

**基于 SVD 分解以及基于 SGNS**

**两种方法构建汉语词向量**

**自然语言处理课程实验报告**

班级：110

姓名：耿翊中

学号：2021213382

2024年4月

目录

[基于 SVD 分解以及基于 SGNS 1](#_Toc18053)

[两种方法构建汉语词向量 1](#_Toc2024)

[自然语言处理课程实验报告 1](#_Toc2904)

[1. 实验要求 3](#_Toc31403)

[2. SVD分解方法 4](#_Toc24638)

[2.1数据预处理 4](#_Toc3911)

[2.3应用svd分解得到词向量 6](#_Toc2050)

[2.4计算余弦相似度 7](#_Toc6971)

[2.5类结构化 9](#_Toc31872)

[2.6模型参数调整与计算 10](#_Toc3670)

[3. SGNS方法 11](#_Toc25753)

[3.1构建skip-gram 11](#_Toc30831)

[3.2实现SGNS模型 11](#_Toc2878)

[3.3训练SGNS模型 11](#_Toc28999)

[3.4参数调整 11](#_Toc1918)

# 实验要求

作业内容：

一、概述：分别基于 SVD 分解以及基于 SGNS 两种方法构建汉语词向量并进行评测。

二、具体说明：

1、语料：training.txt。

2、基于 SVD 分解的方法：获取高维 distributional 表示时 K=5，SVD 降维后的维数自定，获 得子词向量 vec\_sta。之后基于该向量计算 pku\_sim\_test.txt 中同一行中两个子词的余弦相似 度 sim\_svd。当 pku\_sim\_test.txt 中某一个词没有获得向量时(该词未出现在该语料中)，令其

所在行的两个词之间的 sim\_svd=0。

3、基于 SGNS 的方法：SGNS 方法中窗口 K=2，子词向量维数自定，获得向量 vec\_sgns。之后基于该子词向量计算 pku\_sim\_test.txt 中同一行中两个词的余弦相似度 sim\_sgns。当 pku\_sim\_test.txt 中某一个词没有获得向量时(该词未出现在该语料中)，令其所在行的两个词之间的 sim\_sgns=0。

4、两种方法的结果输出要求(因为是机器判定，请一定按如下格式输出)：

4.1 保持 pku\_sim\_test.txt 编码(utf-8)不变，保持原文行序不变

4.2 每行在行末加一个 tab 符之后写入该行两个词的 sim\_sv，再加一个 tab 符之后写入该行

两个词的 sim\_sgns。

4.3 输出文件命名方式：学号。

5、所有输出文本均采用 Unicode(UTF-8)编码

6、算法采用 Python (3.0 以上版本) 实现

# SVD分解方法

## 2.1数据预处理

从training.txt中读取句子数据，并转化为列表；统计得到词表。为后续的文本分析（如构建共现矩阵和应用SVD分解）创建基础。

由于在统计奇异值具体信息时，不能使用稀疏矩阵来进行分解，而完整的共现矩阵太大，所以需要设定低频词过滤，得到过滤后较小的词表。

1. **def** preprocess\_file(self):
2. """从文件中读取文本并预处理：统计词频"""
3. **print**("开始预处理文件...")
4. word\_frequency = {}  # 新增：用于统计词频的字典
5. processed\_sentences = []
6. with open(self.train\_path, 'r', encoding='utf-8') as f:
7. **for** line **in** f:
8. words = line.strip().split()
9. filtered\_words = []
10. **for** word **in** words:
11. **if** word.strip() != '':
12. filtered\_words.append(word)
13. **if** word **in** word\_frequency:
14. word\_frequency[word] += 1
15. **else**:
16. word\_frequency[word] = 1
17. processed\_sentences.append(" ".join(filtered\_words))
18. # 移除低频词
19. min\_frequency = 2  # 设定最低频率阈值
20. filtered\_vocab = {word **for** word, freq **in** word\_frequency.items() **if** freq >= min\_frequency}
21. **return** processed\_sentences, filtered\_vocab

具体过程解释：

初始化词频字典: word\_frequency = {} 创建了一个空字典，用于存储每个词及其对应的出现频次。

初始化处理后的句子列表: processed\_sentences = [] 创建了一个空列表，用于存储处理后的句子。

读取训练文件: 使用 with open(self.train\_path, 'r', encoding='utf-8') 来打开指定路径的训练文件，这里使用了 'utf-8' 编码以支持多种语言文本。

逐行处理文本:

分词: words = line.strip().split() 将每一行文本首先去除首尾空白（strip()），然后按空格分割成单词列表。

过滤空字符串: 通过 if word.strip() != '' 检查每个词是否为空字符串（仅包含空白的字符串），确保只处理有实际内容的词。

统计词频: 如果词非空，将其添加到 filtered\_words 列表中，并在 word\_frequency 字典中更新该词的频次。如果该词已存在于字典中，频次加一；否则，将该词及其频次设为1。

保存处理后的句子: 使用 " ".join(filtered\_words) 将过滤后的词重新组合成句子，然后添加到 processed\_sentences 列表中。

移除低频词:

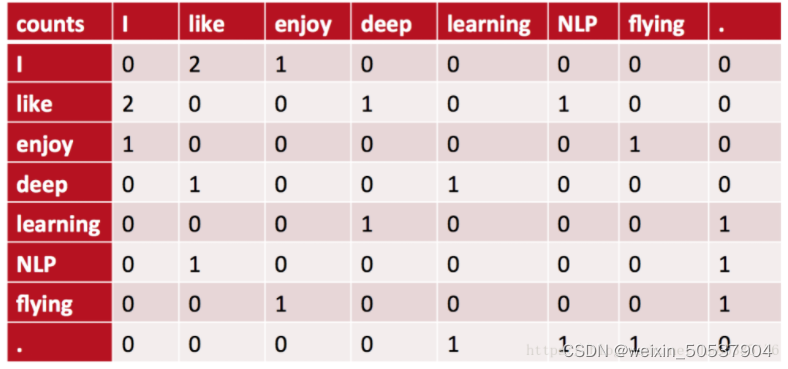
设定最低频率阈值 min\_frequency = 20，意味着只有出现频次达到或超过这个阈值的词才会被保留。

filtered\_vocab = {word for word, freq in word\_frequency.items() if freq >= min\_frequency} 通过列表推导式创建一个新的集合，包含所有出现频次达到或超过20次的词。

返回结果: 方法返回两个值：processed\_sentences 和 filtered\_vocab。前者是一个经过预处理（去除空字符串并统计词频）的句子列表，后者是一个集合，包含了所有满足频次要求的词。

**2.2构建词共现矩阵**

共现矩阵是自然语言处理（NLP）中的一种重要数据结构，用于表示词汇项在特定上下文中同时出现的频率。其中的行和列代表语料库中的词汇，而每个元素的值表示对应行和列的词汇项在预定的窗口大小内共同出现的次数。共现矩阵捕获了词汇之间的关系，这些关系将用于后续的相似度分析中。



1. **def** build\_cooccurrence\_matrix(self, sentences, filtered\_vocab):
2. """基于过滤后的词汇表构建共现矩阵"""
3. **print**("构建共现矩阵...")
4. vocab\_index = {word: i **for** i, word **in** enumerate(filtered\_vocab)}
5. cooccurrence\_matrix = np.zeros((len(filtered\_vocab), len(filtered\_vocab)), dtype=np.float32)
6. **for** sentence **in** sentences:
7. words = sentence.split()
8. **for** i, word **in** enumerate(words):
9. **if** word **in** vocab\_index:  # 确保单词在过滤后的词汇表中
10. target\_word\_index = vocab\_index[word]
11. start = max(0, i - self.window\_size)
12. end = min(len(words), i + self.window\_size + 1)
13. **for** j **in** range(start, end):
14. **if** i != j **and** words[j] **in** vocab\_index:  # 同样确保上下文词在词汇表中
15. context\_word\_index = vocab\_index[words[j]]
16. cooccurrence\_matrix[target\_word\_index, context\_word\_index] += 1
17. **print**("共现矩阵构建完成。")
18. **return** cooccurrence\_matrix, vocab\_index

具体过程解释：

初始化词汇索引: 通过 {word: i for i, word in enumerate(filtered\_vocab)} 创建一个字典，为每个过滤后的词汇分配一个唯一的索引。这为构建矩阵提供了一个映射，将词汇映射到矩阵的行和列上。

创建空的共现矩阵: 使用 np.zeros((len(filtered\_vocab), len(filtered\_vocab)), dtype=np.float32) 初始化一个方阵，其维度等于过滤后的词汇表的大小。这个矩阵用于记录词汇之间的共现频次。

填充共现矩阵:

遍历每个句子中的每个词。

使用一个窗口（由 self.window\_size 定义）来确定每个目标词周围考虑的上下文范围。

对于窗口内的每个上下文词，如果它也在过滤后的词汇表中，则在共现矩阵的相应位置（即目标词和上下文词的索引处）增加1，表示它们在给定窗口大小内共同出现。

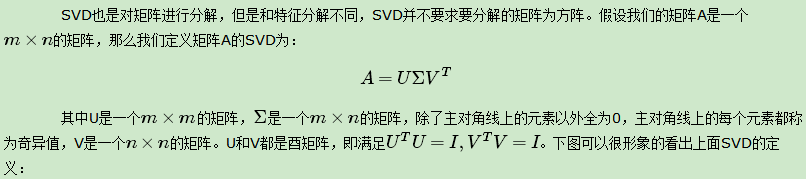
完成提示与返回: 在构建完成后打印提示，并返回构建好的共现矩阵及词汇索引。

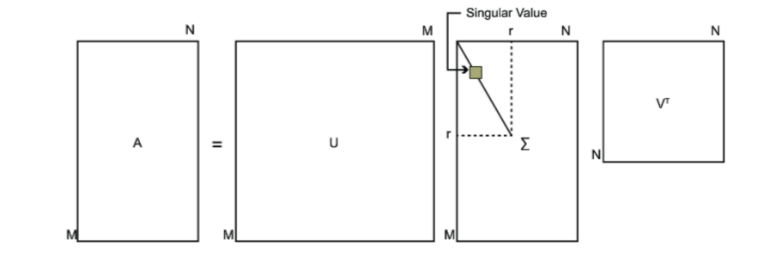
## **2.3应用svd分解得到词向量**

1. 为什么要进行svd分解

共现矩阵通过统计词汇之间的共现频次，能够捕捉到词汇之间的语义关系。通常，语义上相关或功能上相似的词在文本中更频繁地共同出现。但共现矩阵通常是高维的，并且稀疏。通过应用降维技术（如SVD），可以从中提取出更为紧凑、高效的词向量表示，这些表示捕获了词汇的主要语义特征。

1. svd分解的基本原理





对于奇异值,它跟我们特征分解中的特征值类似，在奇异值矩阵中也是按照从大到小排列，而且奇异值的减少特别的快，在很多情况下，前10%甚至1%的奇异值的和就占了全部的奇异值之和的99%以上的比例。也就是说，我们也可以用最大的k个的奇异值和对应的左右奇异向量来近似描述矩阵。

1. svd分解的具体代码实现

在计算较大的矩阵时，可以使用scipy库中的svds方法，对稀疏矩阵进行分解

1. # 使用 lil\_matrix 初始化稀疏共现矩阵，并指定为浮点数据类型
2. cooccurrence\_matrix = lil\_matrix((len(vocab), len(vocab)), dtype=np.float32)
4. **def** apply\_svd(self, cooccurrence\_matrix):
5. k = self.k
6. # 对稀疏共现矩阵应用 SVD 分解
7. U, Sigma, VT = svds(cooccurrence\_matrix, k=k)
8. U\_k = U
10. **return** U\_k

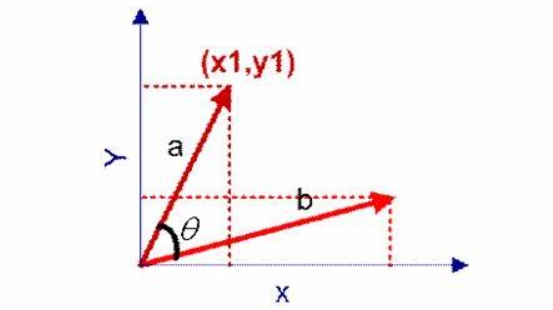
而在分析完整的奇异值性质的时候，需要进行完整的奇异值分解，可以使用scipy库中的svd方法

1. cooccurrence\_matrix = np.zeros((len(filtered\_vocab), len(filtered\_vocab)), dtype=np.float32)
3. **def** apply\_svd(self, cooccurrence\_matrix):
4. **print**("应用SVD分解...")
5. k = self.k
6. # 将稀疏矩阵转换为稠密矩阵，因为 scipy.linalg.svd 不接受稀疏矩阵
7. dense\_matrix = cooccurrence\_matrix
8. # 应用完整的 SVD 分解
9. U\_full, Sigma\_full, VT\_full = svd(dense\_matrix, full\_matrices=False, lapack\_driver='gesvd')
10. # 从完整的奇异值中获取前k个奇异值和向量
11. U\_k = U\_full[:, :k]
12. Sigma\_k = Sigma\_full[:k]
13. **return** U\_k

最终分解后得到每一个向量k个维度的词向量，这里的k就是新词向量的维度

## **2.4计算余弦相似度**

余弦相似度是衡量两个非零向量在方向上的相似度或者差异度的一种度量方法，不受两个向量的长度影响。在自然语言处理(NLP)和信息检索中，余弦相似度经常被用来计算两个文档或者两个词向量之间的相似度。

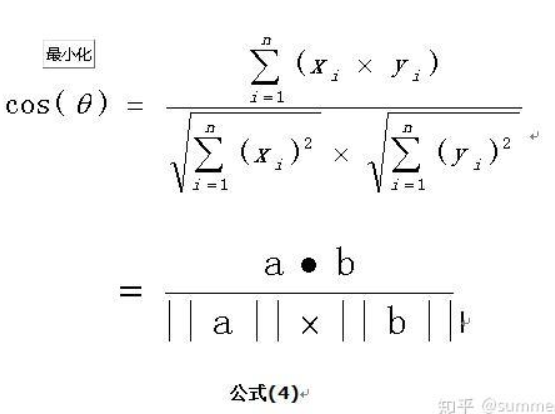


计算过程

计算点积：首先，计算两个向量的点积，即相应元素的乘积之和。

计算向量的范数：接着，分别计算这两个向量的欧几里得范数，即每个向量的元素平方之和的平方根。

计算余弦值：将两个向量的点积除以它们范数的乘积，得到的值即为余弦相似度。



1. **def** cosine\_similarity(self, vec1, vec2):
2. """计算两个向量之间的余弦相似度"""
3. dot\_product = np.dot(vec1, vec2)
4. norm\_vec1 = np.linalg.norm(vec1)
5. norm\_vec2 = np.linalg.norm(vec2)
6. similarity = dot\_product / (norm\_vec1 \* norm\_vec2)
7. **return** similarity
9. **def** calculate\_similarity\_pairs(self, reduced\_matrix, vocab\_index):
10. """计算pku\_sim\_test.txt中每一行两个词的余弦相似度"""
11. similarities = []
12. with open(self.test\_path, 'r', encoding='utf-8') as f:  # 确保使用正确的编码
13. **for** line **in** f:
14. word1, word2 = line.strip().split()
15. # 检查词汇是否存在
16. **if** word1 **in** vocab\_index **and** word2 **in** vocab\_index:
17. vec1 = reduced\_matrix[vocab\_index[word1]]
18. vec2 = reduced\_matrix[vocab\_index[word2]]
19. similarity = self.cosine\_similarity(vec1, vec2)
20. **else**:
21. similarity = 0  # 如果任一词汇不存在，相似度为0
22. similarities.append((word1, word2, similarity))
23. **return** similarities

cosine\_similarity函数实现了上述计算余弦相似度的过程。它接收两个向量vec1和vec2作为输入，首先计算它们的点积，然后计算各自的范数，最后用点积除以范数的乘积得到余弦相似度。

calculate\_similarity\_pairs函数遍历一个文件中的所有词对，对于文件中的每一行，提取两个词word1和word2。如果这两个词都在词汇索引vocab\_index中，它会用它们的向量（从reduced\_matrix中获取）来计算余弦相似度。如果任一词不存在于词汇索引中，它们的相似度被设置为0。最后，所有的相似度值被收集到一个列表中并返回。

## **2.5类结构化**

在分步按照流程完成各个步骤后，构建类Svd\_dec方法。这个过程包括文本的预处理、构建共现矩阵、应用SVD分解以及计算相似度。

类结构与参数

\_\_init\_\_: 初始化函数接收五个参数：停用词文件路径stopwords\_path、训练集路径train\_path、测试集路径test\_path、窗口大小window\_size和保留的奇异值数k。这些参数为处理文本和计算相似度提供了必要的配置。

preprocess\_file: 从训练集中读取文本，移除停用词，统计词频，并过滤掉低频词。这一步确保了后续处理的文本是干净且有意义的。

build\_cooccurrence\_matrix: 基于预处理后的词汇构建共现矩阵，这一矩阵记录了词汇之间的共现信息，为SVD分解提供了基础数据。

apply\_svd: 对共现矩阵应用SVD分解，以降低词向量的维度，并保留最重要的k个维度。这一步是整个过程的核心，它使用了scipy.linalg.svd来执行分解。

cosine\_similarity: 计算两个向量之间的余弦相似度，用于评估词向量之间的相似性。

calculate\_similarity\_pairs: 从测试集中读取词对，并使用之前计算得到的降维词向量和cosine\_similarity函数来计算它们之间的相似度。



## **2.6模型参数调整与计算**

实验要求中window\_size=5，故只需要调整降维后的维数即可

为了评估不同维数下的效果，我采取了计算输出结果的平均相似度和人工抽样观察法，来评估输出的相似度结果是否合理。

并在svd分解后计算总共有多少个非零奇异值，选取了多少个奇异值，选取的奇异值之和、全部奇异值之和以及二者的比例。来观察是否实践结果和理论一样，前几个奇异值之和占所有奇异值之和的绝大部分。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 降维后维数 | 平均相似度 | 选取奇异值数 | 选取奇异值之和 | 全部奇异值之和 | 二者比例 |
| 5 |  |  |  |  |  |
| 10 |  |  |  |  |  |
| 20 |  |  |  |  |  |
| 100 |  |  |  |  |  |

# SGNS方法

## **3.1构建skip-gram**

## **3.2实现SGNS模型**

## **3.3训练SGNS模型**

## **3.4参数调整**