**Report of Deep learning for Natural language Processing**

唐宗润 ZY2303211

**Abstract**

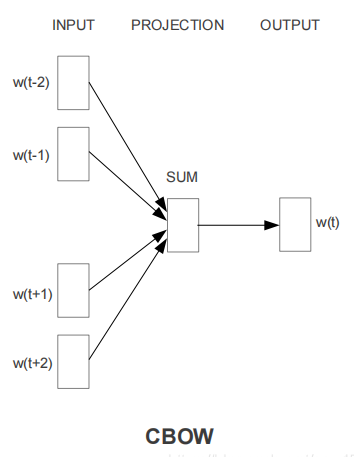
利用给定语料库，基于Word2Vec神经语言模型来训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联来验证词向量的有效性。

**Introduction**

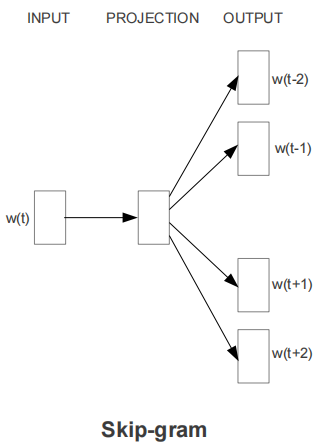
[Word2vec](https://so.csdn.net/so/search?q=Word2vec&spm=1001.2101.3001.7020)是一种用于将词语表示为向量的技术，它可以将每个词语映射到一个固定长度的向量空间。这些向量可以用于许多自然语言处理应用程序，如词语相似度计算和分类。Word2vec的主要优势在于它可以将相似的词语映射到接近的向量空间中，因此在许多情况下可以更准确地表示语义相似性。

Word2vec有两种主要模型：一种是基于神经网络的模型，称为Continuous Bag-of-Words (CBOW)模型；另一种是基于深度学习的模型，称为Skip-Gram模型。这两种模型都用于将词语表示为向量，但是它们的训练方法略有不同。CBOW模型通常更快，因为它的训练需要更少的迭代，而Skip-Gram模型通常更准确，因为它能够更好地捕捉上下文语境中的词语关系。

1.CBOW模型  
CBOW（Continuous Bag-of-Word Model）又称连续词袋模型，是一个三层神经网络。如下图所示，该模型的特点是输入已知上下文，输出对当前单词的预测。

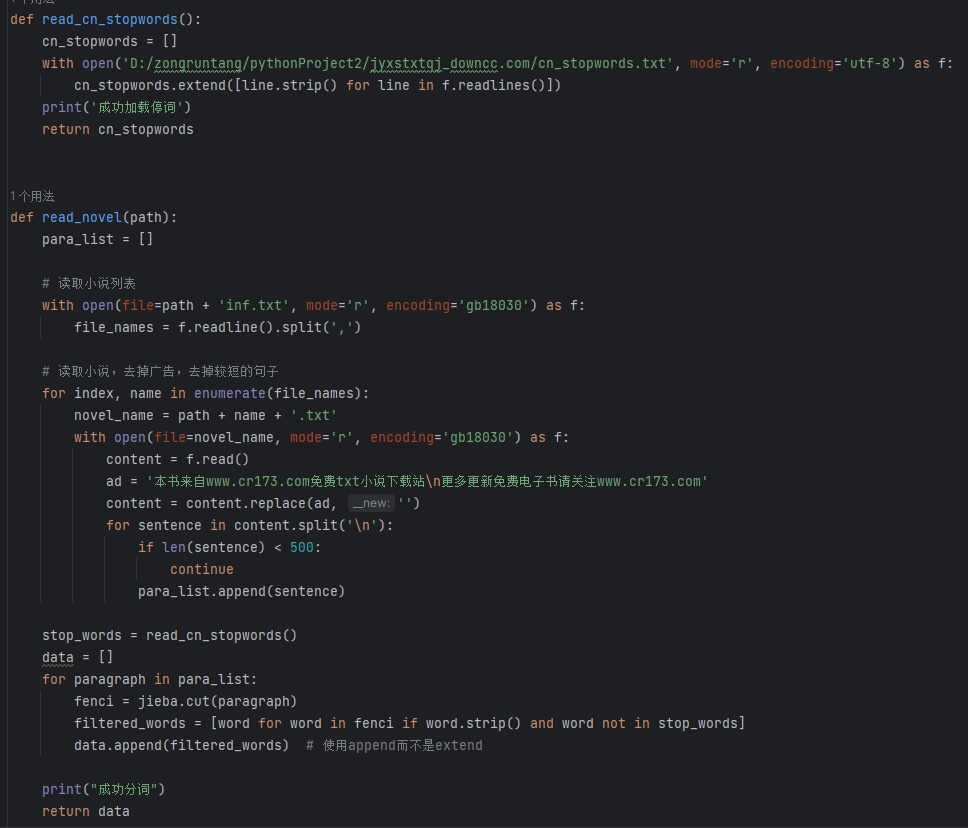


1. Skip-gram模型  
   Skip-gram与CBOW恰恰相反，即已知当前词语，预测上下文



**Methodology**

首先对数据进行预处理，将文件中停词去掉，广告去除，进行jieba分词。



其次，训练word2vec模型。其中

Sentences为需要处理的文本，vector\_size为词向量的维度，维，经过测试维度越高，越准确，但如果过高，则会出现泛化能力差或者过拟合的现象。

Window指窗口函数的大小，经测试，窗口函数大一些（如20）更倾向于提取文章话题上的特征，窗口函数小一些（如5）则更倾向于提取句子中语法上的特征。Min\_count为计入测试的最小出现次数，workers为并行线程数，epochs为迭代次数

图形用户界面, 文本

中度可信度描述已自动生成

接着，验证词向量的准确性。通过找与某个词最相近的是个词或者计算两个词的相似度，这两种方法进行验证

文本

描述已自动生成

主函数



Conclusion

（1）模型中window大小的选取

Window指窗口函数的大小，经测试，窗口函数大一些（如20）更倾向于提取文章话题上的特征，窗口函数小一些（如5）则更倾向于提取句子中语法上的特征。

文本

中度可信度描述已自动生成左图为window=5的结果 右图为window=20的结果

（2）运行结果分析

与 '丐帮' 最相似的词:

八袋: 0.5165

帮众: 0.5036

长老: 0.4707

群丐: 0.4570

帮主: 0.4554

帮中: 0.4536

本帮: 0.4373

化子: 0.4282

史火龙: 0.4253

韩山童: 0.4084

与 '内功' 最相似的词:

境界: 0.4990

修习: 0.4724

所习: 0.4575

武功: 0.4372

功力: 0.4340

内力: 0.4329

经脉: 0.4174

修为: 0.4012

根基: 0.3868

内息: 0.3858

与 '魔教' 最相似的词:

身败名裂: 0.4316

曲洋: 0.4230

东方不败: 0.3966

旁门左道: 0.3945

刘师兄: 0.3784

妖邪: 0.3738

黑木崖: 0.3677

正教: 0.3661

阴谋: 0.3567

与 '赵敏' 最相似的词:

张无忌: 0.5815

鹿杖客: 0.4632

周芷若: 0.4494

赵敏道: 0.3926

金花婆婆: 0.3576

鹤笔翁: 0.3543

玄冥: 0.3472

赵敏之: 0.3470

赵姑娘: 0.3454

谢大侠: 0.3449