**GloVe模型原理和实现方法**

目录

[一、 实验简介 1](#_Toc43380055)

[二、 预备知识 2](#_Toc43380056)

[1. Glove词向量概念和作用 2](#_Toc43380057)

[2. GloVe模型原理和实现步骤 2](#_Toc43380058)

[3. GloVe模型训练过程及可调参数列表 3](#_Toc43380059)

[4. GloVe模型的数学推导 4](#_Toc43380060)

[三、 实验环境和两种实现方式 5](#_Toc43380061)

[1. 实验环境 5](#_Toc43380062)

[2. 实现方式 5](#_Toc43380063)

[四、 实验结果 5](#_Toc43380064)

[五、 GloVe与Word2Vec的优劣势分析 6](#_Toc43380065)

[六、 参考资料和数据集 6](#_Toc43380066)

[八、 小结 6](#_Toc43380067)

[练习 6](#_Toc43380068)

# 实验简介

➢ 实验所属系列: 人工智能大数据

➢ 实验对象: 本科/专科计算机专业

➢相关课程及专业: 自然语言处理

➢ 实验时数(学分):2学时

➢ 实验类别: 实践实验类

# 预备知识

## Glove词向量概念和作用

Glove: Global Vectors for Word Representation，全局词向量，一个基于全局词频统计的词表征工具，是一个将词表示为含有语义信息的向量的模型。该模型使用了词与词之间的共现(co-occurrence)概率，可以捕获单词之间的语义特性，如相似性、类比性等，并通过对向量的运算，计算两个单词之间的语义相似性。

## GloVe模型原理和实现步骤

1. 模型整体结构
2. 根据语料库构建一个共现矩阵（Co-occurrence Matrix）X，其中元素表示词j与词i在特定大小的上下文窗口(context window)的次数。一般次数为1，Glove根据两个单词在上下文窗口的距离d提出了一个衰减函数：decay=1/d用于计算权重，距离越远的单词所占的总计数(total count)的权重越小。

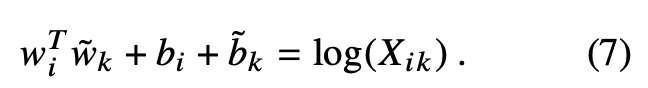
如context window size=1,则“处理”这个词环境中的词为“语言”和“包含”。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **自然** | **语言** | **处理** | **包含** | **很多** | **任务** |
| **任务** | 非常 | 有趣 |  |  |  |

则context window size=1条件下，共现矩阵如下：

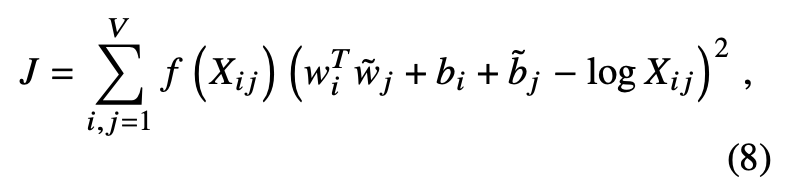
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 自然 | 语言 | 处理 | 包含 | 很多 | 任务 | 非常 | 有趣 |
| 自然 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 语言 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 处理 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 包含 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 很多 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 任务 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 非常 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 有趣 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

1. 构建词向量(word vector)和共现矩阵(co-occurrence)之间的近似关系

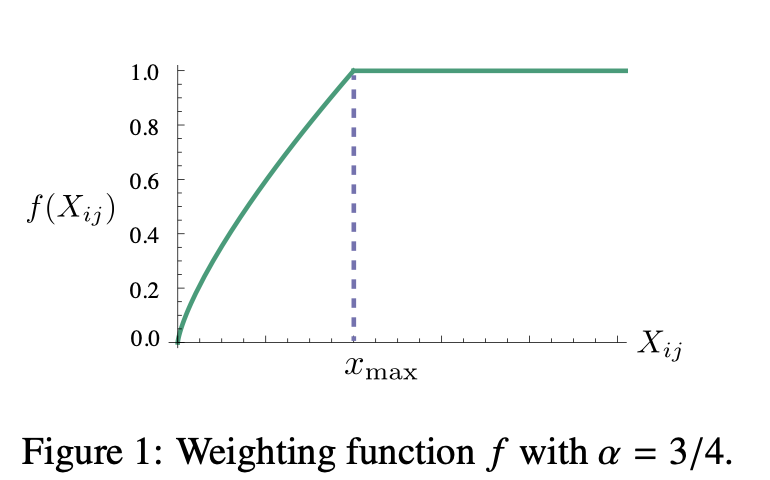


其中前两项w为待求解的词向量，b为两个词向量的偏置项。

1. 构造损失函数：



在均方损失函数基础上增加了一个权重函数f(Xij)，用于处理一个语料库中经常一起出现的单词（frequent occurrence），这些单词的权重是有上限的，该函数是非递减的，若两单词没有同时出现，则Xij=0,f(0)=0



论文作者采用了一个分段函数来描述这种单词的权重，实验中α的取值都是0.75，xmax的取值都是1000.

## GloVe模型训练过程及可调参数列表

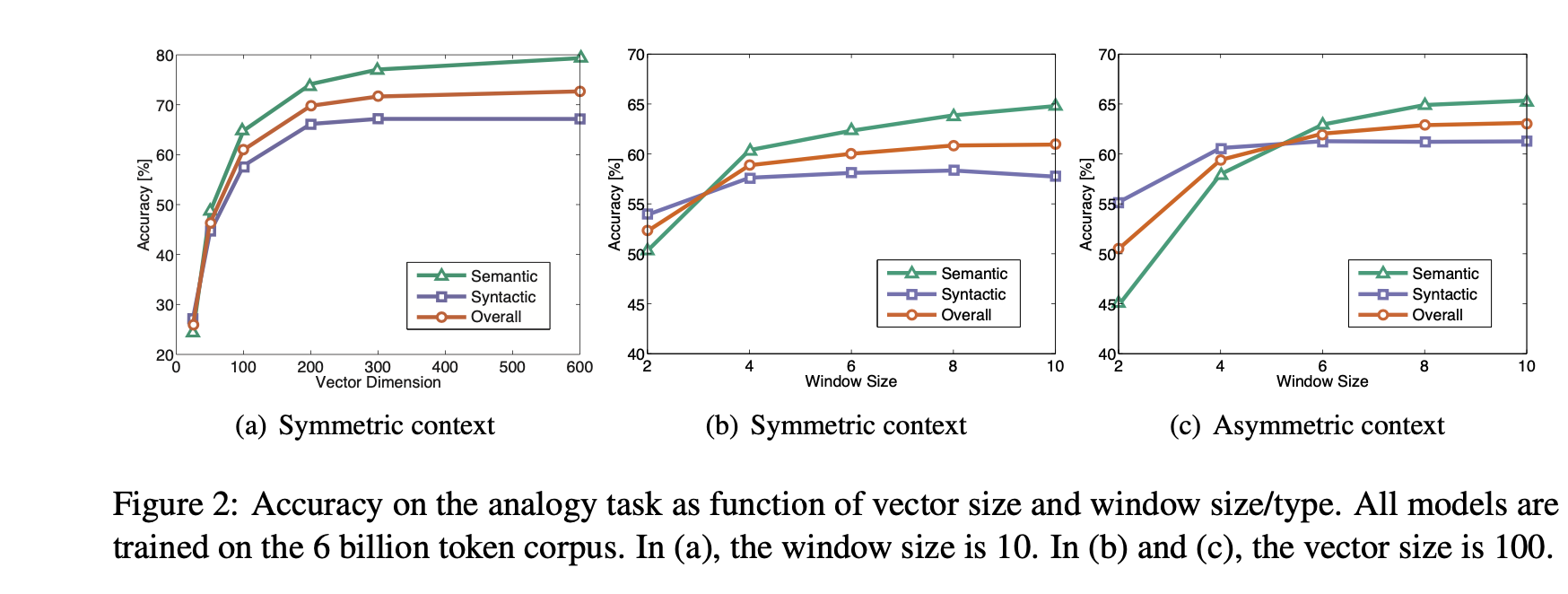
1. **模型训练**

训练过程：采用AdaGrad的梯度下降算法，对矩阵X中的所有非零元素进行随机采样，学习率设为0.05，在vector size小于300的情况下迭代50词，大于300时迭代100次，直到收敛。

训练结果 ：得到两个对称的词向量和，最终选择作为最终词向量。初始化值不同相当于加了不同的随机噪声提高结果鲁棒性。

衡量指标：语义准确度、语法准确度以及总体准确度。

结果分析：向量维度300左右最佳，context Window size大致6-10之间。



1. **可调参数列表**

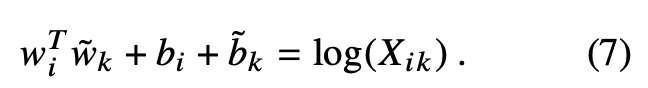
Learnin rate：学习率

Vector size:向量维度

Batch\_size: 每次小批量训练的同时进行的次数。默认值为512。

## GloVe模型的数学推导

公式(7)和公式(8)的推导



表示单词k在单词i的上下文中出现的次数；

表示单词i的上下文中所有单词出现的总次数，即；

，即单词j出现在单词i的上下文中的概率

具体来说，对于语料库中的一个句子：

**自然语言处理包含很多任务，任务非常有趣 。**

每个词在语料库中出现的概率如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probability and Ratio | k=处理 | K=包含 | K=很多 | K=任务 |
| P(k|自然) | 1.9×10-4 | 6.6×10-5 | 3×10-3 | 1.7×10-5 |
| P(k|语言) | 2.2×10-5 | 7.8×10-4 | 2.2×10-3 | 1.8×10-5 |
| P(k|自然)/P(k|语言) | 8.9 | 8.5×10-2 | 1.36 | 0.96 |

共现概率的比值越不靠近1，则表示其中一个共现概率与词i关系越大，越靠近1，则表示关系越小。用一个函数来表达共现概率的比值：

依赖三个词向量，给定i出现k，给定j出现k，前两个向量i和j类似于中心词，最后一个k类似于背景词。

上述f函数不是唯一的，以下展示一种设计思路：两个中心词i和j做减法再做点积，使得自变量为标量，

由于上式左侧是向量，右侧是标量，为了避免线性结构模糊，我们首先取参数的点积：

约束条件：1)X矩阵为对称矩阵，因为词j出现在i中，那么i也出现在j中；2)中心词和背景词共同出现，对调也不影响结果。

故：

(1)(2)式联立，推导出

替换掉，用偏置项表达共现概率：

构建并最小化平方差损失函数：

为权重函数，建议，当x<c时，。损失函数计算复杂度与共现词频矩阵X中非零元素的数目呈线性关系。

# 实验环境和两种实现方式

## 实验环境

Linux 虚拟机

Python3、jupyter notebook

## 实现方式

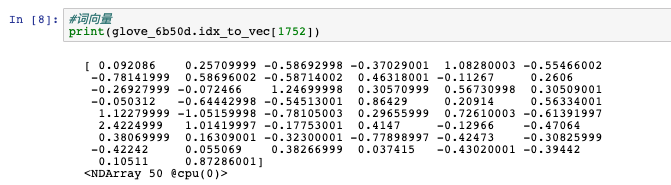
<https://github.com/xjtuerz0/glove_pre-train_tensorflow>

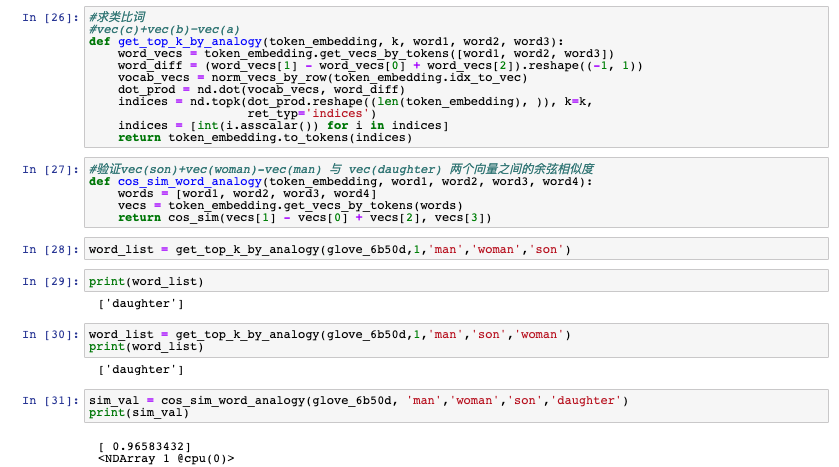
其中，[1-glove-pre-train-model.ipynb](https://github.com/xjtuerz0/glove_pre-train_tensorflow/blob/master/1-glove-pre-train-model.ipynb" \o "1-glove-pre-train-model.ipynb)是直接对100维的词向量进行相似度计算的效果演示；2-tensorflow-glove是基于TensorFlow的GloVe模型实现，为方便演示，本文只使用了只有672K的RC\_2005-12数据集。

# 实验结果

衡量指标：余弦相似度、类比词相似度

结果：





# GloVe与Word2Vec的优劣势分析

1. GloVe在进行矩阵分解时是基于全局信息进行计算的，Word2Vec是基于局部语料进行预测的；
2. GloVe和Word2Vec都是静态词向量，无法解决多义词的问题；
3. GloVe的词向量加上一个任意常数向量得到的值依然是GloVe损失函数的解。这可能会导致最终的词向量失去某些含义。

# 参考资料和数据集

参考资料：GloVe: Global Vectors for Word Representation

官方文档和代码：<https://github.com/stanfordnlp/GloVe>

数据集来源：<https://files.pushshift.io/reddit/comments/RC_2015-01.bz2>

# 八、 小结

Glove是一个基于全局词频统计的词表征工具，通过词向量与共现矩阵之间的距离来衡量词与词之间相似度，通过AdaGrad来训练损失函数，最终得到维度在300左右的词向量。

# 练习

1. Glove模型对同一窗口中距离较远的单词做了什么处理？

【A】增加了一个衰减函数，decay=1/d，距离越远的单词所占总计数的权重越小

【B】将他们丢弃

【C】随机初始化远距离单词的权重

【D】不作任何处理

答案：A

1. Glove模型是如何训练的？

【A】通过神经网络反向传播进行训练

【B】采用AdaGrad梯度下降算法进行迭代训练

【C】采用最大期望算法进行计算

【D】采用前向-后向算法进行计算

答案：B

1. 模型对最终结果和做了什么处理？有什么好处？

【A】直接将和进行拼接

【B】取为最终结果，如此可提高结果的鲁棒性

【C】对和取加权平均，衡量不同方向词的重要性

【D】评估和对模型的重要性，选择重要性高的模型

答案：B