**大致思路：**

由于仅给出两个文件，没有大量文件库，因此TF-IDF算法中的逆频率失去意义，使用余弦定理比较。

将两个文件中所有涉及到的词语整理出来，统计它们再两个文件中出现的次数，整理为两个向量，计算两个向量之间夹角的余弦值（广义），余弦值越接近1，说明两个文档相似度越高。

**具体实现：**

**读取过程：**由于用c++读取word文档过于繁琐，因此用vbscript编写读取过程，先转换为txt文件，再由c++读取。

**算法过程：**首先进行单词分离，由于顺序存储再后期查找复杂度过高，因此使用关联式容器map,分辨存储两个文档中所有的单词与其出现的次数。取两个文档中单词的并集，依据它们在两个文档中出现的次数得到两个向量，计算这两个向量夹角的余弦值，余弦值越大，说明相似度越高。