自然语言和金融AI：向量化和时间序列

顾立平

01自然语言AI

- 自然语言AI是指能够理解和处理自然语言的人工智能技术。这种技术可以让计算机像人类一样理解语言，从而实现自然语言的交互、语音识别、自动翻译等功能。

- 自然语言AI的核心技术包括自然语言处理、语音识别、情感分析等。它已经被广泛应用于智能客服、智能家居、智能医疗等领域，成为推动数字化转型的重要力量之一。

02为机器处理准备自然语言数据

- 为机器处理准备自然语言数据是指将自然语言文本转化为计算机能够理解和处理的形式。
- 这个过程包括标记化、词干标注、柠檬化、大小写规范化等步骤，以便将完整的单词、成对或多单词等有意义的标记转换为数字标记词典。
- 这样可以使机器更好地理解自然语言，进而实现各种自然语言任务，如机器翻译、情感分析、文本摘要等。

03统计模型和对数函数

- 统计模型和对数函数是一种常用的自然语言处理方法。当将文档表示为数字向量时，我们需要考虑术语出现的时间误差率，然后对文档进行矢量化。由于这是基于频率的，所以我们可以将其视为一种统计模型。
- 在处理术语频率时，最好将对数函数应用于计数，而不是使用原始计数。
- 这是因为当术语出现次数非常大、非常小或者具有极端变化时，对数函数可以帮助我们将计数带回正常的领域内。
- 此外，Zipf定律也强调了对数函数的重要性，因为它可以将术语的频率转换为线性尺度，从而更好地反映语料库中术语的出现情况。

04齐普夫术语计数定律

- 齐普夫术语计数定律是指，在自然语言语料库中，根据频率排序的术语中，第一个项目的频率是第二个项目的三倍，第三个项目的九倍，以此类推。

$$f_1 = 2f_2 = 3f_3 = \ldots$$

- 这意味着在语料库中出现的条目数量与其排名成正比。这个定律可以用公式 $f=kr^{(-1)}$ 来表示，其中 f 是术语的频率，k 是常数，r 是术语的排名。

$$f_r = f(r) = f_1 r^{-1}.$$

- 这个定律也可以用来验证语料库中术语的频率分布是否符合幂律分布。为了更好地展示这种关系，可以绘制术语的频率和它们各自的排名的对数-对数图。

$$f_r = f(r) = f_1 r^{\alpha}.$$

05自然语言文档的各种向量表示

- 自然语言文档的各种向量表示指的是将自然语言文本转换为向量形式的方法。这些方法包括词袋模型、TF-IDF、word2vec等。

✓词袋模型将文本看做是一个无序的词集合，统计每个词在文本中出现的次数作为向量的维度；

✓TF-IDF则是在词袋模型的基础上加入了词的重要性权重，用于衡量某个词对于整个文本的重要程度；

✓word2vec则是基于神经网络的词向量表示方法，能够将语义相近的词语映射到相近的向量空间中。

- 这些向量表示方法可以帮助机器学习算法更好地理解和处理自然语言文本。

06文档或单词袋的术语频率向量表示

- 文档或单词袋的术语频率向量表示是指：把文档或单词袋中的每个单词出现的次数作为向量的一个维度，然后将每个单词出现的次数作为对应维度上的数值，形成一个向量。
- 这个向量表示可以帮助机器学习算法更好地理解和处理自然语言文本。

07项频率逆文档的频率矢量表示

- 项频率逆文档的频率矢量表示是一种用于文本挖掘和信息检索的技术，也称为TF-IDF表示。它是由词频(Term Frequency)和逆文档频率(Inverse Document Frequency)两部分组成。

$$\text{IDF for cat} = \frac{\text{number of documents in corpus}}{\text{number of documents containing cat}+1}$$

- 其中，词频指的是某个词在文档中出现的次数，逆文档频率则是用来衡量这个词在整个语料库中的重要程度。通过将这两个指标结合起来，可以得到一个向量表示，用于描述文档或单词的重要性。这种表示方式被广泛应用于搜索引擎、推荐系统等领域。

08基于潜在语义分析的文档主题向量表示

- 基于潜在语义分析的文档主题向量表示是一种常用的文本处理技术，它可以将文档转换成一个低维向量表示，同时保留文档的主题信息。

- 它通过构建一个词汇-文档矩阵，利用奇异值分解(SVD)算法将矩阵分解成三个矩阵的乘积，从而得到文档的主题向量表示。

- 这种表示方式被广泛应用于文本分类、信息检索等领域。

09由潜在Dirichlet分配确定的文档的主题向量表示

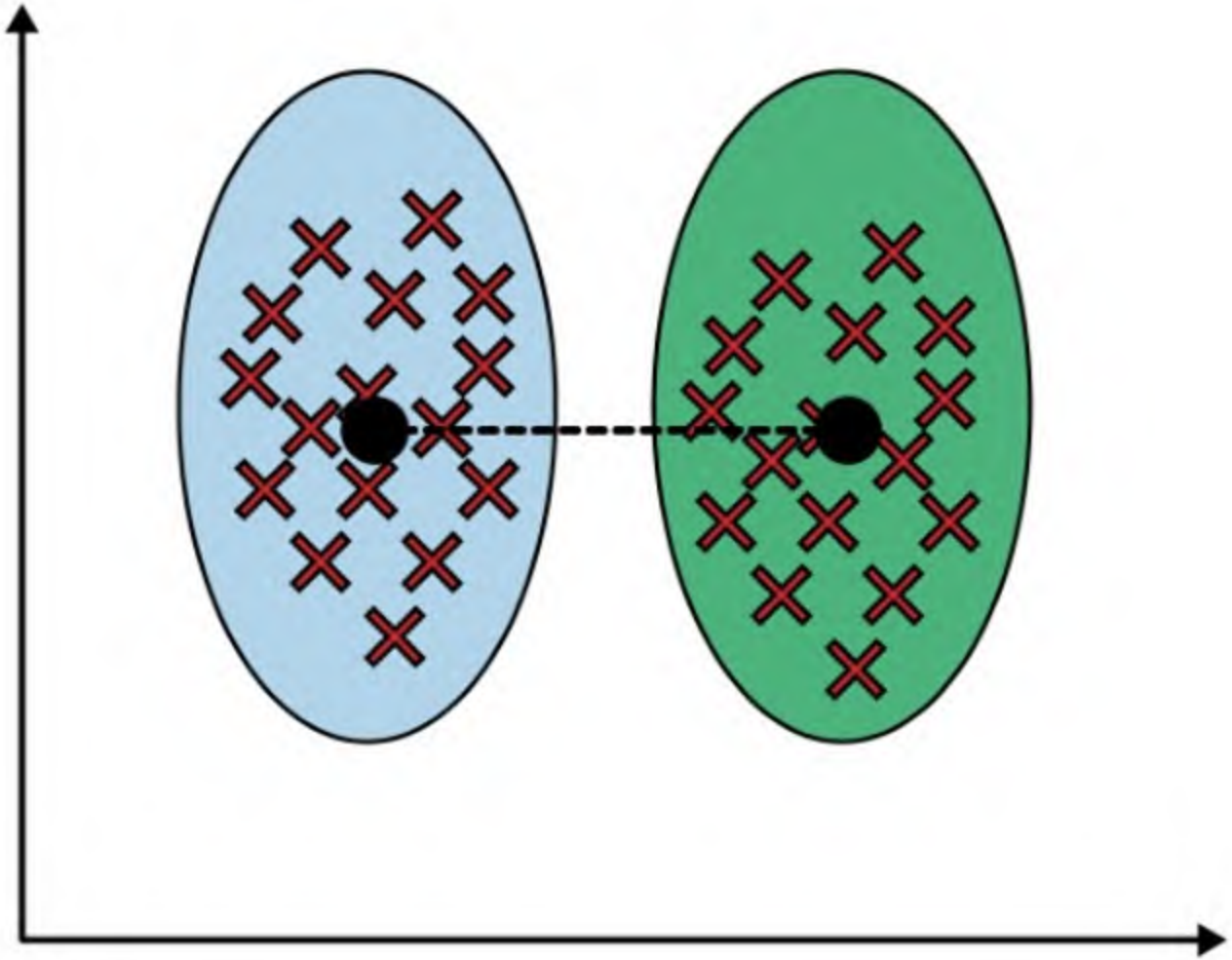
- 潜在Dirichlet分配(PDA)是一种概率模型，用于将文档表示为主题的混合。

✓它假设每个文档都由多个主题组成，每个主题又由一组词项分布组成。

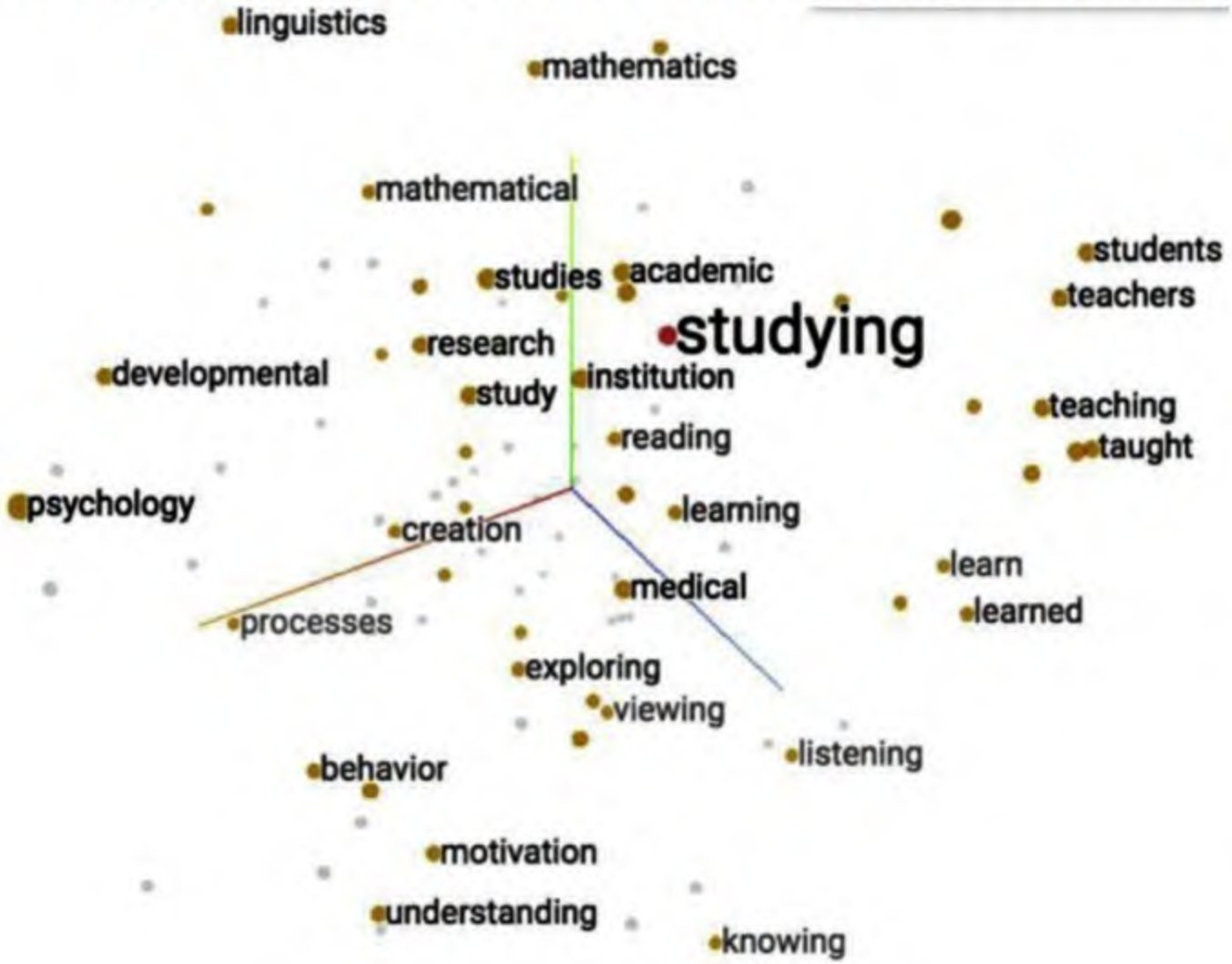
✓PDA的主要优点是可以处理多模态文档，即文档可以属于多个主题。相比之下，潜在语义分析(LSA)只能处理单模态文档，即每个文档只能属于一个主题。

✓但是，由于PDA需要估计大量的参数，因此其计算复杂度较高。

10由潜在判别分析确定文档的主题向量表示



11用神经网络嵌入方法确定单词和文档的意义向量表示



- 神经网络嵌入方法是一种用于确定单词和文档意义向量表示的有效方法。这种方法使用神经网络来学习单词和文档的向量表示，使得语义相似的单词和文档在向量空间中距离较近。

11用神经网络嵌入方法确定单词和文档的意义向量表示

- 具体来说，神经网络通过输入单词或文档的上下文信息来预测下一个单词或文档，并根据预测误差来更新单词或文档的向量表示。
- 这种方法不需要手动设计特征，而是自动学习特征，从而提高了向量表示的质量。

$$Prob(w_n = w \mid w_1, w_2, \cdots, w_{n-1}) = \frac{exp\left(w^t_{vocab_i} w\right)}{exp\left(w^t_{vocab_1} w\right) + exp\left(w^t_{vocab_2} w\right) + \cdots + exp\left(w^t_{vocab_{vocab_{size}}} w\right)}$$

11用神经网络嵌入方法确定单词和文档的意义向量表示

- 常用的神经网络嵌入方法包括Word2Vec和GloVe等。

Probability and ratio	k = solid	k = gas	k = water	k = fashion
P (k|ice)	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
P (k|steam)	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
P (k|ice)/P (k|steam)	8.9	8.5×10^{-2}	1.36	0.96

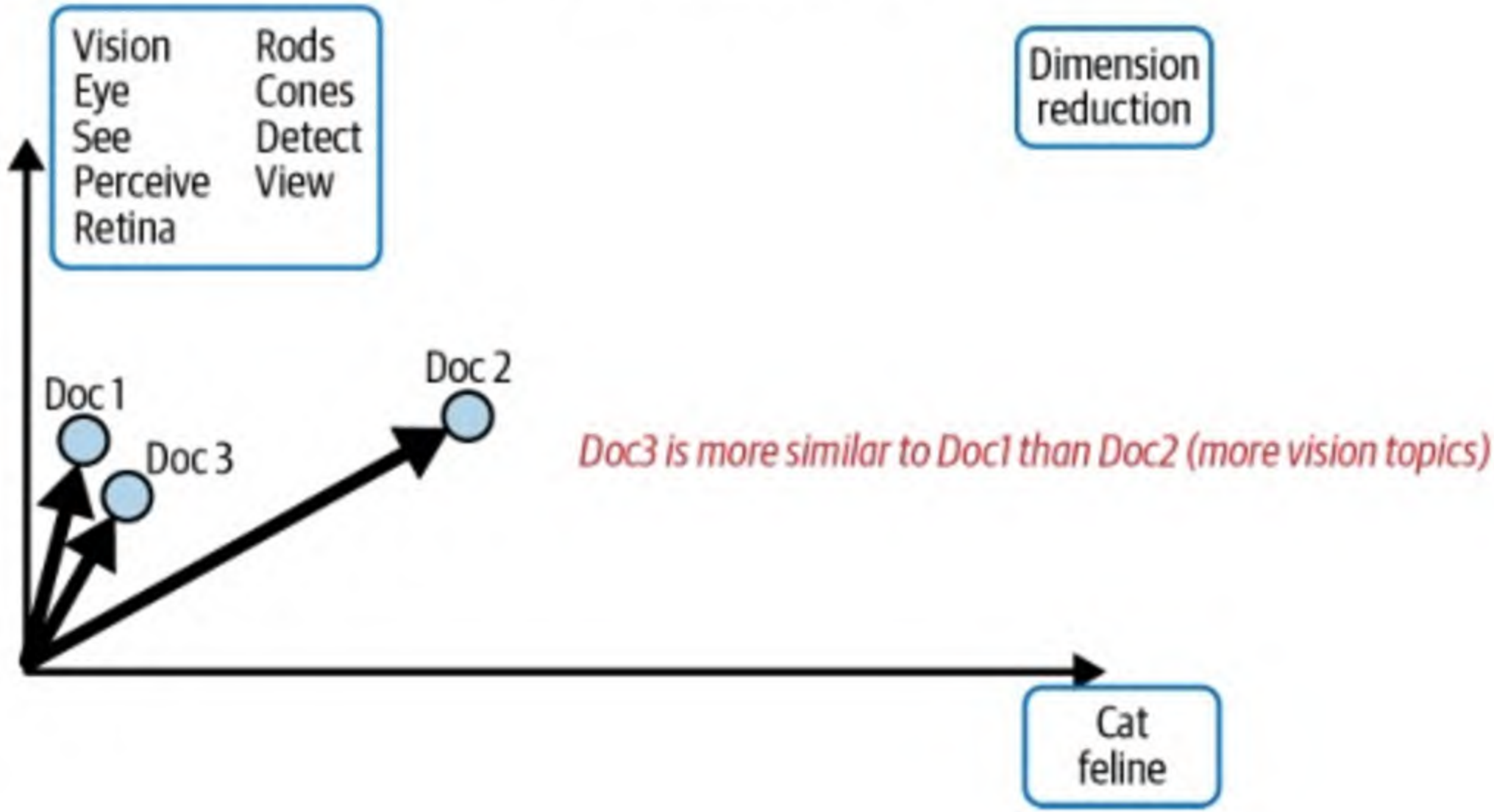
12余弦相似

- 余弦相似是一种常见的文本相似度计算方法，它可以用来衡量两个文本之间的相似程度。该方法基于向量空间模型，将文本表示为向量形式，然后计算这两个向量之间的夹角余弦值。

$$\cos\left(\text{angle between } \overrightarrow{doc_1} \text{ and } \overrightarrow{doc_2}\right) = \frac{\overrightarrow{doc_1}^{t}}{length\left(\overrightarrow{doc_1}\right)} \frac{\overrightarrow{doc_2}}{length\left(\overrightarrow{doc_2}\right)}$$

- 如果余弦值越接近1，说明两个文本越相似；反之则说明它们不相似。余弦相似常用于搜索引擎、推荐系统等领域。

12余弦相似


Vision
Eye
See
Perceive
Retina
Rods
Cones
Detect
View
Dimension reduction
Doc 1
Doc 2
Doc 3
Doc3 is more similar to Doc1 than Doc2 (more vision topics)
Cat
feline

13自然语言处理应用程序

- 自然语言处理（NLP）应用程序是指利用计算机技术对自然语言进行分析、理解和生成的应用程序。

✓这些应用程序可以实现很多功能，例如语音识别、机器翻译、情感分析、文本分类等。

✓在商业领域，NLP 应用程序可以帮助企业自动化客户服务、提高营销效果、优化供应链管理等。

✓在医疗领域，NLP 应用程序可以帮助医生更快速准确地诊断病情、制定治疗方案。总的来说，NLP 应用程序已经成为现代生活中不可或缺的一部分。

14情绪分析

- 情绪分析是一种常见的自然语言处理应用程序，用于从文本中提取情感信息。常用的情绪分析方法包括硬编码规则和机器学习模型。
- 其中，VADER 是一种成功的硬编码规则算法，它能够正确处理标点符号和表情符号，从而传递更多的情感信息。
- 此外，还有基于朴素贝叶斯分类器的情绪分析方法，以及基于潜在判别分析的情绪分析方法等等。通过情绪分析，我们可以更好地了解人们在社交媒体上的态度和情感倾向，为企业和个人提供更好的服务和体验。

15垃圾邮件过滤器

- 垃圾邮件过滤器是一种常用的网络安全技术，用于识别和阻止不需要的电子邮件。通常，垃圾邮件过滤器使用一系列规则和算法来确定一封邮件是否是垃圾邮件。
- 这些规则和算法可以根据不同的特征进行分类，例如发件人地址、主题行、正文内容、附件类型等等。最近，随着深度学习技术的发展，许多研究人员开始探索使用神经网络来进行垃圾邮件过滤。
- 这种方法通常涉及将邮件转换成数字向量，然后使用卷积神经网络或循环神经网络对其进行分类。总的来说，垃圾邮件过滤器对于保护用户隐私和安全非常重要。

16搜索和信息检索

- 搜索和信息检索是指：通过互联网或其他信息系统来查找相关信息的过程。
- 常见的搜索和信息检索方法包括基于文档TF-IDF之间的余弦相似性、基于语义的搜索、基于特征向量迭代等。
- 这些方法可以帮助人们快速找到他们需要的信息，提高工作效率和生活质量。同时，随着技术的进步，人工智能也在不断地改善搜索和信息检索的效果和速度。

17机器翻译

- 机器翻译是一种利用计算机程序自动将一种语言转换成另一种语言的技术。它通常使用统计机器翻译或神经机器翻译等算法来进行翻译。
- 机器翻译的应用范围广泛，例如在国际贸易、旅游、文化交流等领域都有着重要的作用。
- 然而，由于语言的复杂性和多义性等问题，机器翻译仍然存在一些挑战和局限性，需要不断改进和完善。

18图像字幕

- 图像字幕是指在图片上添加文字说明或者描述的一种技术。
- 它可以用于帮助视力障碍者更好地理解和识别图片内容，也可以作为图片的重要补充信息。
- 图像字幕可以手动添加，也可以通过自动化工具自动生成。在数字媒体时代，图像字幕已经成为了一个越来越重要的话题，许多网站和应用程序都提供了图像字幕的功能。

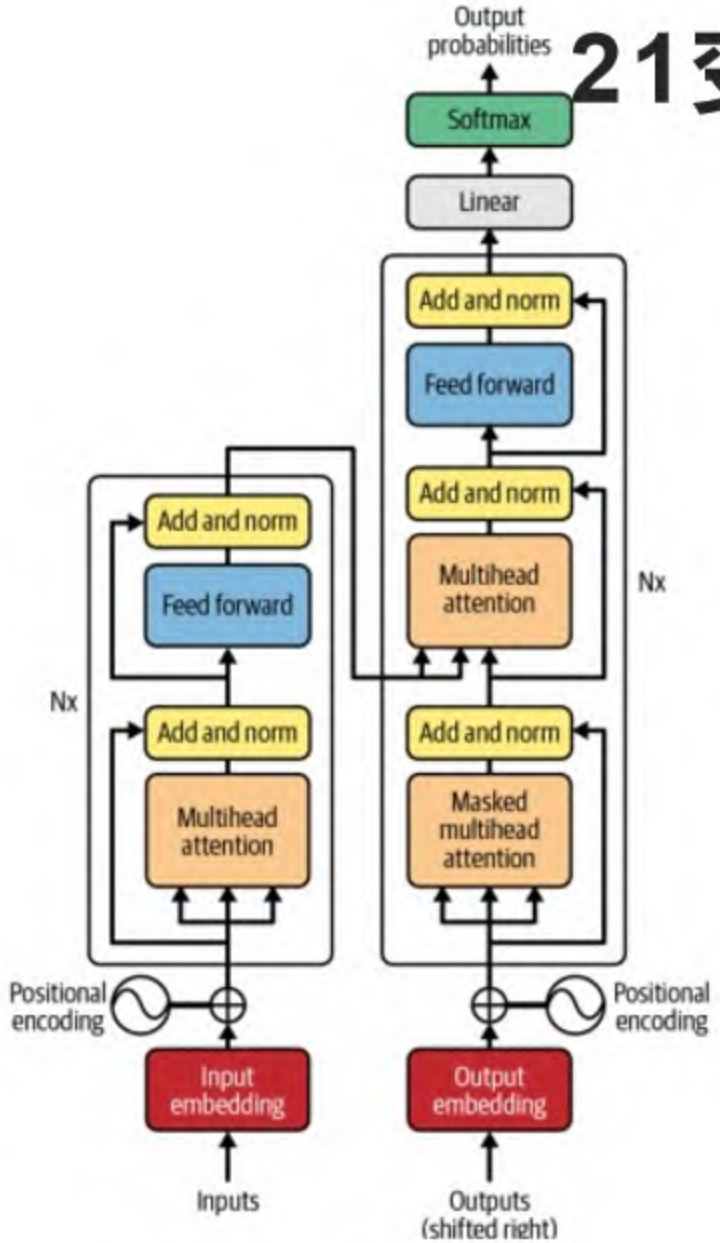
19聊天机器人

- 聊天机器人是一种基于人工智能技术的应用程序，能够模拟人类的对话行为。它们可以通过语音或文本输入与用户交互，解答问题、提供服务、娱乐等。
- 聊天机器人通常采用自然语言处理技术和机器学习算法来实现对用户输入的理解和回应。
- 近年来，随着人工智能技术的发展，聊天机器人已经在各个领域得到了广泛的应用，如客户服务、在线购物、医疗咨询等。

20其他应用程序

- 除了聊天机器人外，自然语言处理技术还可以应用于许多其他领域，如信息抽取、情感分析、机器翻译、自动摘要、语音识别等。其中，信息抽取可以用来从大量文本中提取出结构化的信息；情感分析可以帮助企业了解消费者对其产品和服务的态度和情感倾向；机器翻译可以将一种语言的文本转换成另一种语言的文本；自动摘要是从一篇较长的文章中自动生成简短的摘要；语音识别则可以将人的语音转化为文字形式。
- 这些应用程序都是基于自然语言处理技术的基础上发展起来的，通过不断地研究和发展，将会给人们的生活带来更多的便利和创新。

21变换器和注意力模型

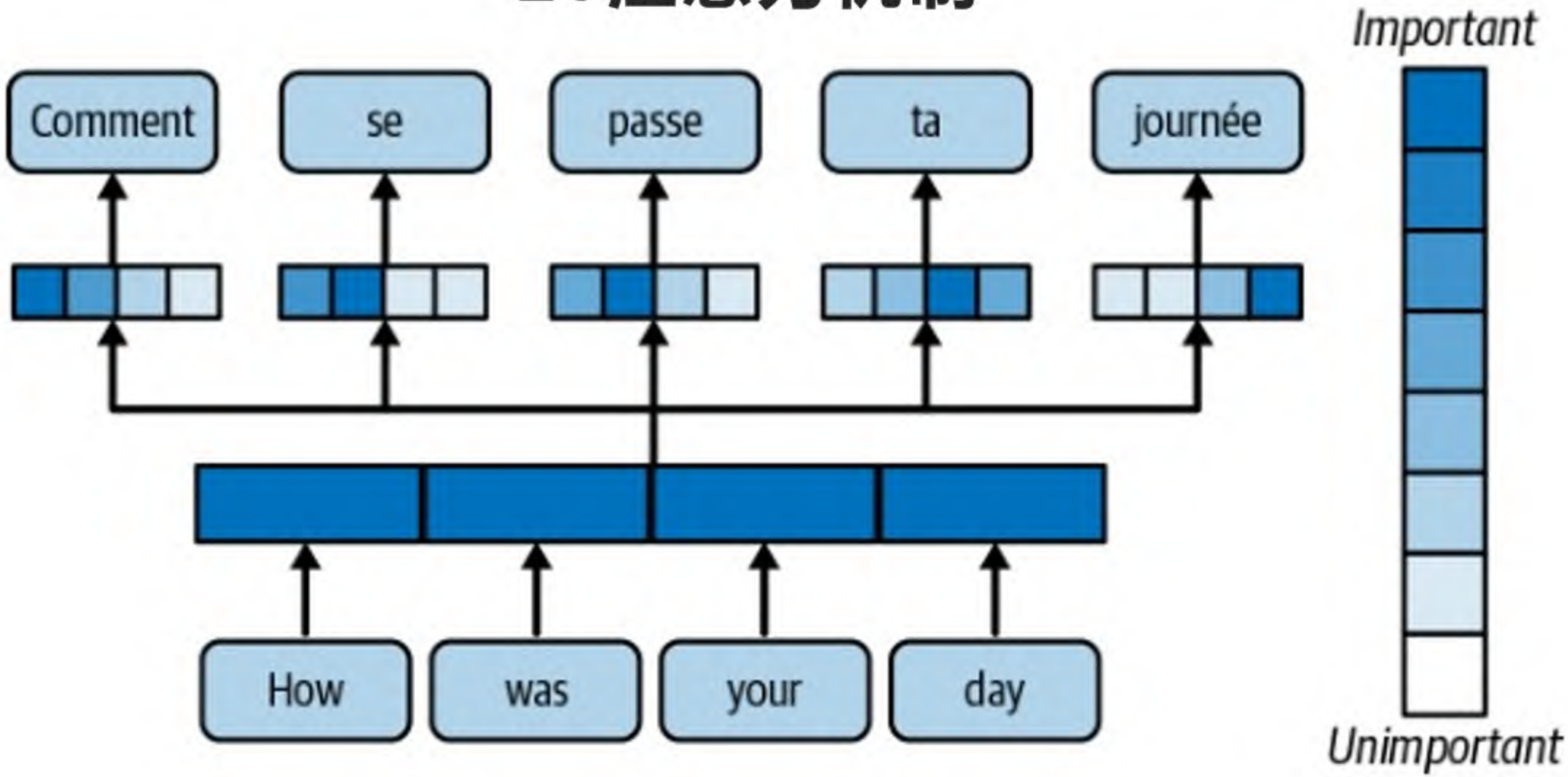


- 变换器是一种神经网络模型，用于处理自然语言处理任务，如机器翻译、文本摘要、问答系统等。它采用了多头自注意力机制，能够同时考虑输入序列中所有单词之间的关系，从而更好地捕捉长距离依赖关系。
- 注意力机制通过对输入序列中每个单词赋予不同的权重，使得模型更加关注与当前预测相关的单词，而不是所有单词。这种机制可以有效地提高模型的表现，并且具有很好的解释性。

22变换器架构

- 变换器架构由编码器和解码器两部分组成。编码器负责将输入序列转换成一系列隐藏状态，其中每个隐藏状态都包含了输入序列中所有单词的信息。
- 解码器则根据编码器产生的隐藏状态和上一个时间步的输出来生成下一个时间步的输出。
- 在编码器和解码器内部，变换器采用了多头自注意力机制，通过多次计算不同类型的注意力权重来提取输入序列的不同特征。
- 此外，变换器还引入了残差连接和层归一化技术，以加速训练过程并提高模型性能。

23注意力机制


Comment
se
passe
ta
journée
How
was
your
day
Important
Unimportant

23注意力机制

- 注意力机制是一种用于处理序列数据的技术，它可以自动地对输入序列中的不同部分赋予不同的权重，以便更好地捕捉序列中的关键信息。在自然语言处理中，注意力机制被广泛应用于机器翻译、语音识别、文本摘要等任务中。

$$\overrightarrow{query} = W_q \vec{w}$$

$$\overrightarrow{:key} = W_k \vec{w}$$

$$\overrightarrow{value} = W_v \vec{w},$$

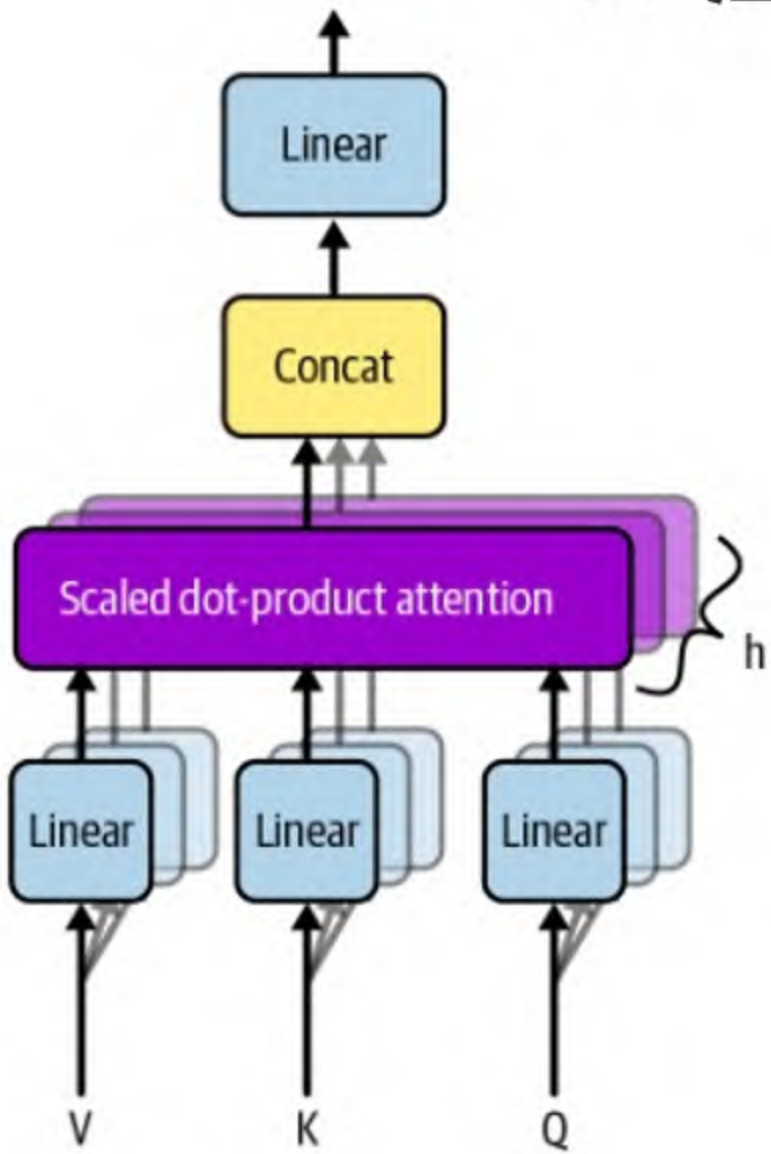
23注意力机制

- 注意力机制通过计算输入序列中每个位置与目标位置的相关性得分，从而确定对于当前目标位置最为重要的输入位置。这种机制可以帮助模型更好地理解输入序列的含义，从而提高模型的性能。

$$alignment_{cooking,kitchen} = \frac{1}{\sqrt{l}} \overrightarrow{query}^{t}_{cooking} \overrightarrow{key}_{kitchen}$$

$$context_{cooking} = \omega_{cooking,I} \overrightarrow{value}_{I} + \omega_{cooking,love} \overrightarrow{value}_{love}$$
$$+ \omega_{cooking,cooking} \overrightarrow{value}_{cooking} + \omega_{cooking,in} \overrightarrow{value}_{in} + \omega_{cooking,my} \overrightarrow{value}_{my}$$
$$+ \omega_{cooking,kitchen} \overrightarrow{value}_{kitchen}$$

23注意力机制



$$\omega_{cooking,kitchen} =$$

$$softmax\left(alignment_{cooking,kitchen}\right) =$$

$$\frac{exp\left(alignment_{cooking,kitchen}\right)}{\left\{exp\left(alignment_{cooking,I}\right) + exp\left(alignment_{cooking,love}\right) + exp\left(alignment_{cooking,cooking}\right) + exp\left(alignment_{cooking,in}\right) + exp\left(alignment_{cooking,my}\right) + exp\left(alignment_{cooking,kitchen}\right)\right\}}$$

24变换器远非完美

- 变换器（Transformer）是一种基于自注意力机制（Self-Attention）的神经网络结构，在自然语言处理领域取得很大成功。然而，尽管变换器在很多任务上表现优异，但它仍然存在一些局限性和不足之处。

✓变换器在处理长序列时可能存在效率问题。由于其采用了自注意力机制，导致在计算过程中需要考虑所有输入序列的位置，因此当序列长度增加时，计算复杂度呈指数级增长，训练时间和计算资源都会受到很大影响。

✓变换器可能会出现过拟合问题。由于其具有很强的表达能力，如果训练数据量不足或者噪声较多，变换器容易在训练集上表现出色，但在测试集上的泛化能力却不够强。

✓变换器对于输入序列的顺序敏感程度较低，可能导致一些顺序相关的任务表现不佳。例如，在机器翻译中，源语言和目标语言之间的语序差异可能会影响翻译质量，但变换器并没有专门考虑到这一点。

25时间序列数据的卷积神经网络

- 时间序列数据的卷积神经网络（Convolutional Neural Network for Time Series Data）是一种常用于处理时间序列数据的神经网络结构。它主要利用卷积操作来捕捉时间序列数据中的局部特征，从而实现对整个序列的建模。
- 时间序列数据的卷积神经网络通常包含多个卷积层和池化层，其中卷积层主要用于提取局部特征，池化层则用于降低特征维度。在网络的最后一层，一般采用全连接层来进行分类或回归等任务。
- 相比于传统的循环神经网络（Recurrent Neural Network），时间序列数据的卷积神经网络具有更少的参数数量和更快的训练速度，同时也能够更好地处理大规模的时间序列数据。因此，在实际应用中，时间序列数据的卷积神经网络已经被广泛应用于股票预测、天气预报等领域。

26时间序列数据的递归神经网络

- 时间序列数据的递归神经网络（Recursive Neural Network for Time Series Data）是一种常用于处理时间序列数据的神经网络结构。它主要利用递归操作来建立时间序列数据之间的依赖关系，从而实现对整个序列的建模。
- 时间序列数据的递归神经网络通常包含一个递归单元，该单元接收当前时刻的输入和前一时刻的状态，然后根据一定的规则计算出当前时刻的状态。在网络的最后一层，一般采用全连接层来进行分类或回归等任务。
- 相比于时间序列数据的卷积神经网络，时间序列数据的递归神经网络能够更好地处理序列中的长期依赖关系，但同时也会面临梯度消失和梯度爆炸等问题。因此，需要根据具体情况选择合适的神经网络结构。

27递归神经网络是如何工作的？

- 递归神经网络（Recurrent Neural Network, RNN）是一种常用的神经网络结构，主要用于处理序列数据。它的核心思想是通过引入循环结构，使得网络可以处理任意长度的输入序列，并且能够在处理序列的过程中保存之前的状态信息。工作方式如下：

1. 对于给定的输入序列，将其按照时间步依次送入网络中；

2. 在每个时间步，将当前的输入和前一时刻的状态作为输入，经过一系列的非线性变换后得到当前时刻的状态；

3. 将当前时刻的状态作为下一时刻的输入，不断重复上述过程，直至遍历完整个输入序列；

4. 最后，根据网络的输出，可以进行分类、回归等任务。

28门控递归单元和长短期存储器单元

- 由于递归神经网络中存在循环结构，因此在反向传播过程中会出现梯度消失或梯度爆炸的问题。为了解决这个问题，后来的研究者们提出了一些改进的递归神经网络结构，例如LSTM和GRU等。这些结构在一定程度上缓解了梯度消失和梯度爆炸的问题，同时也提高了网络的性能。

- 门控递归单元（Gated Recurrent Unit, GRU）和长短期存储器单元（Long Short-Term Memory, LSTM）都是递归神经网络的改进版本，用于解决递归神经网络中存在的梯度消失和梯度爆炸等问题。

28门控递归单元和长短期存储器单元

- GRU是一种比较简单的门控机制，它包括一个更新门和一个重写门，用来控制信息的流动和更新状态。而LSTM则是一种更加复杂的门控机制，它包含了三个门控：输入门、遗忘门和输出门，用来控制信息的流入、流出和保留。相比于GRU，LSTM具有更强的记忆能力和更好的性能表现。
- 在实际应用中，选择使用哪种门控递归单元主要取决于具体的任务需求和数据特征。一般来说，如果需要处理较短的序列或者数据特征比较简单，可以选择GRU；如果需要处理较长的序列或者数据特征比较复杂，可以选择LSTM。

29自然语言数据示例

- 自然语言数据示例是指用于训练和测试自然语言处理模型的文本数据集合。比如，IMDb公司的电影评论数据集、WaveNet的音频数据集等。这些数据集通常由大量的文本数据组成，可以用于训练和评估各种自然语言处理模型，如情感分析、机器翻译、语音识别等。
- 通过使用这些数据集，可以帮助研究人员更好地理解和探索自然语言处理技术的本质，从而推动这一领域的发展。

30金融AI

- 金融AI是指利用人工智能技术来解决金融领域中的问题和挑战的一种方法。在金融领域，人工智能可以应用于风险管理、投资组合管理、信用评估等方面。例如，可以使用机器学习算法来预测股票市场的走势，或者使用自然语言处理技术来分析新闻报道和社交媒体上的信息，以便更好地了解市场情绪和趋势。
- 此外，还可以使用强化学习算法来进行资产配置和交易决策，以提高投资组合的表现。
- 金融AI是一种强大的工具，可以帮助金融机构更好地管理和分析数据，提高业务效率和准确性。

谢谢！

gulp@mail.las.ac.cn