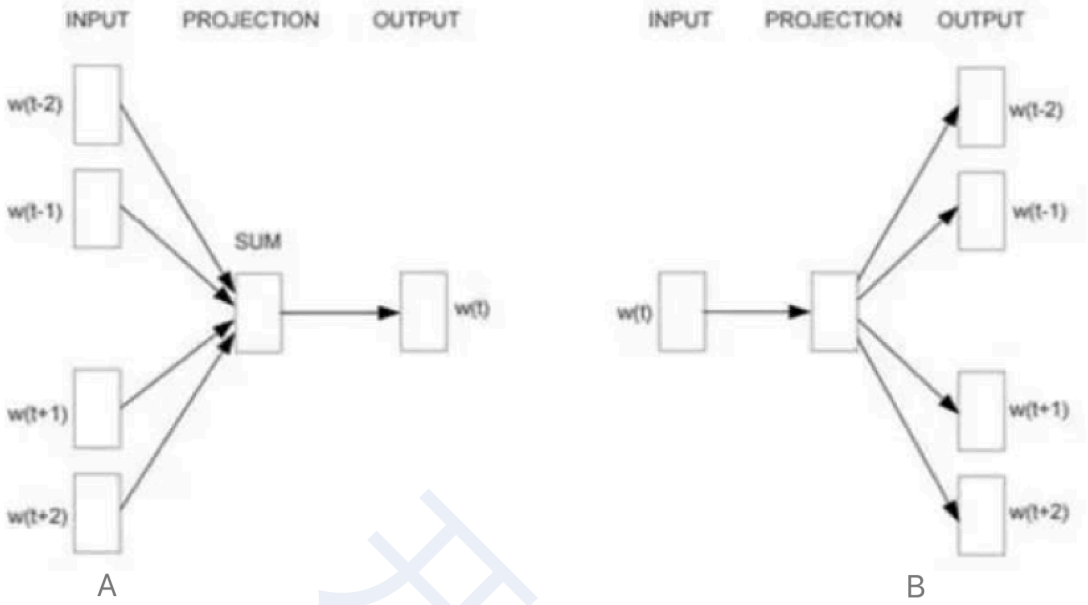
1. Skip gram 模型是在 Word2vec 算法中为词嵌入而设计的最优模型。以下哪一项描绘了 Skip gram 模型？



A、A
B、B
C、A 和 B
D、以上都不是

2. 中文同义词替换时，常用到 Word2Vec，以下说法错误的是

A、Word2Vec 基于概率统计
B、Word2Vec 结果符合当前预料环境
C、Word2Vec 得到的都是语义上的同义词
D、Word2Vec 受限于训练语料的数量和质量

3. 下列方法中，不可以用于特征降维的方法包括

A、主成分分析 PCA
B、线性判别分析 LDA
C、深度学习 SparseAutoEncoder
D、矩阵奇异值分解 SVD

4. 代码补全题：

对中文进行分词

"我爱北京天安门"————>"我 爱 北京 天安门"

设计如下函数，将对应输入转换为我们需要的分词之后的字符串。使用 jieba 补全如下代码：

```
def cut_word(text):
    # 用结巴对中文字符串进行分词
    text = _____
    return text
```

5. **（判断）TF-IDF 作用：用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。**

6. **（填空题）假如一篇文件的总词语数是 100 个，而词语"非常"出现了 5 次，那么"非常"一词在该文件中的词频就是（）。而计算文件频率（IDF）的方法是以文件集的文件总数，除以出现"非常"一词的文件数。所以，如果"非常"一词在 1,000 份文件出现过，而文件总数是 10,000,000 份的话，其逆向文件频率就是（）。最后"非常"对于这篇文档的 tf-idf 的分数为（）**

7. **Word2Vec 案例练习**

由于语料比较大，就提供了一个下载地址：http://www.sogou.com/labs/resource/cs.php
搜狗新闻中文语料(2.7G)
做中文分词处理之后的结果
保存加载模型，完成以下测试：

```
model.most_similar("警察")
model.similarity('男人','女人')
model.most_similar(positive=['女人', '丈夫'], negative=['男人'], topn=1)
```