一、 有一个文档集合包含 1700 篇文档，其中平均文档长度为 200 个 词，其中有一篇文档 d 包含了 300 个词。有一名用户在系统中输入查询 Q:cat care tips for healthy cat。已知 cat 在 100 篇文档中出现，在文档 d 中出现 2 次，求 cat 对于查询 Q 和文档 d 的 BM25 得分，结果保留两位小数。(b=0.6, k1=1.2, k3=8)。

二、 (1) 请简要解释 Word2Vec 模型中的 CBOW（Continuous Bag of Words）和 Skip-gram 两种架构的主要区别。

(2) 假设有以下 5 个单词的 Word2Vec 向量:

apple = [1,2,0,1]

book = [2,0,1,1]

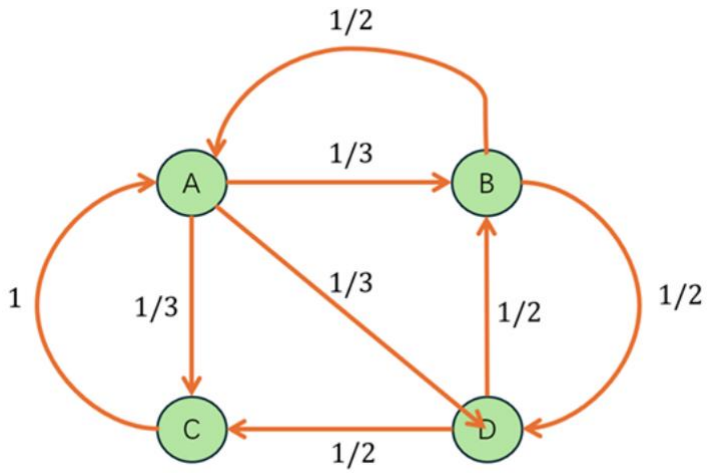
car = [1,0,2,1]

desk = [2,1,1,0]

pencil = [1,1,2,0]

请根据余弦相似度，给出与 desk 最相似的词。

三、 现有互联网网页 A、B、C、D，每个网页的初始跳转概率相同， 网页间的链接关系图如下所示，进行迭代计算，获取图中每个页面对应的 PageRank 。



四、 给定两个文档 D1 和 D2 的词频向量，使用余弦相似度,欧氏距离,Jaccard 相似度分别计算文档的相似性。

假设两个文档的词汇表已经统一为：

V={t1,t2,t3,t4,t5}

文档 D1 和 D2 的词频向量分别为：

D1=[3,0,2,5,1] , D2=[1,2,0,4,3]

五、 对于一个只基于返回网页的标题文本进行相关反馈的 web 搜索系统。用户给定查询 banana slug。返回的前三个网页标题分别是：

banana slug Ariolimax columbianus

Santa Cruz mountains banana slug

Santa Cruz Campus Mascot

该用户认为前两篇文档相关，而第三篇文档不相关。该搜索系统只基于词项频率(不包括长度归一化因子和 IDF 因子)进行权重计算，也不对向量进行长度归一化，并且使用 Rocchio 算法对原始查询进行修改，其中 a=1,B=0.75,y=0.25。请给出最终的查询向量(按照字母顺序依次列出每个

词项所对应的分量)。