1、自然语言处理实现各种应用目标最终需要解决的关键问题是

答:歧义消解问题和未知语言现象的处理问题.

2、语料库定义

答:语料库:存放语言材料的数据库。语料库是为一个或者多个应用目标面专门收集的,有一定结构的、有代表的可被计算机程序检索的、具有一定规模的语料集合。

3、文本清洗的目的

答:主要是将非文本内容剔除，只保留文本部分

*4、分词的主要目的

答:是将连续的文本字符串分割成有意义的词语序列

*5、基于统计的分词方法基本思想

答:词是字的稳定的组合，相邻的字同时出现的频率越大，就越有可能构成一个词

*6、基于统计的分词方法算法实现

答:预先统计训练文本中相邻出现的各个字的组合的频度，然后计算它们之间的互现信息(它体现了汉字之间结合关系的紧密程度)，当该值高于某一个阈值时，便可以认为此字组可能构成了一个词

7、停用词定义

答:停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据(或文本)之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop·Words(停用词)

*8、停用词种类并举例说明

答:(1)一类是人类语言中包含的功能词，这些功能词极其普遍，与其他词相比，功能词没有什么实际含义、比如“the'、'is'、'at'、'which'、'on'等

(2)另一类词包括词汇词，比如'want'等，这些词应用十分广泛，但是对这样的词搜索引擎无法保证能够给出真正相关的搜索结果,难以帮助缩小搜索范围，同时还会降低搜索的效率，所以通常会把这些词从问题中移去，从而提高搜索性能

9、词性标注的难点在于

答:兼类词和未登录词的词性标注

*10、隐马尔科夫模型五个参数在中文词性标注中的含义是什么?

答:(1)隐藏状态 S 对应词性标注中的词的状态词性集合，如 nr、p、qt

(2)可观察状态 O 对应词性标注中的所有语料库中的词集合，如张三、于、1942 年...

(3)初始状态概率矩阵π对应词性标注中的词中各种隐藏状态的初始概率

(4)隐藏状态转移概率矩阵 A 对应词性标注中的词性之间的相互转移概率，如

P(nr,p)

(5)观测状态转移概率矩阵 B 对应词性标注中的每一个词到各自词性的概率，如 P(张三，nr)、P(于，p)

*11、隐马尔科夫模型解决词性分析问题的实质

答:隐马尔科夫模型解决词性分析问题的实质是对每一个词隐藏的词性求最大联合概率密度问题

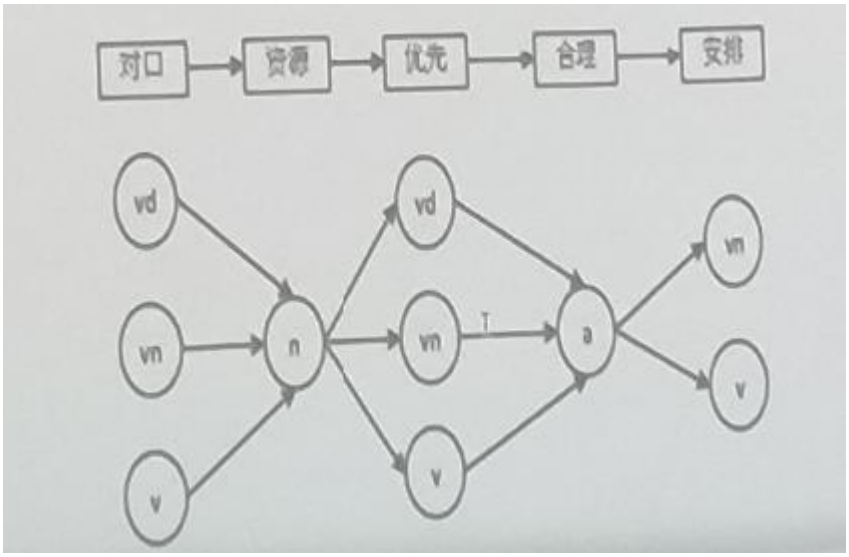
*12、基于隐马尔可夫模型进行词性标注 (不全)

给定句子“对口资源优先合理安排”，通过分词得到结果“对口·|·资源·|·优先·|·合理·|·安排”。

词语“对口”包含词性“vd”、“n”、“v”；

词语“资源”包含词性“n”；

每一个词语的词性之间产生有向依赖关系形成一个概率有向图模型



上图的概率有向图模式中的边，即为隐马尔科夫模型的隐藏状态转移概率，是整个隐藏状态转移矩阵中的一部分,如下所示为隐藏状态的转移概率矩阵

	n	vd	vn	v	a
n	0.45	0.12	0.13	0.25	0.23
vd	0.06	0.10	0.20	0.16	0.12
vn	0.11	0.26	0.23	0.31	0.10
v	0.39	0.13	0.26	0.21	0.33
a	0.11	0.15	0.20	0.28	0.05

将有向概率图与隐藏状态转移概率矩阵结合后形成的转移概率图如下图所示

13、文本表示的目的

答:将文本中的文字转换为可以处理的数据类型

14、词袋模型的“袋”的含义

答:以术语频率作为权重，由于丢弃了词序，所以相当于把文本看作是词的堆积，即，多集(multiset)，又称袋(bag)

*15、词义的分布式表示的理论基础

答:(1)上下文相似的词，其语义也相似

(2)词的语义由其上下文决定

16、分布式表示方法的本质

答:分布式表示方法的本质要做的就是利用上下文信息把每一个词映射成一个维度固定的短向量

17、词向量定义

答:词向量(word*vector)是对词语义或含义的数值向量表示，包括字面意义和隐含意义,词向量可以捕捉到词的内涵，将这些含义结合起来构成一个稠密的浮点数向量，这个稠密向量支持查询和逻辑推理

18、以 CBOW 的训练过程为例，写出 word2vec 模型的训练过程

答:语料库词表规模·V，上下文长度·C，隐含层维数·N

(1)输入上下文各单词的 one-hot 向量(C 个 V 维向量)，其中词汇表维度为 V，上下文单词数量为 C.

(2)上下文中各词的 one-hot 编码分别乘以输入共享的权重矩阵 W*{V*N 维}(N 为隐藏神经元个数，也就是生成的词向量维度)，得到 C 个维度为 1*N 的向量。

(3)将 C 个 1*N 维的向量进行累加求和再平均，得到隐藏层向量(维度为 1*N)

(4)隐藏层向量乘以输出权重矩阵 W'{N*V}得到向量·{1*V};

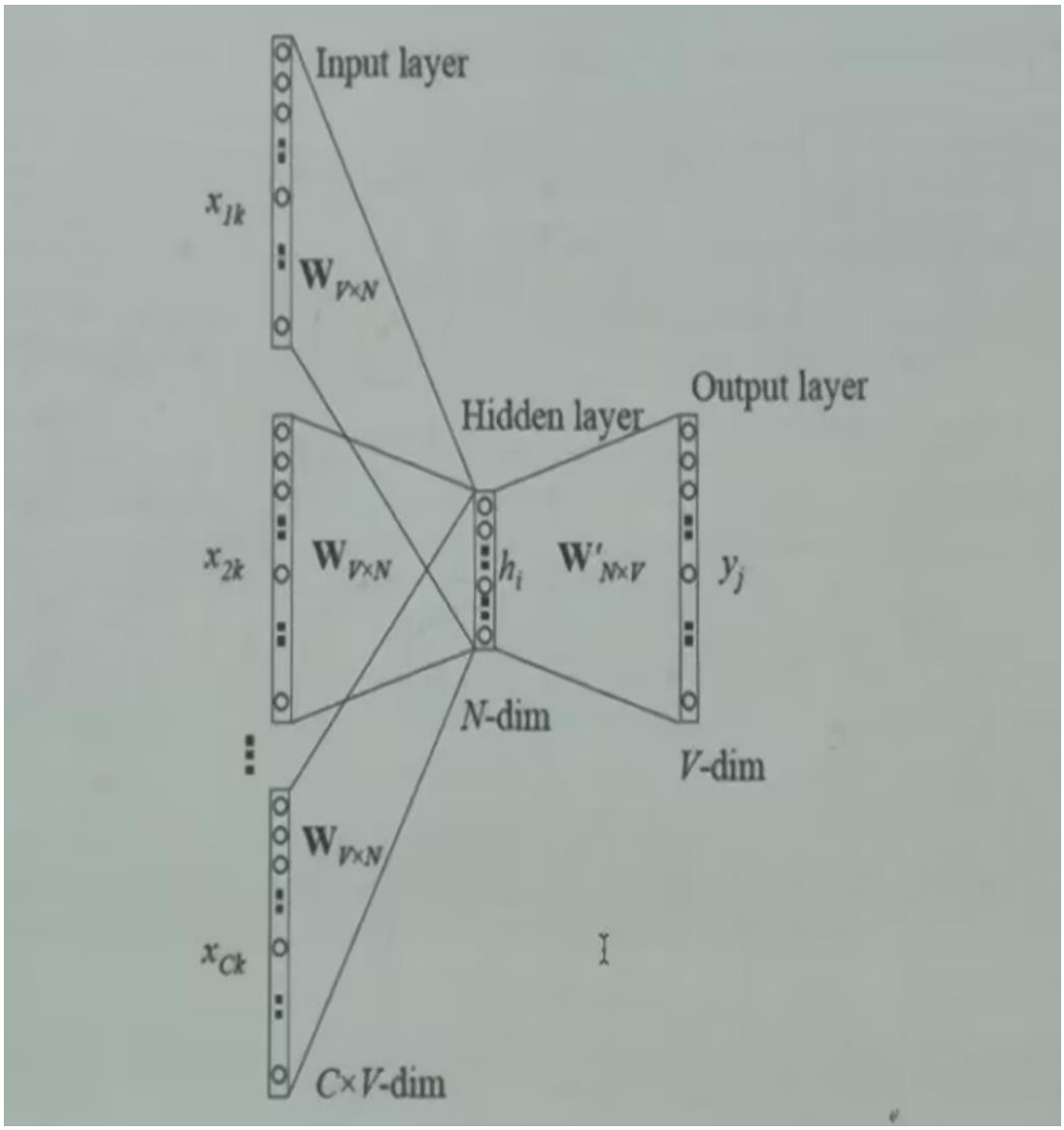
(5)该向量经 SofMax 激活函数处理，得到每个分量的概率分布(V 维)即预测词的 one-hot 表达式.

(6)概率最大的索引所指示的单词为预测出的中间词,计算生成的 one-hot 真实

标答的 one-hot 编码之间的误差(损失值)，误差越小越好.

(7)根据误差，采用梯度下降算法反向更新权重矩阵·W·和·W'的权值.

(8)模型训练完毕后，得到的输入权重矩阵 W 就是词向量、任何一个单词的 one-hot 表示乘以这个矩阵 W 都将得到自己的词向量



19、Transformer 中 Multi-Head 的作用

答:Multi-Head 的作用是去关注句子的不同方面，将一个词的词向量切分成 h

个维度，计算每个 h 维度的注意力权重，文本中的词被映射到高维空间中生成词向量，不同维度空间表示了词的不同方面，其可以学到不同的特征，相邻空间所学结果更相似，相较于全体空间放到一起对应更加合理

20、Bert 中 mask 策略的任务是什么

答:mask 策略的任务是给定一句话，随机抹去这句话中的一个或几个词,要求根据剩余词汇预测被抹去的几个词分别是什么

21、Bert 中 NSP 的任务是什么

答:NSP 的任务是判断句子 B 是否是句子 A 的下文,如果是的话,输出 IsNext

否则输出 NotNext

22、BERT 的特征

答:(1)BERT 是一个预训练模型，通过对大量无监督语料进行训练，为下游任务提供更准确表示语义、句法的信息，预训练的语言模型已被证明可有效改善许多自然语言处理任务

(2)BERT 模型很深，其 base 版本有 12 层，large 版本有 24 层

(3)BERT 采用的是双向语言模型，不同于之前普遍使用的单项模型，双向模型能更好的学习到前后文两侧的知识.

(4)BERT 是以 transformer 作为模型的内核进行特征抽取

23、关键词提取技术整体流程

答;第一步是获取文本的候选词,第二步则是对候选词进行打分，输出的关键词是

候选词中得分比较高的.

24、基于 TF-IDF 对给出的关键词进行排序.

假定某本书共有 50 万个词，其中“词向量”共出现 9800 次，“文本”出现 14000 次，“自然语言"出现了 17000 次，假设我们的语料库，中共有 1 万个文档，包含“词向量”的文档数为 347 个,包含“文本”的文档数为 621 个，包含"自然语言"的文档数为 440 个，计算这三个词的 TF-IDF 值

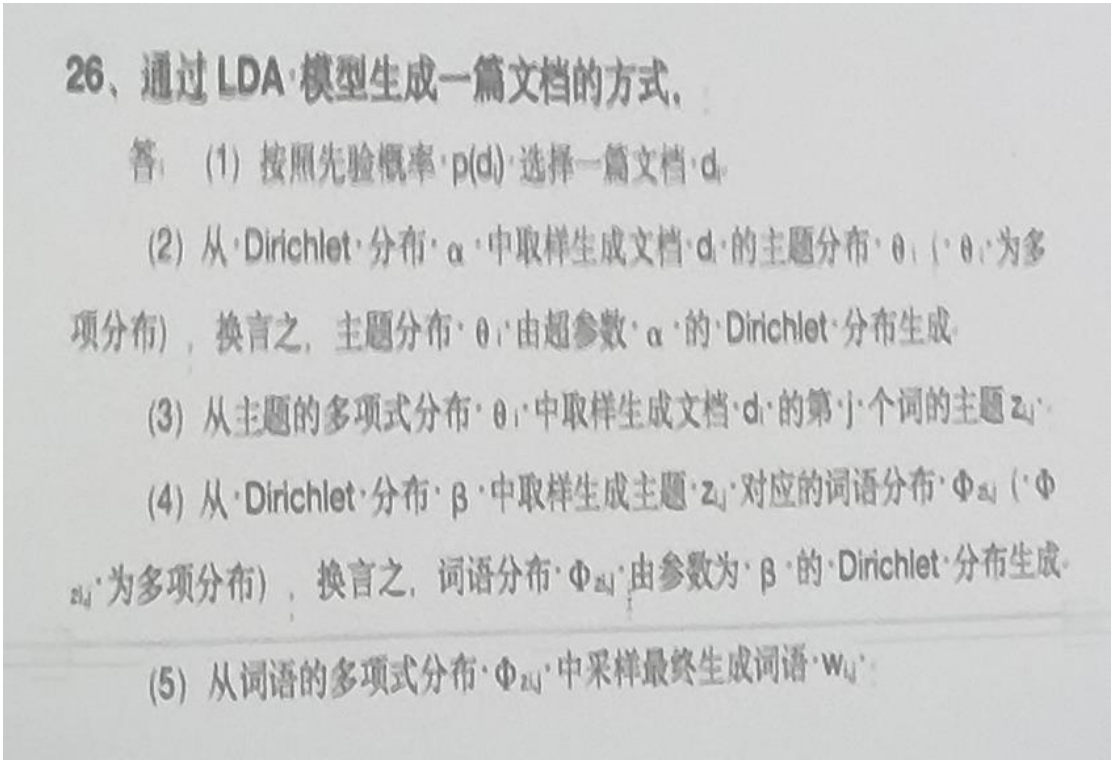
$$\text{词频 TF} = \frac{\text{单词出现的次数}}{\text{该文档的总单词数}}$$

$$\text{逆向文档频率 IDF} = \log \frac{\text{文档总数}}{\text{该单词出现的文档数} + 1}$$

TF-IDF=TF*IDF

25、主题模型的核心假设是什么,

答:存在隐含变量，即文本主题，决定子文本中词汇的出现



27、PLSA 与 LDA 的区别

答:在 PLSA 中，选主题和选词都是两个随机的过程，主题分布和词分布是唯一确定的，能明确的指出主题分布和词分布。

在 LDA 中，主题分布和词分布不再唯一确定不变，即无法确切给出。LDA 在 PLSA 的基础上给主题分布和词分布加了两个先验分布的参数，LDA 是先从众多的主题分布中选择出一种主题分布概率,之后过程就是 PLSA 主题生成过程,主题获得后，再是从众多的词分布中选择出一种词分布概率，之后再是 PLSA 的词生成过程。

28、正向最大匹配法的基本思想

答:(1)查找词典中的最长词的长度，设为 n

(2)由左向右(即从句首开始)的方向取长度为 n 的字符申，与词典进行匹配

(3)如匹配成功，则作为一个词

(4)若不成功，则将该字符串的最后面一个汉字去掉再匹配，循环重复,直至完成匹配

*29、利用正向最大区分算法对下列待切分串进行分词，写出分词过程(写到第二次取最长字串即可)及最终结果

待切分串 “海南一直是我向往的地方，椰子树、大海在我的心里是那样的神秘”

词典: 海南、/、、、一直、、/、、是、、/、、我、、、、、/、、向往、/、、的、、/、、地方、、向往的地方、、、/、、椰子树、、/、、、大海、、、/、、在、、/、、我的、、/、、心里、/、、那样的神秘

答:(1)由于是最大匹配，首先确定准备的词典的最大词长，词典的最大词长是5(个字)

(2)由于是正向最大匹配算法，则由左向右的方向取待处理字符串的前 5 个字串，即:“海南一直是”，用该子串去匹配词典中的词.

(3)匹配不成功，去掉最后一个字，得到较小子串“海南一直”，用该子串去匹配词典中的词.

(4)匹配仍不成功，去掉最后一个字，得到较小子串“海南一”，用该子串去匹配词典中的词

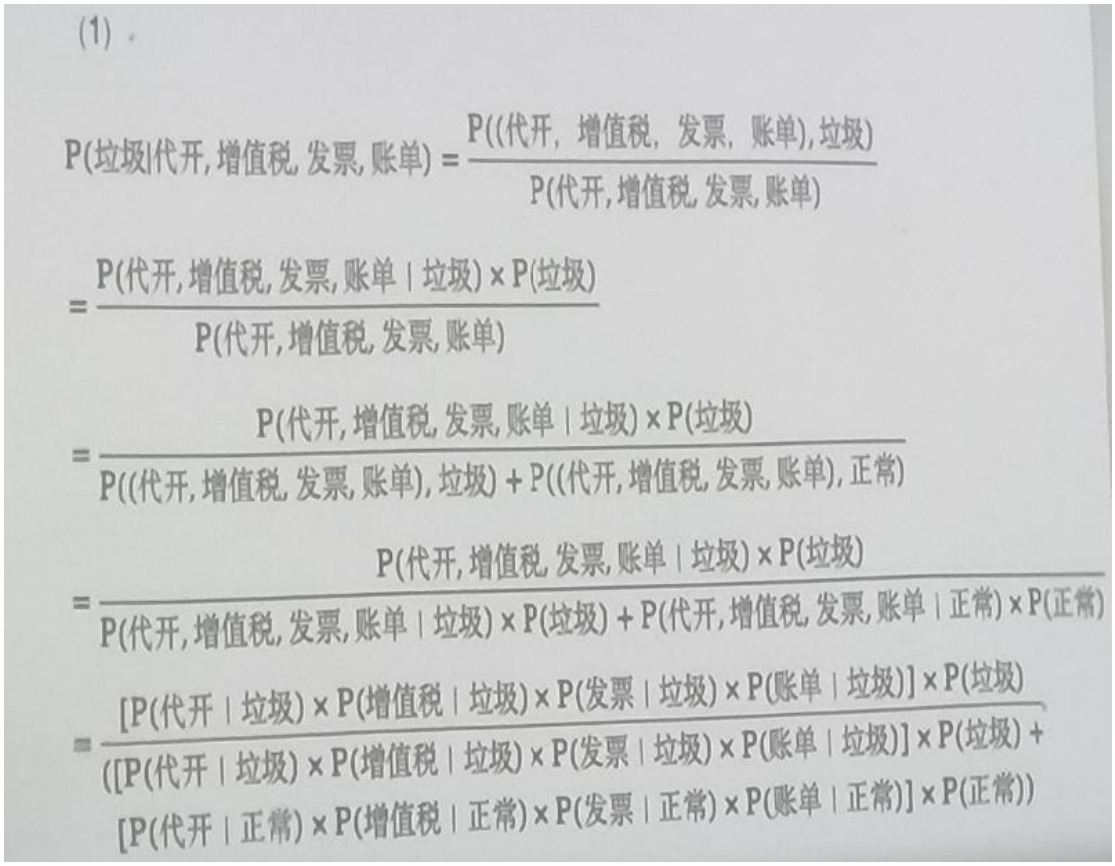
(5)匹配仍不成功，去掉最后一个字，得到较小字串“海南”，用该字串去匹配

词典中的词。

(6)匹配成功，则得到一个词“海南”

(7)继续取接下来的最长字串，即：“一直是我向”，循环该过程，得到“一直”，然后“是我向往的”,。分词结果：海南/一直/是/我/向往的地方/，椰子树/、大海/在/我的/心里/是/那样的神秘/。

30、你有一封邮件包含:代开，增值税，发票，账单，这几个词,要预测其为垃圾邮件的概率?



(1)

$$P(\text{垃圾}|\text{代开, 增值税, 发票, 账单}) = \frac{P((\text{代开, 增值税, 发票, 账单}), \text{垃圾})}{P(\text{代开, 增值税, 发票, 账单})}$$

$$= \frac{P(\text{代开, 增值税, 发票, 账单} \mid \text{垃圾}) \times P(\text{垃圾})}{P(\text{代开, 增值税, 发票, 账单})}$$

$$= \frac{P(\text{代开, 增值税, 发票, 账单} \mid \text{垃圾}) \times P(\text{垃圾})}{P((\text{代开, 增值税, 发票, 账单}), \text{垃圾}) + P((\text{代开, 增值税, 发票, 账单}), \text{正常})}$$

$$= \frac{P(\text{代开, 增值税, 发票, 账单} \mid \text{垃圾}) \times P(\text{垃圾})}{P(\text{代开, 增值税, 发票, 账单} \mid \text{垃圾}) \times P(\text{垃圾}) + P(\text{代开, 增值税, 发票, 账单} \mid \text{正常}) \times P(\text{正常})}$$

$$= \frac{[P(\text{代开} \mid \text{垃圾}) \times P(\text{增值税} \mid \text{垃圾}) \times P(\text{发票} \mid \text{垃圾}) \times P(\text{账单} \mid \text{垃圾})] \times P(\text{垃圾})}{([P(\text{代开} \mid \text{垃圾}) \times P(\text{增值税} \mid \text{垃圾}) \times P(\text{发票} \mid \text{垃圾}) \times P(\text{账单} \mid \text{垃圾})] \times P(\text{垃圾}) + [P(\text{代开} \mid \text{正常}) \times P(\text{增值税} \mid \text{正常}) \times P(\text{发票} \mid \text{正常}) \times P(\text{账单} \mid \text{正常})] \times P(\text{正常}))}$$

(2) 采用拉普拉斯变换计算得到：

P（代开|垃圾）=2/（5+1）=0.333

P（增值税|垃圾）=2/（5+1）=0.333

P（发票|垃圾）=5/（5+1）=0.833

P（账单|垃圾）=1/（5+1）=0.167

P（代开|正常）=1/（5+1）=0.167

P（增值税|正常）=1/（5+1）=0.167

P（发票|正常）=1/（5+1）=0.167

P（账单|正常）=2/（5+1）=0.333

(3)

P(垃圾|代开，增值税，发票，账单)

$$= \frac{[P(\text{代开}\mid\text{垃圾}) \times P(\text{增值税}\mid\text{垃圾}) \times P(\text{发票}\mid\text{垃圾}) \times P(\text{账单}\mid\text{垃圾})] \times P(\text{垃圾})}{([P(\text{代开}\mid\text{垃圾}) \times P(\text{增值税}\mid\text{垃圾}) \times P(\text{发票}\mid\text{垃圾}) \times P(\text{账单}\mid\text{垃圾})] \times P(\text{垃圾}) + [P(\text{代开}\mid\text{正常}) \times P(\text{增值税}\mid\text{正常}) \times P(\text{发票}\mid\text{正常}) \times P(\text{账单}\mid\text{正常})] \times P(\text{正常}))}$$

$$= \frac{0.333 \times 0.333 \times 0.833 \times 0.167 \times \frac{5}{10}}{0.333 \times 0.333 \times 0.833 \times 0.167 \times \frac{5}{10} + 0.167 \times 0.167 \times 0.167 \times 0.333 \times \frac{5}{10}}$$

$$= \frac{0.277389}{0.277389 + 0.027889} = 0.909$$

31、为什么 RNN 能够这么快在 NLP 流行并且占据

了主导地位呢?

答;主要原因还是因为 RNN 的结构天然适配解决 NLP 的问题

(1)在处理某个词的上下文关系时，不仅与该单词自身有关，还和该单词之前出现的单词有关，通过建立具有上下文关系的单词链，就可以把上下文信息包含在这个结构中，循环神经网络在同一隐藏层面的节点也建立连接，对于隐藏层的一个节点，其输入不仅包括上一层的输出，还包括上一时刻隐藏层的输出，通过这种模式，循环神经网络就可以实现对上一时刻信息的记忆，并把这样一个信息用于当前节点的输入进行计算

(2)NLP 的输入往往是个不定长的线性序列句子，而 RNN 本身结构就是个可以接纳不定长输入的由前向后进行信息线性传导的网络结构，而在 LSTM 引入三个门后，对于捕获长距离特征也是非常有效的，所以 RNN 特别适合 NLP 这种线形序列应用场景,这是 RNN 为何在 NLP 界如此流行的根本原因

32、文本数据是序列数据的原因

答:序列是一种相互依赖的(有限或无限)数据流，人类的自然语言，是符合某个逻辑或规则的字词拼凑排列起来的,文本数据在处理某个词的上下文关系时，不仅与该单词自身有关，还和该单词之前出现的单词有关，

33、情感分析中“情感”的理解

答:情感分析里所说的情感(sentiment)是指文本作者对文中所指的某个特定事件、物品、观点等目标的喜好倾向，通常用喜欢/赞同/支持(positive)、厌恶/反对/排斥 negative)和中立(neutral)三种状态表示

34、简述神经机器翻译模型采用“编码-解码”方式进行序列到序列转换时面临的问题，并介绍引入注意力机制的作用

答:一般的神经机器翻译模型采用“编码-解码”的方式进行序列到序列的转换，这种方式有两个问题:一是编码向量的容量瓶颈问题，即源语言所有的信息都需要保存在编码向量中，才能进行有效地解码;二是长距离依赖问题，即编码和解码过程中在长距离信息传递中的信息丢失问题

通过引入注意力机制，将源语言中每个位置的信息都保存下来,在解码过程中生成每一个目标语言的单词时,通过注意力机制直接从源语言的信息中选择相关的信息作为辅助，这样的方式就可以有效地解决上面的两个问题，一是无需让所有的源语言信息都通过编码向量进行传递,在解码的每一步都可以直接访问源语言的所有位置上的信息;二是源语言的信息可以直接传递到解码过程中的每一步，缩短了信息传递的距离