清华大学深圳研究生院

计算智能实验室 周报

|  |  |  |
| --- | --- | --- |
| **姓名** | **研究方向** | **报告覆盖时间** |
| 李晨辉 | NLP（法律文本分类） | 2010.3.4.26-2019.3.10 |
| **本周主要完成的工作** | | |
| 1. 深入理解word2vec； 2. LSTM等RNN模型的学习； | | |
|  | | |
| **一、深入理解word2vec** | | |
| word2vec是google在2013年推出的一个NLP工具，它的特点是将所有的词向量化，这样词与词之间就可以定量的去度量他们之间的关系，挖掘词之间的联系。在word2vec的实现中，有两种不同实现方式：   * 基于Hierarchical Softmax的模型   从隐藏层到输出的softmax层进行了改进。为了避免要计算所有词的softmax概率，采样了霍夫曼树来代替从隐藏层到输出softmax层的映射。  CBOW模型  **输入：**基于CBOW的语料训练样本，词向量的维度大小M，CBOW的上下文大小2c,步长η  **输出：**霍夫曼树的内部节点模型参数θ，所有的词向量w  对于训练样本中的每一个词，其前面的c个词和后面的c个词作为了CBOW模型的输入,该词本身作为样本的输出，期望softmax概率最大。  **步骤**  　1. 基于语料训练样本建立霍夫曼树。  2. 随机初始化所有的模型参数θ，所有的词向量w  3. 进行梯度上升迭代过程，对于训练集中的每一个样本(context(w),w)做如下处理：  a) e=0， 计算  Lw为从根节点到w所在叶子节点包含的的节点总数    c) 对于context(w)中的每一个词向量xi(共2c个)进行更新：xi=xi+e  d) 如果梯度收敛，则结束梯度迭代，否则回到步骤3继续迭代  Skip-Gram模型  输入：基于Skip-Gram的语料训练样本，词向量的维度大小M，Skip-Gram的上下文大小2c,步长η  输出：霍夫曼树的内部节点模型参数θ，所有的词向量w  对于训练样本中的每一个词，该词本身作为样本的输入， 其前面的c个词和后面的c个词作为了Skip-Gram模型的输出,，期望这些词的softmax概率比其他的词大。  步骤  1. 基于语料训练样本建立霍夫曼树。  2. 随机初始化所有的模型参数θ，所有的词向量w  3. 进行梯度上升迭代过程，对于训练集中的每一个样本(w,context(w))做如下处理：    b)如果梯度收敛，则结束梯度迭代，算法结束，否则回到步骤a继续迭代。   * 基于Negative Sampling的模型   输入：one-hot的词表示，输出：对应词的得分，优化目标是使得正样例的得分接近于1，负样例的得分接近于0.  优化完成后得到的隐含层的词表示就是我们想要的词向量  输入层到隐藏层是词向量平均值，隐藏层到输出层为二元逻辑回归  CBOW模型  输入：基于CBOW的语料训练样本，词向量的维度大小Mcount，CBOW的上下文大小2c,步长η, 负采样的个数neg  输出：词汇表每个词对应的模型参数θ，所有的词向量xw  步骤  1. 随机初始化所有的模型参数θ，所有的词向量w  2. 对于每个训练样本(context(w0),w0),负采样出neg个负例中心词wi,i=1,2,...neg  3. 进行梯度上升迭代过程，对于训练集中的每一个样本(context(w0),w0,w1,...wneg)做如下处理：  Skip-Gram模型  输入：基于Skip-Gram的语料训练样本，词向量的维度大小Mcount，Skip-Gram的上下文大小2c,步长η , 负采样的个数neg。  输出：词汇表每个词对应的模型参数θ，所有的词向量xw  步骤  1. 随机初始化所有的模型参数θ，所有的词向量w  2. 对于每个训练样本(context(w0),w0),负采样出neg个负例中心词wi,i=1,2,...neg  3. 进行梯度上升迭代过程，对于训练集中的每一个样本(context(w0),w0,w1,...wneg)做如下处理： | | |
| **二、LSTM网络学习** | | |
| * 最简单simple RNN引入：   D:\桌面文件\研究生\周报\SRNN.png   * LSTM (Long Short-Term Memory )   D:\桌面文件\研究生\周报\LSTM.png | | |
| * GRU（Gated Recurrent Unit）   D:\桌面文件\研究生\周报\GRU.png | | |
| **下周工作主要安排**   1. 统计自然语言处理方法的学习； 2. 继续学习一些其他的数据结构； | | |