Senparc 盛派®

密级：公开

自然语言处理（NLP）简介

2021.01.27 · 苏州盛派网络科技有限公司

主持人/分享人：Kyle

什么是自然语言处理（NLP）

Senparc 盛派®

自然语言处理(Natural Language Processing, NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。

自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。因而它是计算机科学的一部分 。

自然语言处理主要应用于机器翻译、舆情监测、自动摘要、观点提取、文本分类、问题回答、文本语义对比、语音识别、中文OCR等方面 。

应用场景

Senparc 盛派®

一、词法分析

基于大数据和用户行为的分词后，对词性进行标注、命名实体识别，消除歧义 。
识别文本中具有特定意义的实体，主要包括：人名、地名、职位名、产品名词等。

Senparc 盛派®

二、文本分类

对文章按照内容类型（体育、教育、财经、社会、军事等等）进行自动分类，为文章聚类、文本内容分析等应用提供基础支持。

文章分类对文章内容进行深度分析，输出文章的主题一级分类、主题二级分类，在个性化推荐、文章聚合、文本内容分析等场景具有广泛的应用价值。

Senparc 盛派®

三、获取摘要

实现文本内容精简提炼，从长篇的文章中自动提取关键句和关键段落，构成摘要内容，进而生成指定长度的新闻摘要。

四、文本审核

判断一段文本内容是否符合网络发文规范，识别文本中是否包含违禁类型里面的关键字/词，能够实现自动化、智能化的文本审核，大幅节省内容审核的人力成本。

应用场景：

（1）用户信息审核

对网站的注册信息进行检测，过滤筛查用户提交注册的用户名或网名昵称，避免通过用户名的方式恶意推广。

（2）用户评论监控

对网站用户的评论信息检测，一旦发现用户提交恶意垃圾内容，可以做到文本的自动审核与过滤，保证产品良好用户体验

（3）文章内容审核

媒体文章的文本内容审核，自动识别文章中可能存在的推广、反动、色情信息，避免已发布文章的线上风险

文本是一种非结构化的数据信息，是不可以直接被计算的。
文本表示的作用就是将这些非结构化的信息转化为结构化的信息，这样就可以针对文本信息做计算，来完成我们日常所能见到的文本分类，情感判断等任务。

整数编码

假如我们要计算的文本中一共出现了4个词：猫、狗、牛、羊。向量里每一个位置都代表一个词。用一种数字来代表一个词：
猫：1
狗：2
牛：3
羊：4

整数编码的缺点如下：
1.无法表达词语之间的关系
2.对于模型解释而言，整数编码可能具有挑战性。

独热编码（one-hot representation）

4个词：猫、狗、牛、羊。向量里每一个位置都代表一个词。所以用 one-hot 来表示就是

猫：［1，0，0，0］
狗：［0，1，0，0］
牛：［0，0，1，0］
羊：［0，0，0，1］

one-hot 的缺点如下：
1.无法表达词语之间的关系
2.这种过于稀疏的向量，导致计算和存储的效率都不高

词嵌入(Word embedding)

概念上而言，它是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量。

word embedding 是文本表示的一类方法。跟 one-hot 编码和整数编码的目的一样，不过他有更多的优点。

词嵌入并不特指某个具体的算法，跟上面2种方式相比，这种方法有几个明显的优势：

1.他可以将文本通过一个低维向量来表达，不像 one-hot 那么长。

2.语意相似的词在向量空间上也会比较相近。

3.通用性很强，可以用在不同的任务中。

Senparc 盛派®

Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在word2vec中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

这种算法有2种训练模式：

1.通过上下文来预测当前词（CBOW）

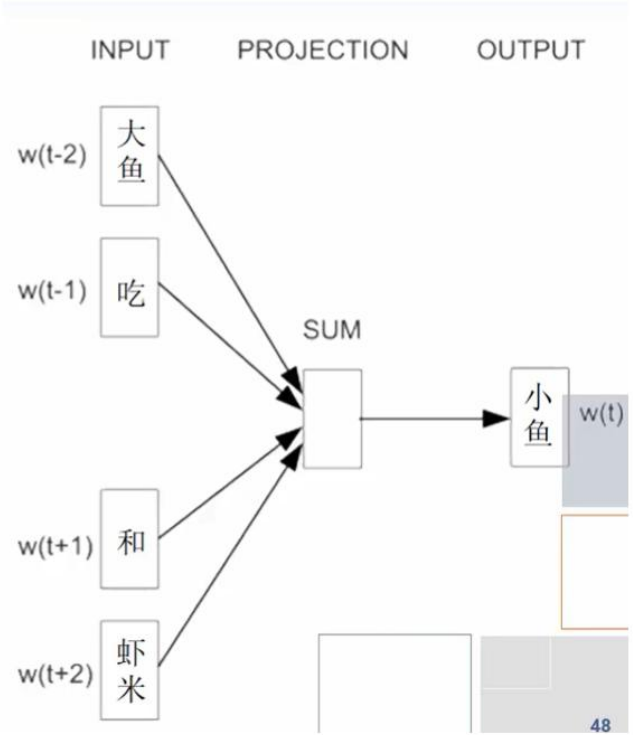
2.通过当前词来预测上下文(skip-gram)

优点：

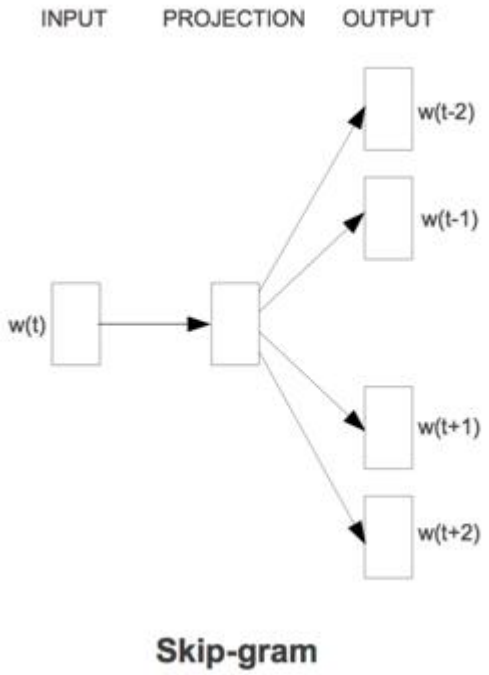
1.由于 Word2vec 会考虑上下文，跟之前的 Embedding 方法相比，效果要更好比之前的 Embedding方 法维度更少，所以速度更快

2.通用性很强，可以用在各种 NLP 任务中

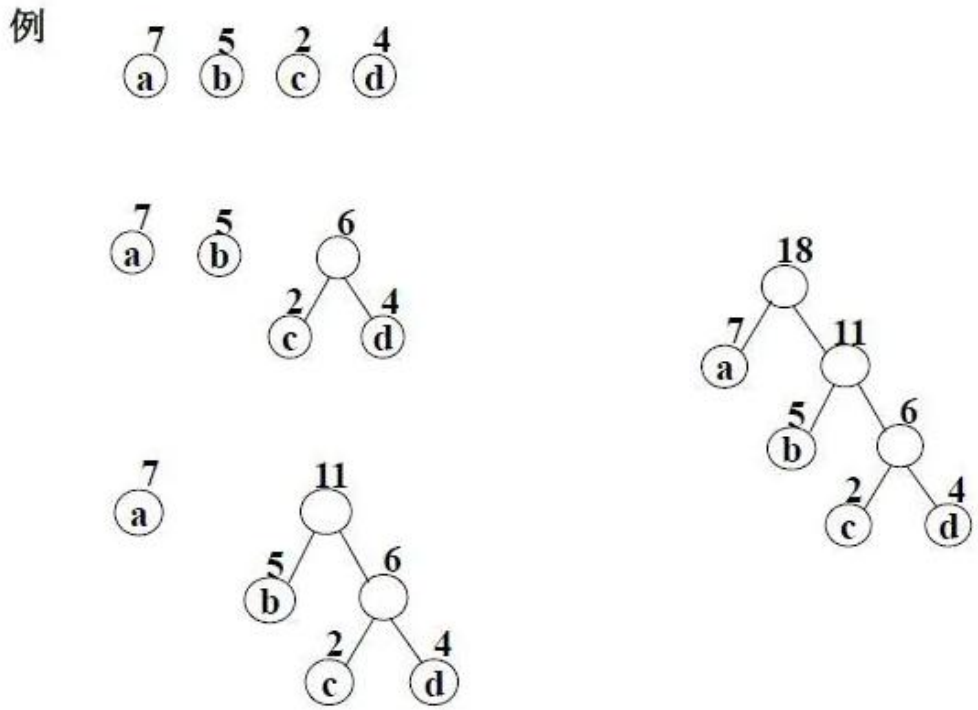
CBOW模型的训练输入是某一个特征词的上下文相关的词对应的词向量，而输出就是这特定的一个词的词向量。比如下面这段话，我们的上下文大小取值为2，特定的这个词是"小鱼"，也就是我们需要的输出词向量,上下文对应的词有4个，前后各2个，这4个词是我们模型的输入。由于CBOW使用的是词袋模型，因此这4个词都是平等的，也就是不考虑他们和我们关注的词之间的距离大小，只要在我们上下文之内即可。



Skip-Gram模型和CBOW的思路是反着来的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量。还是上面的例子，我们的上下文大小取值为2， 特定的这个词"小鱼"是我们的输入，而这4个上下文词是我们的输出。



词频越高的词，希望编码长度越短。
在huffman树中，叶子节点是各个词，叶子节点的权重是词频。每个词都有权重×距离根节点的长度，即词频*编码长度。Huffman树保证了上述所有词的上述乘积的和是最小的，即该编码方式的总代价是最小的。



霍夫曼编码示例

实战

Senparc 盛派®

谢谢!

Kyle
E-mail: qwu@senparc.com