主要步骤：

第一步 open 读入

第二步 预处理：读入每一个文本，用 while 去除每一个标点符号及换行符，调整大写字母为小写字母

第三步 分词 将预处理后的文本用 split 函数执行分割（空格）

第四步 统计词频 使用 count 函数 统计每个函数出现的频率 转换成 dict 形式

第五步 根据单词出现频率编码：将文本的 dict 执行排序，用以下这个代码。sorted(dic.items(), key=lambda d:d[1], reverse = True)；选择前 k 个高频词汇代表这个文本；对这 k 个高频词汇制作 dummy 编码；比如 L 文章的高频词汇是"today"和"yesterday", M 文章的高频词汇是"tomorrow"和"today"，K 文章的高频词汇是"today"和"yesterday"；如果想要再细分成 one-hot 编码。那么可以根据第(3)步中的表格执行种类划分。