1. Representation语言表示
2. 如何让语言表示成为神经网络可以处理的数据类型？

语音中，用音频频谱向量所构成的矩阵作为输入，输入到NN

图像中，用图片的像素构成矩阵展平成向量序列后，输入到NN

NLP中，将每个词用一个向量表示出来

1. 词的表示方法
2. One-hot representation 词的独热表示

是NLP中最直观最常用的表示方法

Ex:”话筒“表示为[0 0 0 0 1 0 0 0 0 0 0 0 …]

“麦克风“表示为[0 0 0 0 0 0 0 0 0 1 0 0 0 …]

每个词都是茫茫词海中的一个，每个词都有一个唯一的数字ID

存储：稀疏矩阵。很简洁

缺点：向量的维度会随着句子的词的数量类型增大而增大

致命的是，任何两个词之间都是孤立的，无法表现出在语义层面上的词语关系

1. Distributed representation 词的分布式表示

Harris在1954年提出distributed hypothesis: 上下文相似的词，其语义也相似

Firth在1957年进一步提出： a word is characterized by the company it keeps

核心思想in 词表示法：①选择一种方式描述上下文 ②选择一种模型刻画某个词与其上下文之间的关系

几种词表示方法：

* 基于矩阵的分布表示

每行对应一个词，每列表示一种不同的上下文，矩阵中的每个元素对应相关词和上下文的共现次数。

* 基于聚类的分布表示

根据两个词的公共类别判断这两个词的语义相似度。

最经典的是布朗聚类：

布朗聚类是一种自底向上的层次聚类算法，基于n-gram模型和马尔科夫链。

① n-gram模型： 是一种基于统计语言模型的算法。将文本里面的内容按照字节进行大小为N的滑动窗口操作，形成了长度是N的字节片段序列。

每个字节片段成为gram，对所有gram的出现频度进行统计，并按照事先设定好的阈值进行过滤，形成关键gram列表——即这个文本的向量特征空间。

整个句子的概率就是各个词出现的概率的乘积。

应用：搜索引擎/输入法的猜想、提示。

② 布朗聚类的输入是一个词序列构成的语料库，输出一个二叉树。树的叶子节点是一个个词，中间节点是我们想要的类。

* 基于神经网络的分布表示

也叫word embedding, distributed representation

通过神经网络技术对上下文、以及上下文与目标词之间的关系建模。

优势：可以表示复杂的上下文

1. 语言模型

语言模型包括文法语言模型和统计语言模型，一般指后者。

统计语言模型：计算一个句子的概率的大小的模型。概率越高，则这个句子是更合乎人说出来的自然句子。

1. 词嵌入（基于神经网络的分布表示）

基于分布假说：上下文相似的词，语义也相似

核心：上下文的表示；目标词之间的关系的建模

为什么叫“嵌入”：在one-hot基础上的改进：

将vector每一个元素由整形变成浮点型

将原来稀疏的巨大矩阵嵌入到一个更小维度的空间

1. Word2vec:是语言模型 ，是Continuous Bagof-words 和Skip-gram的实现工具

统计语言模型：Given几个词，在这几个词连续出现的情况下计算某个词出现的概率

Continuous Bagof-words: 属于统计语言模型，根据某个词前后的n个词，来计算某个词出现的概率

Skip-Gram model: 根据某个词，分别计算它前后出现某几个词的概率。

1. 训练方法：

分为无/弱监督的预训练；word2vec,auto

端对端的有监督训练

1. 做具体NLP任务时，建议自己训练自己的词向量；使用别人训练好的词向量的坑比较多。