基于典型相关度分析的词向量模型

一个文档集D包含个术语（词）,这些词来源于一个词典V,词典中包含词汇的个数为.词典集中的每个词用一个维的指示向量表示,也就是对于第个词,表示为: (0,0…0,1,0,…0).其中第个位置为1,其余为0.文档集D可以表示为一个矩阵表示,其中行向量表示第个词。

例1. 文档中包含3句话

I ate harm. You ate cheese. You ate.

则可以表示为一个矩阵（包含3个.号）

词的上下文指的是一个词的前后（左右）词，上下文的前后长度为。如果，则为3-gram，也就是一个词的上下文用其左右的2个词（前后各1个）表示。如harm的left词为ate, 右词为.号。那么可以定义文档的上下文矩阵L和R分别是词的左上下文矩阵和右上下文矩阵。

左上下文矩阵L是文档中每个词的左（前）词一个矩阵，矩阵中每一个行向量表示文档中第个词的前面是词，同理有。合并LR可以得到词的上下文矩阵。

由于词典V包含词汇的个数为很大，所以每个词的向量的维度很高。另外词在词典中的向量没有任何的语义上下文信息。所以利用文档D构建一个嵌入式词向量。这个词向量的长度（维数）应该很低，达到降维的目的，同时包含了语义上下文信息。一个词的嵌入式向量是一个长度为的实值向量。

典型关联分析（Canonical Correlation Analysis）提供了很好的理论基础。假如一个对象有2个不同的多维观察角度，则可以得到一组数据，其中，。CCA度量2个不同观察角度和的相关性，也就是求2个矩阵之间的相关系数。

设分别表示向量的协方差矩阵。定义交叉相关矩阵为：



对矩阵进行SVD分解，得到：



其中

是一个正交矩阵；

是奇异值对角矩阵，且满足；

是一个正交矩阵。

称为典型相关系数。设，称矩阵的行向量是的典型变量。设，称矩阵的行向量是的典型变量。对于高维矩阵和，可以选择最大的个奇异值对应的典型向量作为原始矩阵的近似。

与上下文矩阵对比，我们可以认为一个词的左上下文和右上下文分别是对词的2个观察。对上下文和进行CCA计算，然后选择最大的个奇异值对应的典型向量，计算：



作为词的低维近似向量。

算法LRMVL:2viewCCA 是将采用光滑处理，得到上下文矩阵（不是直接文档的LR矩阵），然后计算的CCA，得到词的近似向量。

算法2viewWvsLR是对上下矩阵和进行CCA计算，然后得到词的近似向量。