# NLP大作业——Word2Vec模型

**学院：** 自动化科学与电气工程学院 **姓名：**王明贤 **学号：**ZY2103526

## 一、Word2Vec模型简介

在NLP的课程中已经学习了N-gram语言模型和LDA语言模型。Word2Vec模型是一种神经语言模型（NNLM），用一个一层的神经网络把one-hot形式的稀疏词向量映射称为一个n维的稠密向量的过程。

NLP中的词语，是人类的抽象总结，是符号形式的（比如中文、英文、拉丁文等等），所以需要转换成数值形式，或者说——嵌入到一个数学空间里，这种嵌入方式是词嵌入(word embedding), Word2Vec就是词嵌入的一种。

词向量基于语言模型的假设：“一个词的含义可以由它的上下文推断得出”，提出了词的Distributed Representation表示方法。相较于传统NLP的高维、稀疏的表示法(One-hot Representation)，Word2Vec训练出的词向量是低维、稠密的。

Word2Vec里面有两个重要的模型CBOW模型(Continuous Bag-of-Words Model)与Skip-gram模型。其中CBOW模型是根据前后文推断中间缺失的内容，Skip-gram则是根据中间内容联想前后文内容。

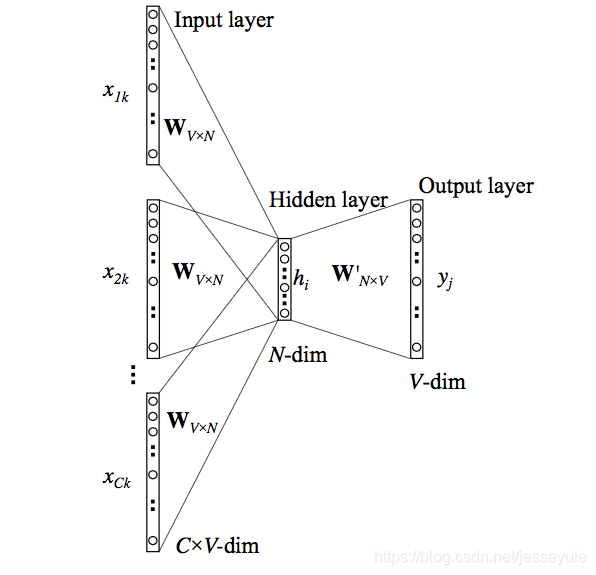


图 1 CBOW示意图

* Input layer层：是上下文单词的one hot。假设单词向量空间的维度为V，即整个词库corpus大小为V，上下文单词窗口的大小为C。
* 假设最终词向量的维度大小为N，则图中的权值共享矩阵为W。
* 将得到的Hidden layer向量与W转置相乘，并且用softmax处理，概率中最大的index所代表的单词为预测出的中间词
* 与groud truth中的one hot比较，求loss function的极小值

## 二、问题描述与分析

### 1.问题描述：

利用给定语料库（或者自选语料库），利用神经语言模型（如：Word2Vec， GloVe等模型）来训练词向量，通过对词向量的聚类或者其他方法来验证词向量的有效性。 截至日期：5月20日晚12点前

### 2.问题分析

整个问题包含三个部分：文本预处理、Word2Vec模型训练、训练结果分析。

文本预处理包括文本过滤、jieba分词等；模型训练调用gensim库中Word2Vec函数；结果分析主要为词向量的相似度分析，结合常识对结果进行判断。

## 三、运行结果

### 1.运行结果

本实验选取了数据库中所有小说，采用gensim库中Word2Vec模块，设置词向量维度为200，预测词上下文长度为5，选用CBOW模型进行训练。

表 1 小龙女人物关系相近度分析

|  |  |  |
| --- | --- | --- |
|  | 人物 | 相近度 |
| Top1 | 周芷若 | 0.801 |
| Top2 | 杨过 | 0.800 |
| Top3 | 丁典 | 0.788 |
| Top4 | 郭襄 | 0.777 |
| Top5 | 王语嫣 | 0.776 |
| Top6 | 闵柔 | 0.761 |
| Top7 | 思索 | 0.759 |
| Top8 | 瑛姑 | 0.752 |
| Top9 | 赵敏 | 0.750 |
| Top10 | 石清 | 0.748 |

表 2 张无忌人物相近度分析

|  |  |  |
| --- | --- | --- |
|  | 人物 | 相近度 |
| Top1 | 周芷若 | 0.801 |
| Top2 | 杨过 | 0.800 |
| Top3 | 丁典 | 0.788 |
| Top4 | 郭襄 | 0.777 |
| Top5 | 王语嫣 | 0.776 |
| Top6 | 闵柔 | 0.761 |
| Top7 | 思索 | 0.759 |
| Top8 | 瑛姑 | 0.752 |
| Top9 | 赵敏 | 0.750 |
| Top10 | 石清 | 0.748 |

### 2. 结果分析

从最终训练得到的结果中可以看出，与小龙女和张无忌相近的人物和印象中基本一致，同时也出现了一些不相关的词汇（也可能是我对小说不够熟悉）。试验记过基本可以说明算法模型的有效性。

## 四、总结体会

Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学词文本。在深度学习介入自然语言处理前，这个模型是一个很好的开拓性的模型，现在依然可以作为NLP的预处理为相关研究提供支撑。