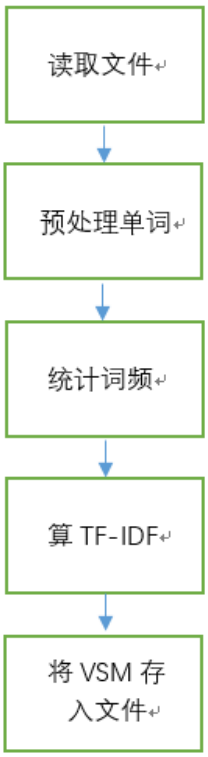
一、Vector Space Model (VSM)

VSM 是一种表示文档的方式，它用文档中单词出现次数构成的向量来表示文档，一个文档的维数为构建的词典的大小，向量中元素的大小为每个单词出现的频数。

二、Project Flow



三、问题及解决方案

1. Python

由于之前没学过 python，所以还是花了一些时间熟悉 python 的语法。但好在语法比较简单，而且目前也没有用到较难的语法，所以还挺顺利。

2. 文件编码问题

对编码问题本来就很迷，在编译程序时遇到 UnicodeDecodeError 时，查找了很多解决方案。在我的程序中采用了，读取文件时用 rb 模式读取、try-except 解决异常问题的方法来解决编码错误。

3. 内存溢出问题

减少了字典的规模（去掉词频为 0 的单词），但是内存占用率还是很高，决定在做 KNN 的时候，视情况再减小字典的规模。

四、实验结果

实验结果的输出格式为每个文件均给出字典中每个单词的 TF-IDF 值，并存入 VSM.txt 文件中。工程中附带 VSM.txt 文件。