大作业1 关键词提取

To do list

1. 提取出所有句子中的有用单词(去重，去掉冠词，网址等冗余信息)
2. 提取出所有的关键词，统计出现次数
3. 排序所有单词，关键词，建立唯一索引对应
4. 统计每个关键词下，每一个单词的出现次数，进一步计算出在每一个关键词出现时出现每个单词的概率
5. 学习朴素贝叶斯方法，并训练改进

朴素贝叶斯方法的基本思路：

给定新的句子，找出其中的有用单词，在已有数据的基础上，计算出每个单词出现时每一个关键词出现的概率，将其中所有单词下每个关键词的概率加权求和（权值个人感觉是这个算法需要不断训练的调整的关键），进而找出其中概率最大的关键词（朴素想法取最大的几个，在不要求关键词个数的情况下应该有对应阈值，在超过阈值的关键词内取前若干名）。

基本训练方法：

将原本给定的数据5000个句子随机取出其中若干，作为测试数据输入，程序给出所有结果调用评估程序与标准结果比对，以某种比较科学的算法给出分数，进而相应调整参数（主要是每个单词的权值），重复上述过程（个人感觉应该保持原始输入数据不变的情况下调整参数，一直调整到分数达到某个阈值，重新选取若干数据继续调整，最终最好情况下最好使得任意的组合都能得到比较理想的分数）。

待解决的问题：

1.权值调整算法

2.评估函数

3.对于有陌生单词的情况，应给出较为合适的权值，关于这个权值的估算算法（考虑和已有单词相似度等等）

4.是否还有更高级的不朴素的算法?