# 第二单元　文本的智能生成

## 第 7 课　文本的语义编码

学校名称：

教师姓名：

# 学习目标

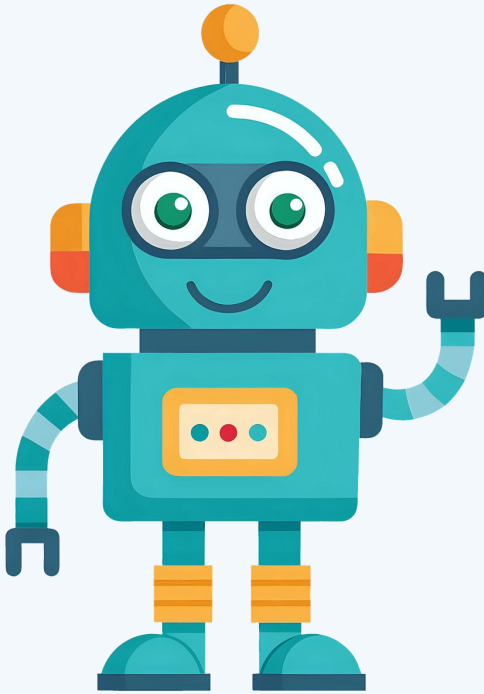
1. 掌握通过语义编码反映语义的方法。
2. 掌握动态调整语义编码的方法。
3. 理解相似度的计算方法及其意义。

# 情境导入

大语言模型在处理文本时，会先将文本分解为基本单位——词，而分词的结果会对应不同的数字编号。下图展示了某个模型分词结果对应的数字编号。


| token | 数字编号 |
| --- | --- |
| 数据 | 3374 |
| 、 | 410 |
| 算法 | 15801 |
| 和 | 548 |
| 算 | 1834 |
| 力 | 910 |
| 是 | 389 |
| 人工智能 | 33574 |
| 的 | 301 |
| 三大 | 23233 |
| 技术 | 2823 |
| 基础 | 4435 |
| 。 | 320 |

这些数字只是编号，无法反映词的含义，因此无法帮助模型理解语义。该怎么解决这个问题呢？



# 学习内容


1. 语义与编码
2. 文本语义编码与相似度

# 语义与编码

用一个数字编号无法充分表示语义特征，研究人员就想到了用一组数字（语义编码）来表示语义特征，进而实现语义上的区分，如表所示。

| 词 | 语义编码 | 词 | 语义编码 |
| --- | --- | --- | --- |
| 猫 | （3，0） | 狗 | （3，1） |
| 汽车 | （2，1） | 飞机 | （1，3） |

# 语义与编码

1. 参照表格，把语义编码视为二维坐标，绘制在坐标系中。

| 词 | 语义编码 | 词 | 语义编码 |
| --- | --- | --- | --- |
| 猫 | （3，0） | 狗 | （3，1） |
| 汽车 | （2，1） | 飞机 | （1，3） |

# 语义与编码

2. 从种类、外形等角度进行观察，查看语义相近的词在坐标系中是否更贴近。

完善的语义编码可以提供多种信息。例如，猫和狗的距离较近、汽车和飞机的距离较近，可能因为它们在种类上更为接近；另外猫、狗与汽车的距离小于与飞机的距离，可能因为猫、狗与汽车都在陆地上活动。

# 语义与编码

如果词的语义编码固定不变，那就无法解决多义词问题，例如：

- 房间**窗口**朝南。
- 关闭计算机**窗口**。

这里的“窗口”虽然形式一样，但意义完全不同。此时应该根据上下文，如“房间”“计算机”等，动态调整“窗口”的语义编码。

# 语义与编码

1. 参照下面的规则，计算句1和句2中的“窗口”的动态语义编码，并填写表格。

2. 在坐标系中把句1和句2中的“窗口”绘制出来，然后说一说现在能否把它们区分开。

**规则1：**

“朝南”使得方向属性增强，即*x*轴数值+1

**规则2：**

“计算机”使得科技属性增强，即*y*轴数值+1

| 环境 | 语义编码 | 环境 | 语义编码 |
| --- | --- | --- | --- |
| 初始 | （0.5，0.5） | 句1中 | |
| 句2中 | | | |

# 2 文本语义编码与相似度

得到单个词的语义编码后，就可以使用不同的组合方式，得到一段文本的语义编码。

例如：把每个词各个角度的特征值加起来，然后对每个角度求平均值；根据重要性给不同词分配不同的权重，再进行加权平均；提取每个角度的最大值或最小值……

通过不同的操作方式，就可以获取文本的整体语义表示，从而用于分类、相似度计算、情感分析等自然语言处理任务。

# 2 文本语义编码与相似度

1. 启动Mixly AI学习平台中的《文本编码》模块，输入一段文本，选择一种编码方式，然后单击“获取编码”按钮，观察获得的编码。

2. 切换不同的编码方式，看看获得的编码是否随之变化。

# 2 文本语义编码与相似度

通过词的语义编码获取文本的语义编码后，就可以进行多种运算，如相似度运算。例如，假定开心的语义编码为（4，3），快乐为（5，4），难过为（1，2），悲伤为（2，1），那么可以通过计算距离来衡量相似度

$$d_{\text{快乐–开心}} = \sqrt{(5-4)^2 + (4-3)^2} = \sqrt{1^2 + 1^2} = \sqrt{2} \approx 1.41$$

一般来说，距离越小表示文本越相似。

# 2 文本语义编码与相似度

1. 运行可以计算文本相似度的软件，输入任意两段文本，查看计算出的相似度。

2. 尝试用不同方式表达相同的意思，或用相近方式表达相反的意思，然后看看这些情况下相似度的计算结果。

# 课堂总结

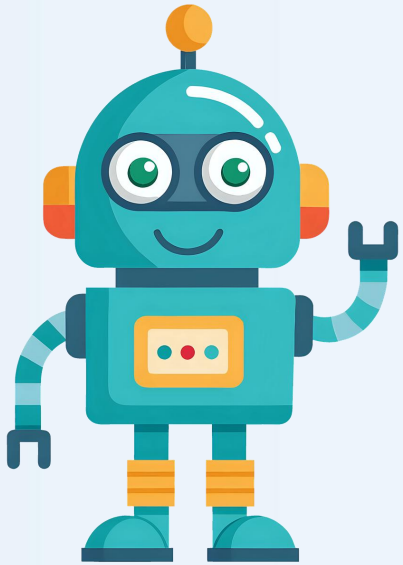
（1）可以用一组数表示词语的语义特征，形成语义编码。

（2）如果词的语义编码固定不变，那就无法解决多义词问题。因此会根据上下文，动态调整语义编码。

（3）获得文本的语义编码后就可以进行多种运算，如相似度运算。我们可以通过计算距离来衡量相似度，一般来说，距离越小表示文本越相似。

# 拓展提升

假如你是一名国际会议的助理，今天收到了一份重要文稿的多个语言版本，如中文、英文和日文。现在，你需要快速判断这些文稿内容是否一致。



请基于文本语义编码和相似度计算方法设计一个软件，用于自动判断多语种文稿内容的一致性。说出软件的设计思路即可。