



词与分词

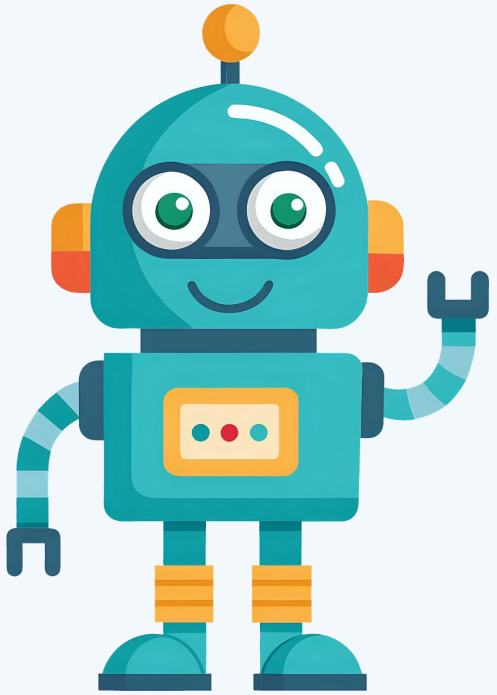
对比采用不同技术分词的异同

1. 运行Mixly AI 学习平台中的《文本分词器》模块。

2. 输入需要分词的文本，然后单击“分词”按钮，比较不同分词方式得到的结果。

词与分词

Jieba 编程库旨在将文本切分为适合人理解的词，但一般认为其分词结果并不适合人工智能模型处理文本。
不同模型有不同的分词策略，对应不同的分词结果。分词结果的数量即使用模型时所需的token数，需考虑输入和输出两个阶段。token数可类比为网络流量。



2 自动构建词表

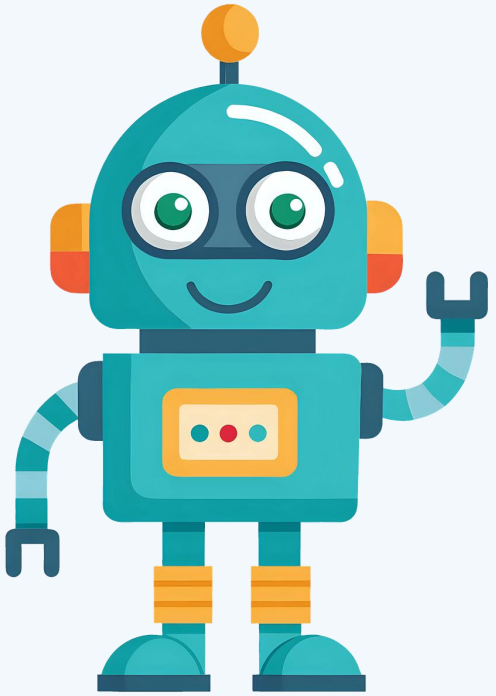
要实现分词，可以先构建一个词表，然后按某种规则把文本中的字符组合跟词表中的条目进行匹配，进而完成分词任务。

不同分词策略对比

词表	人工、人工智能、计算机、人类、智能……
文本	人工智能是利用计算机实现人类智能
长串优先	人工智能 / 是 / 利用 / 计算机 / 实现 / 人类 / 智能
短串优先	人工 / 智能 / 是 / 利用 / 计算机 / 实现 / 人类 / 智能

2 自动构建词表

构建词表是一个耗时巨大的工作，要想准确分词，往往需要构建庞大的词表。而且随着社会的发展，新词不断涌现，靠人工来增补已经非常困难。因此，研究人员提出了很多从文本中自动提取词并构建词表的方法。



2 自动构建词表

基于统计思想自动构建词表

文本数据：

· 人工智能正在改变世界，人工智能技术需要深度学习；

· 人工神经网络通过大量计算实现智能，这种智能需要人工设计算法。

2 自动构建词表

1. 将句子按字符拆解，省略其中的标点，形成初始词表。

2. 检查词表中的字符（或组合）在原文中相邻出现的次数，把次数最多的（至少大于 1）加入其中，同时去除没有单独出现过的。

迭代次数	候选词表
0	人 工 智 能 正 在 改 变 世 界 技 术 需 要 深 度 学 习 神 经 网 络 通 过 大 量 计 算 实 现 这 种 设 法
1	~~人~~ ~~工~~ 智 能 正 在 改 变 世 界 技 术 需 要 深 度 学 习 神 经 网 络 通 过 大 量 计 算 实 现 这 种 设 法 人工 智能

第1次迭代，发现“人工”和“智能”出现4次，这导致作为单字的“人”“工”“智”“能”消失了。

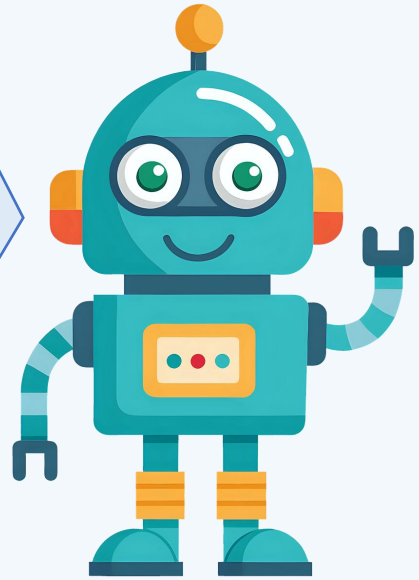
2 自动构建词表

迭代次数	候选词表
2	人 工 智 能 正 在 改 变 世 界 技 术 需 要 深 度 学 习 神 经 网 络 通 过 大 量 计 算 实 现 这 种 设 法 人工 智能 人工智能 需要 计算
3	人 工 智 能 正 在 改 变 世 界 技 术 需 要 深 度 学 习 神 经 网 络 通 过 大 量 计 算 实 现 这 种 设 法 人工 智能 人工智能 需要 计算 _____ _____

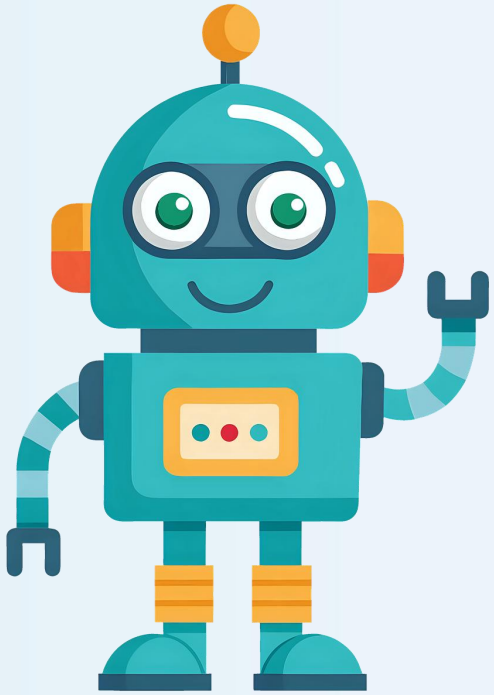
3. 重复第 2 步，继续更新候选词表。

4. 一共找出来 _____ 组字符组合，寻找效果（☐ 好 ☐ 不好）。

第2次迭代，“人工智能”出现2次，使得“人工”和“智能”的出现次数减少，但不至于消失。除此之外，“需要”“计算”也出现了2次。



2 自动构建词表



在实际应用中，现代语言处理系统通常采用人工与机器相结合的方式构建词表。先由机器从大量文本中提取候选词，再由人工进行审核、筛选和优化，如限制词长或剔除无意义组合，从而提高词表质量。

3 自动分词

机器通过大量文本获得词的出现规律，并把规律转换为数值，就可以形成分词模型。分词时，分词模型根据记录进行反复推导，就可计算得出最有可能的切分方式。

3 自动分词

1. 查看词对应的分数，分数越高，表示词出现的可能性越大。

2. 检查“南”：“南”-10，结果为['南']。

3. 检查“京”：“南”+“京”是-30，而“南京”得分是-12，所以结果改为['南京']。

4. 检查“市”：“南”+“京”+“市”是-55，而“南京”+“市”是-37，所以结果为['南京'+'市']。

待分词文本：南京市长江大桥

词	分数	词	分数
南	-10	京	-20
南京	-12	市	-25
长	-12	市长	-20
江	-30	长江	-10
大	-20	桥	-20
大桥	-12		

3 自动分词

待分词文本：南京市长江大桥

5. 检查“长”：“市”+“长”是-37，而“市长”是-20，所以结果改为['南京', '市长']。

6. 检查“江”：“市”+“长”+“江”是-67，“市长”+“江”是-50，而“市”+“长江”是-35，所以结果改为['南京', '市', '长江']。

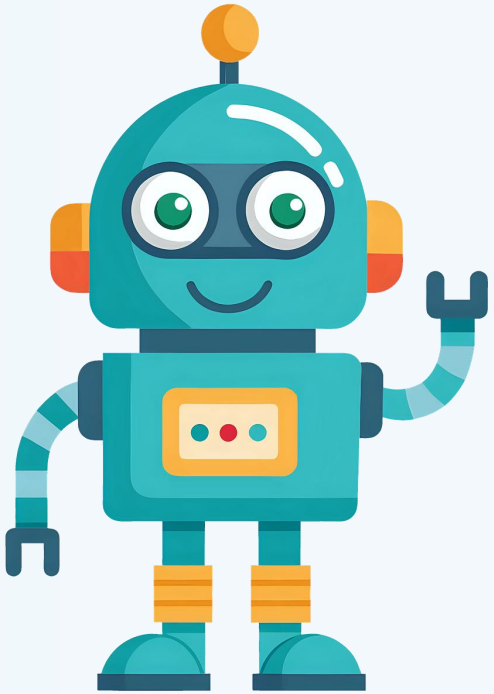
7. 继续向后检查，适当调整分词结果。

8. 记录结果：['南京', '市', '长江', '大桥']

词	分数	词	分数
南	-10	京	-20
南京	-12	市	-25
长	-12	市长	-20
江	-30	长江	-10
大	-20	桥	-20
大桥	-12		

3 自动分词

通过本次分词活动可以发现，机器并不是“理解”句子的意思，而是通过统计和计算词语出现的可能性来完成分词。这种基于得分的方式，是许多人工智能语言处理技术的基础原理之一。



课堂总结

（1）人工智能处理文本时的基本单位为记号（token），日常称其为词或词元。

（2）基于词频统计的分词方法，广泛应用于现代人工智能语言模型中。

（3）机器并不是“理解”句子的意思，而是通过统计和计算词语出现的可能性来完成分词。

拓展提升

假如你是一名人工智能研究员，现在遇到了一个有趣的挑战：有一段来自古代的神秘文字，没人知道它的语言规则，也没有现成的词表可以参考。任务是想办法从这段陌生文本中找出可能的“词”，进而建立初步的词表。

你可以借助人工智能的力量来完成这项任务。