

第二单元 文本的智能生成

第 6 课 统计与文本生成

学校名称：

教师姓名：

▼ 学习目标



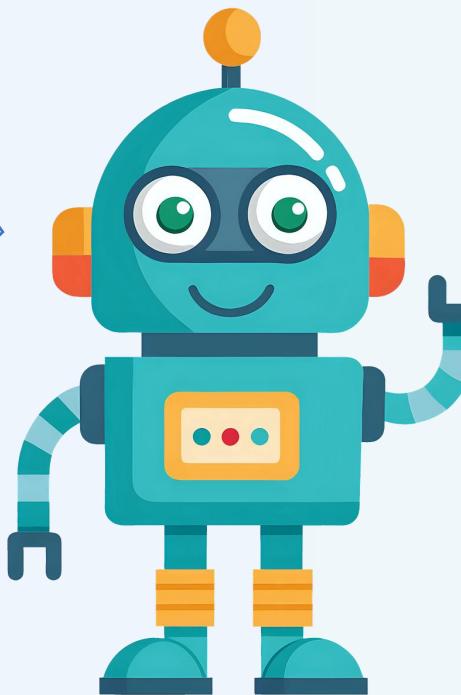
掌握基于前文进行预测的基本原理，并能应用于简单文本生成。



初步了解语言模型如何通过训练数据自动学习语言规律。

情境导入

文本生成中是如何自动预测下一个词的？



为什么前文会影响下一个词的选择？

▼ 学习内容



- 1 根据前文进行预测
- 2 基于相邻字词生成文本
- 3 体验诗句的自动生成

根据前文进行预测

为了清楚地说明这一点，请大家先完成以下语句。

1. 小猫 奔跑 (奔跑 游泳 飞翔)
2. 小鱼 游泳 (奔跑 游泳 飞翔)
3. 小鸟 飞翔 (奔跑 游泳 飞翔)

根据前文进行预测

看到问题1时，会立刻选“奔跑”；看到问题2时，会选“游泳”；看到问题3时，会选“飞翔”。这是因为前面出现的“小猫”“小鱼”“小鸟”对后续内容的选择产生了影响。

也就是说，前文会对后文产生影响。因此，可以根据前文预测后续要出现的内容。

2

基于相邻字词生成文本

前文，尤其是相邻字词，会对后续要出现的字词产生较大影响。因此，统计相邻字词共同出现的次数，就可以进行预测了。

2 基于相邻字词生成文本

文本数据：山深水流远 风起云自闲 月落空林影 水远风声随 云暗月影深 空山声自远 林暗水流深 风起随云闲 影落深水月 月随流水空

行表示前字，列表示后字，组合（山，深）的值为1，表示它们在已有的文本中相邻出现了1次，而（深，水）的值为2，表示相邻出现了2次。

2 基于相邻字词生成文本

根据表格，选择曾经相邻出现的字，尝试组建新的诗句。

深→水→流→ →

空→林→影→ → → →

_____ → _____ → _____ → _____ → _____ → _____

2 基于相邻字词生成文本

根据表格，选择曾经相邻出现的字，尝试组建新的诗句。

深→水→流→ →

空→林→影→ → → →

→ → → → → →

→ → → → → →

2 基于相邻字词生成文本

在适当的地方产生新的关联，然后尝试组建诗句。

山→深→水<自→

月→ ◇云→ 林→影→深

2 基于相邻字词生成文本

在适当的地方产生新的关联，然后尝试组建诗句。

山→深→水<→自→ 闲

月→落 <→云→ 闲 林→影→深

2

基于相邻字词生成文本

不难想象，如果参与训练的数据足够多，建立的统计表足够大，那么就能更加全面、完整地了解诗文中字的组合规律，进而就可以利用组合规律生成诗句。当然，在生成过程中可以有意打破这个规律，从而产生新的组合，带来新的创意。

3 体验诗句的自动生成

1. 观察右表，可以发现很多单元格的值都是0，这些对于生成诗句是无意义的。为了简化，仅记录相邻出现次数至少为1的组合。例如：

'山': {'深':1,'声':1}

'水': {'流':2,'远':1,'月':1,'空':1}

3

体验诗句的自动生成

2. 打开Mixly AI学习平台中的《古诗文本分析器》模块，输入古诗文本，然后单击“开始统计词频”按钮，统计古诗中字的相邻出现次数。

The screenshot displays two main sections of the Mixly AI platform:

- 古诗分析 (Poetry Analysis):** This section shows a poem input field containing a famous Tang Dynasty poem by Wang Wei: "床前明月光，疑是地上霜。举头望明月，低头思故乡。春眠不觉晓，处处闻啼鸟。夜来风雨声，花落知多少。白日依山尽，黄河入海流。欲穷千里目，更上一层楼。红豆生南国，春来发几枝。愿君多采撷，此物最相思。空山不见人，但闻人语响。返景入深林，复照青苔上。君自故乡来，应知故乡事。来日绮窗前，寒梅著花未。移舟泊烟渚，日暮客愁新。野旷天低树，江清月近人。" Below the input field are two buttons: "开始统计字频" (Start Frequency Statistics) and "随机生成诗歌" (Generate Random Poem). A message "字频统计完成!" (Frequency statistics completed!) is displayed above a table titled "字频统计结果" (Frequency Statistics Results).
- AI生成诗歌 (AI Poetry Generation):** This section features a large yellow box containing a generated poem: "‘’ 点击“随机生成诗歌”按钮 AI将基于字频统计 为您创作新的古诗 基于马尔可夫链算法生成，保持古诗韵律" (Click the "Generate Random Poem" button. AI will generate a new poem based on character frequency statistics. It uses the Markov chain algorithm to maintain古诗 rhyme and rhythm.)

当前字	后续字	频率	概率
人	在	4	23.53%
	但	1	5.88%
	语	1	5.88%
	千	1	5.88%
	踪	1	5.88%
	故	1	5.88%
	西	1	5.88%
	物	1	5.88%
	道	1	5.88%
	生	1	5.88%
	间	1	5.88%
	有	1	5.88%
长	1	5.88%	

3

体验诗句的自动生成

3. 基于前面的字频统计结果，点击随机生成诗歌，体验基于字频统计观察生成过程及结果。

古诗分析

床前明月光，疑是地上霜。举头望明月，低头思故乡。
春眠不觉晓，处处闻啼鸟。夜来风雨声，花落知多少。
白日依山尽，黄河入海流。欲穷千里目，更上一层楼。
红豆生南国，春来发几枝。愿君多采撷，此物最相思。
空山不见人，但闻人语响。返景入深林，复照青
君自故乡来，应知故乡事。来日绮窗前，寒梅著花未。
移舟泊烟渚，日暮客愁新。野旷天低树，江清月近人。

开始统计字频
随机生成诗歌

诗歌生成完成！

字频统计结果

当前字	后续字	频率	概率
人	在	4	23.53%
	但	1	5.88%
	落	1	5.88%
	千	1	5.88%
	踪	1	5.88%
	故	1	5.88%
	西	1	5.88%
	物	1	5.88%
	道	1	5.88%
	生	1	5.88%
	间	1	5.88%
	有	1	5.88%
	长	1	5.88%
	家	1	5.88%

AI生成诗歌

乡来日还两
个黄昏点滴
滴这次第怎
敌他晚来风

基于马尔可夫链算法生成，保持古诗韵律

3

体验诗句的自动生成

4. 根据生成过程，思考以下问题：

- 你觉得当前软件生成的诗句存在哪些不足？
- 你认为这个生成诗句的软件理解古诗文吗？

3

体验诗句的自动生成

当前，大语言模型主要基于神经网络捕捉字词间的复杂关系，而非简单记录字词相邻出现次数。例如，大语言模型知道“明月”和“玉盘”同为月亮，还知道“春风”应搭配“温柔”而非“凶猛”。

不过其背后的预测思想是一致的，都是根据已有的文字计算当前可能出现的词。

▼ 课堂总结

- (1) 前文会对后文产生影响，根据前文的内容，可以预测后续可能出现的内容。
- (2) 通过统计相邻字词共同出现的频率，可以进行文本生成和预测，即可以基于训练数据中字词的共现关系来模拟语言行为。

▼ 拓展提升

有时候，要搭配的词并不是紧挨着出现的。例如：

猫吃_____ (鱼 骨头) 狗吃_____ (鱼 骨头)

尽管“猫”和“鱼”、“狗”和“骨头”之间隔着“吃”，但选择的决定因素仍然是“猫”和“狗”。这表明：很多时候，需要考虑更前面的词，才能做出合理预测。

有人提出：可以基于前 n 个词来预测第 $n+1$ 个词。这样就能捕捉更远距离的语言关联。尝试借助人工智能平台，按最多3个词一组训练文本生成模型，并用训练的模型生成几段文本。