speech and language processing 第六章报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. sparse vector和dense vector表示的学习的共通点是什么？

讨论后的理解：一般情况下，sparse vector考虑的是维度为词汇集合大小的维度的向量，而dense vector则是学习出的向量，其维度是根据人为选定的，但是其共同点在于都是用每一个维度去衡量一个word或者document的的特征，相比sparse vector，dense vector的特征更为集中。

1. 基于共现矩阵的表示方法的优势是什么？基于word2vec的窗口式的学习方法的优势是什么？

讨论后的理解：共现矩阵在某种程度上来看使用了全局的信息，比如同一个document中共同出现的情况以及词汇数量的信息，所以其优势是更加全面。而word2vec的特征在于其只考虑了局部特征，那么word的意思以及语法结构可以更好地学习，但是一些全局性的内容被忽略，而且缺乏解释性。

1. 负采样(negative sampling)为什么会有很好的训练效果?

讨论后的理解：基于负采样的skip-gram实际上是一个logistic regression，其被简化为了一个二分类模型，并且有着明确的目标函数，也就是让训练集中出现的内容出现的概率最大，而随机采样获得的样例出现的可能性最小，也就是最小化了随机性，这样设计非常巧妙，也非常简单。其相当于强化了正例的效果，而如果直接训练的话，人为选择负例实际上并不能确保所谓的杜绝随机性以及表现出英文的特征，从而影响训练的效果，而负采样的话则是出于杜绝随机性的考量提出的。

1. 别人提出的问题的理解
2. P15页为什么PMI为负值会造成影响？后面说到，unless our corpora are enormous,语料库大为什么能一定程度上解决这种问题？

自己的理解：这个问题我认为可能主要是负值的话不是很明显，比如书上的例子可以看出判断两个单词是不是真的比预计的出现概率小可能会出现10e-12次方的情况，而计算这种概率必须有较大的语料库才能比较准确的判断出来其数值是不是能达到这么小，而小语料库很难分辨其区别，并且对于人来说，不太同时出现的词语实际上也很难判断其是会随机出现还是会出现的情况更小，因为对于人的判断来说这并不明显，所以PMI的效果也很难衡量。

1. P24页6.11中的gender stereotype是不是指在学习语义的时候，可能出现性别偏见？a property of human reasoning是不是说学习的时候可能产生人种的歧视？

自己的理解：gender stereotype的意思应该是性别偏见，因为学习的材料中存在歧视的现象，比如医生这个职位或者科研工作者一般都联想到男性，女性的话则是护士，或者加上女医生的前缀。a property of human reasoning可能是说人类的联想，比如说美国人喜欢把负面的词汇和非裔美国人联系起来，这其实属于一种根深蒂固的偏见，而其在测试中反映出来，这种联想不一定就是歧视，比如说它也会把一些好的词语和某些人联系起来，所以这可以理解为人的一些偏见。

1. 读书计划

1、本周完成的内容章节：阅读第8章一部分，复习第6章。

2、下周计划：阅读完speech and processing第8章内容。

四、读书摘要总结：

1. Lexical semantics：

最早对语义的研究属于词汇语义，其一般是解释词汇，常用的就是英文字典的形式，并且早期的研究提出了诸如similarity，relatedness，word sense等概念。

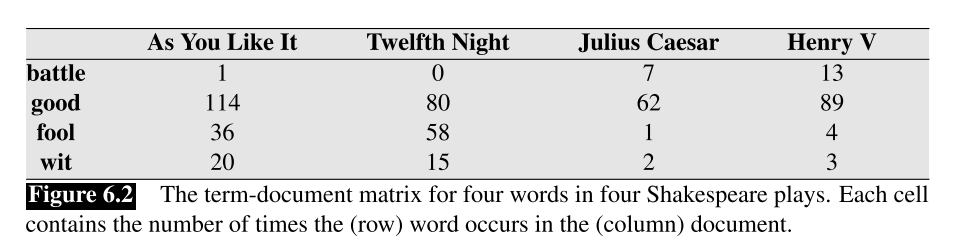
2. Vector Sematics:

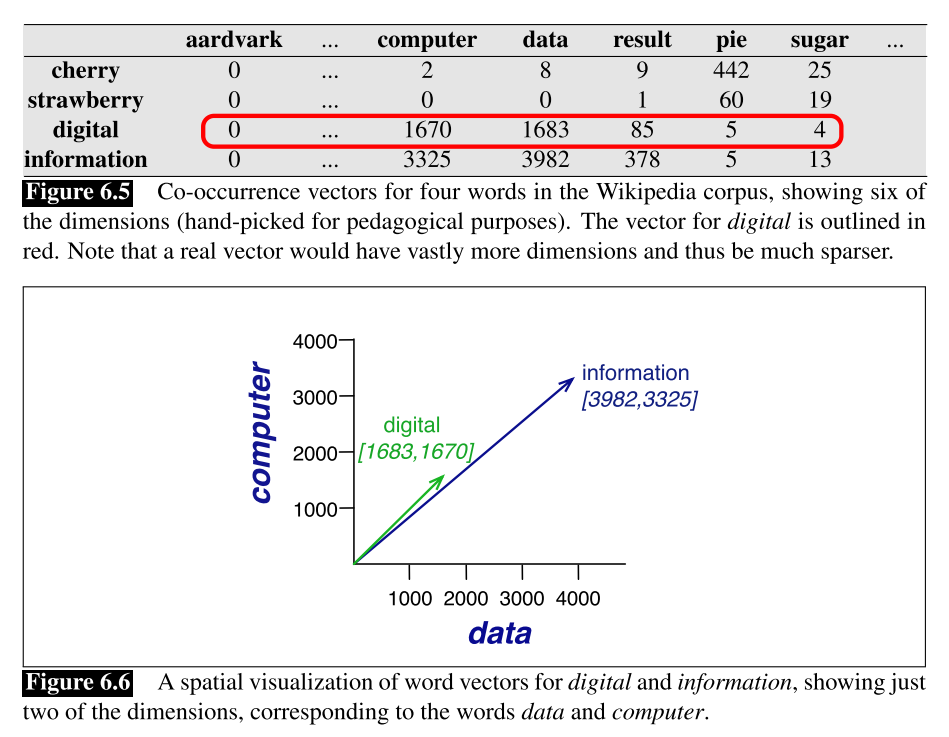
尽管词汇语义可以解决一些问题，但是在计算机领域，计算机无法直接理解词汇的人为描述，所以进行NLP工作也变得困难。最早对Vector Sematics产生启发的是对词义理解方式的转变从人为的解释变成了词汇如何在语言中使用表现了其意义。更进一步来说，可以使用词汇所处的上下文和分布位置来学习其语义，而不用依赖字典。

此外，另外的研究比如将词汇分解为几个维度，情感的好坏，强度，控制度并且就用三个参数表示词汇的做法使得词汇可以被展示在三维空间中。在这两种思路的引导下，逐渐出现了学习vector Sematics的算法。一般vector被称为embedding。

3. Words and Vectors：

最早的vector是基于共现矩阵的基础来学习的，共现矩阵有term-document矩阵和word-word矩阵等。

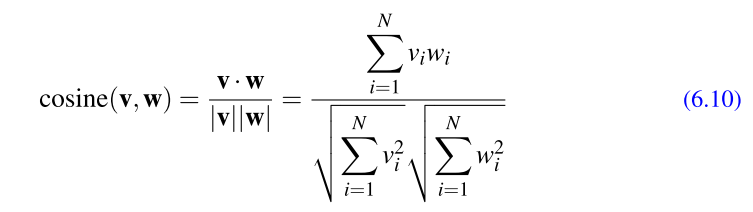




这些矩阵的行可以直接作为词汇的向量表示，因为其表示了词汇的分布特征以及出现频率。但是这样直接的处理的结果就是向量维度过大。但是其依然有着不错的效果。

4. Cosine for measuring similarity：

考虑到词汇的向量表示以及向量长度的影响，最终使用了余弦相似度作为衡量两个词汇是否相似的依据：

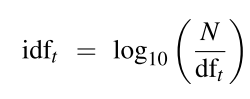


这种情况下，同方向的vector乘积会最大化，也最相似。

5. TF-IDF: Weighing terms in the vector：

余弦相似度有着很好的效果，但是其没有解决词频带来的问题，比如说the，like这样的词汇并不能对词汇的意义产生重要的作用，但是其却会增大相似度，从而造成干扰，为了处理这种问题，要对这样的词进行加权。加权一般使用TD-IDF来计算权值：



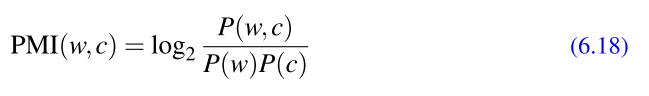




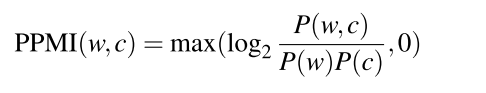
其考虑了一个词汇出现的频率以及其出现在不同document中的次数(越多越不重要)。

6. PPMI参数：

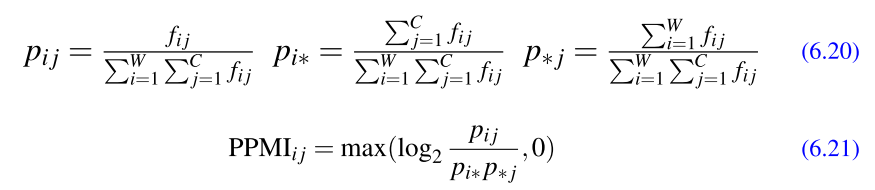
PMI用来衡量的是两个词汇共同出现的概率与两者随机的概率之间的关系：



其和之前的那些参数类似，如果两个词汇共同出现的概率远远大于两者随机，那么其就有明显的关系，而一般不考虑负PMI，所以：



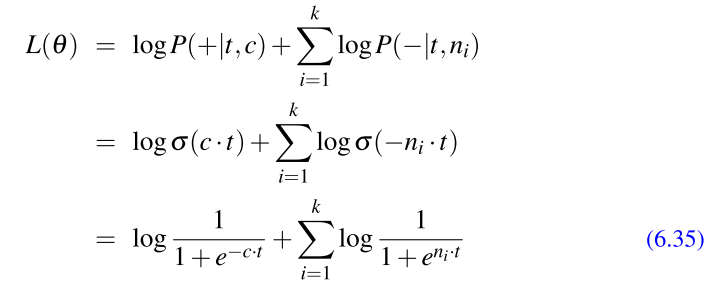
在文档与词汇中，其可以标识为：



7. word2vec：

上面涉及的内容都是sparse vector，而word2vec模型可以学习dense vector并且有着很好的效果。一般维度是50-1000，这是根据经验的计算结果。Skip-gram模型是一个有效率的训练词向量的工具，其训练了一个logistic regression classifier来对”两个词语出现在相邻的文本中”进行分类，其概率就是两个词汇向量的乘积，而模型本身并不重要，向量才是目标，越相似，出现在一起的概率越大。其使用了sigmoid函数进行分类，

其目标函数可以理解如下：



然后使用stochastic gradient descent来进行学习。其突破性的思路是引入了负采样，进行随机抽取，这样一方面减少了训练数据的收集难度，另一方面更加自动化，从而让训练变得快速，有效。学习结束后，将得到的词汇向量矩阵作为dense vector，实验表明，word2vec有着很好的效果。

8. 其他内容：

除了上面的内容，可视化，训练数据集存在的偏见问题以及对数据结果进行评估测试的数据库书中也进行了详细的介绍。