基于结巴分词的情感词典建立

一：结巴分词简介

**特点：**

1. 支持三种分词模式：

* 精确模式，试图将句子最精确地切开，适合文本分析；
* 全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；
* 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

1. 支持繁体分词
2. 支持自定义词典

**主要功能：**

1. 分词
2. 添加自定义词典
3. 关键字提取

参考链接：https://www.cnblogs.com/zhbzz2007/p/6177832.html

1. 词性标注

Jieba使用参考链接：https://github.com/fxsjy/jieba

二：情感词典的建立

购物领域情感词典构建：**评论语料情感关键词+基础情感词典+手工添加情感词。**

训练集中的评论语料情感关键词全部提取出来作为最基本的情感词典情感词。此时我们可以通过人工挑选一些基础情感词典中符合购物领域情感词和人工添加情感词两种途径扩展我们的最基本情感词典。

**基础情感词典**：基础情感词典包括了一些被广泛认同的情感词，比如”好“，”漂亮“，”差“，”烂“这些词。有研究者已经帮我们整理了这么一份情感词典。一个是著名的知网（Hownet）情感词典，还有一个是台湾大学简体中文情感极性词典。

但是这些基础情感词典有很多莫名其妙的词，需要我们人工判断并手动删除。

**改进点**：**拓展情感词典**，拓展情感词典其实就是把基础情感词典通过同义词词典找到情感词的同义词，这样就拓展了基础情感词典。

**人工添加情感词：**从训练集中人工挑选出一些未记录在情感关键词中的情感词，并找一些购物领域评论相关的数据集，人工标注一些情感关键词并加入到我们的情感词典。

**情感词典缺点:**

最后必须说明，利用情感词典来判断一个句子的情感是有着明显不足的。

中文有着丰富的语义表达，很多情感都是隐含的，比如：”我昨天吃了这道菜，今天就拉肚子了“。这句话没有一个情感词，但表达的是消极的情绪。还有各种事正话反说的句子，比如：“你说这里的菜很好吃，我只能呵呵了“。如果用词典匹配，有”好吃“，”呵呵“两个积极词，但这句话表达的绝非积极的情绪。

这里就需要更高级复杂的处理方式，要更深入句子的句法，语法了。

**改进点：**通过规则+情感词典的方式，但是中文语法规则非常多变复杂（我们只实现了最简单最基础的一些规则）。

补充：基于情感词典匹配的情感分析，我们首先需要对测试数据集进行分词，然后与情感词典进行匹配，此时为了加快效率，我们通常将分词后的数据进行去停用词，**停用词典**的构建可以通过一些常用的停用词典，通过合并并人工筛选的方式构建。这样将一些无关简要的词停用掉。